Quantifying Privacy Risks in Synthetic Data: A Study on Black-Box Membership Inference

Giacomo Fantino[1 [0009−0009−5808−604X]], Marco Rondina[1 [0009−0008−8819−3623]], Antonio Vetrò[1 [0000−0003−2027−3308]], and Juan Carlos De Martin[1 [0000−0002−7867−1926]]

Politecnico di Torino, Torino, Italy {giacomo.fantino, marco.rondina, antonio.vetro, juancarlos.demartin}@polito.it

Abstract. The use of synthetic data has grown steadily in recent years, particularly to support AI research and data sharing. However, synthetic data remains vulnerable to privacy risks such as membership inference attacks (MIAs), where an attacker identifies whether a data record was in the original dataset, whose recent variants increasingly exploit overfitting in generative models to boost their accuracy. Privacy metrics have been proposed to assess the protection offered by synthetic datasets and the risk of information leakage. However, their ability to reflect actual risks of MIAs remains unexplored. This study empirically evaluates the trade-offs between utility and privacy in the generation of synthetic tabular data leveraging a variety of black-box MIAs, providing a novel assessment of privacy risks. Using state-of-the-art generative models, we repeatedly generated synthetic datasets, assessed their utility, measured vulnerability to black-box MIAs, and evaluated privacy using commonly used privacy metrics. Our analysis reveals that CTGAN and CTAB-GAN+ can mitigate the risks of membership disclosure without significantly compromising the utility of the data, while the other generators showed weaker privacy-utility trade-offs. However, the analysis of the privacy metrics suggests that their reliance on proximity to training data limits their ability to fully measure an attacker's exploitation capabilities. The results observed in this study highlight the potential applicability of the aforementioned generative models to privacy-sensitive domains, demonstrating their ability to balance utility and privacy even under the challenge of diverse black-box MIAs. Our analysis of privacy metrics provides empirical evidence on the real-world privacy risks of synthetic tabular data and call for developing new, empirically validated privacy metrics.

Keywords: Synthetic Data · Membership Inference Attacks · Privacy.

1 Introduction

Providing access to code and data from AI/ML research studies and industrial applications is a step forward in addressing the reproducibility crisis [15], while also enhancing the transparency and reliability of AI system development in

industrial applications [27]. However, although transparency, accountability and reproducibility are fundamental aspects to achieve, compromising the privacy of individuals is not acceptable: if personal data is used, it cannot be shared, and even if no personal data is included in the training processes, it is still possible to uniquely identify individuals [23].

Synthetic data offers a promising means to address this trade-off: by creating a new anonymized dataset that has the same general statistical characteristics of the original data, it is possible to train a new model and obtain results that are consistent with the use of the original data [12]. Many generators have emerged, most recently based on architectures such as Generative Adversarial Networks (GAN) [14] and Variational Autoencoders (VAE) [22]. However, these approaches are undermined by a number of attacks: using the generated synthetic data and integrating it with additional information, an attacker can infer part of the original data, including sensitive attributes.

Among these attacks, Membership Inference Attacks (MIAs) have emerged as one of the most effective techniques, aiming to determine whether a particular data record was part of the training dataset of the target model [20]. The ability to infer such information is a serious threat to the privacy and fairness of an individual, as it may disproportionately expose vulnerable groups or amplify biases in decision-making. Exposure of sensitive information about an individual's health could be inferred through such attacks [18]. Bias amplification could occur in the employment sector, where sensitive details about a person's socioeconomic or ethical background could lead to potential biases in the hiring process [35]. Therefore, synthetic data generation that is trained on personal data or other types of sensitive information needs to be made robust against privacy attacks. This paper focuses on the MIA attack due to the growing research on this topic in the last years [31] and its impact on fairness and privacy.

Existing work has evaluated synthetic data with respect to membership inference attacks [21, 24, 45]; however, with the advent of new attacks [19, 3], a comprehensive evaluation of the privacy robustness of synthetic data generators for tabular data and privacy metrics remains absent. In this paper, we perform an evaluation of the robustness of synthetic data generators for tabular data against recently proposed MIAs, and analyze the extent to which commonly used privacy metrics reflect the actual success of these attacks.

The main contribution of this study is as follows:

1. We use empirical analysis to assess the risk of using synthetic data to share data containing sensitive information or personal data under multiple black-box membership inference attacks.
2. We evaluate the extent to which commonly used privacy metrics accurately reflect actual membership inference attack success.

To achieve these goals, we evaluated four state-of-the-art synthetic tabular data generators on datasets containing personal information, measuring their resilience to MIAs and analyzing how three privacy metrics correlate with attack success.

The rest of the article is structured as follows. Section 2 provides an overview of synthetic data generation techniques (Sect. 2.1), attacks on synthetic data (Sect. 2.2), and privacy metrics (Sect. 2.3). Section 3 presents the related work. Section 4 introduces our research questions and the privacy metrics used for evaluation. Section 5 describes our methodology in detail. In Section 6, we analyze the results and address our research questions. In Section 7, we discuss the limitations of our approach. Finally, Section 8 presents our conclusions and suggests directions for future research.

2 Background

2.1 Synthetic data generation

Synthetic data is obtained through models that reproduce the statistical properties of real data [12]. While this study focuses on tabular data, similar approaches have also been applied to other modalities such as images and audio [11]. Typical motivations for generating synthetic data include the removal of personal or sensitive information [12], data augmentation in data-scarce contexts [28], and fairness enhancement in machine learning applications [4].

Synthetic data generation methods can be broadly categorized by their underlying mechanisms [12]. Three main classes can be distinguished: Probability-based, where synthetic data is generated by estimating and sampling from a data distribution; Randomized, which involves generating samples through perturbation or interpolation; and Network-based, which encompasses generative methods that utilize neural network architectures.

Simple interpolation-based techniques such as Mixup represent baseline strategies. Mixup generates synthetic samples by linearly combining pairs of real instances using a mixing parameter λ drawn from a beta distribution. While effective for data augmentation, such methods provide limited control over privacy preservation or distributional fidelity.

More sophisticated network-based methods have recently gained prominence, particularly those based on Generative Adversarial Networks (GANs) [5]. GANs comprise two models trained in opposition: a generator and a discriminator. During the learning phase, the generator learns to mimic the training data distribution, while the discriminator attempts to distinguish real from synthetic samples. Through this adversarial process, the generator is optimized to produce synthetic data that are increasingly indistinguishable from real data.

Another prominent family of generative models are Variational Autoencoders (VAEs) [22], which learn probabilistic latent representations of the input data. A VAE consists of an encoder that maps data to a latent space and a decoder that reconstructs new samples by sampling from this latent representation.

These architectures were initially designed for continuous domains such as image and audio synthesis [11, 36]. When applied to tabular data, however, their performance may degrade due to the presence of categorical variables and complex feature dependencies [41]. To address these challenges, specialized models have been proposed.

CTGAN is a state-of-the-art GAN-based generator for tabular data [41]. It employs a preprocessing stage based on a Variational Gaussian Mixture Model (VGM) to normalize continuous variables and applies one-hot encoding to categorical columns. A conditional vector guides the generator toward learning minority categories in unbalanced features.

TVAE, proposed by the same authors, adopts the same preprocessing pipeline but replaces the GAN architecture with a VAE, enabling probabilistic sampling while maintaining compatibility with tabular structures [41].

CTAB-GAN+ extends CTGAN with domain-specific improvements [44]. Continuous features are min–max scaled for better VGM compatibility, long-tailed distributions are log-transformed to prevent unrealistic values, and mixed-column strategies handle highly unbalanced or sparse features. These refinements yield more stable and realistic tabular data generation, particularly in datasets containing heterogeneous variable types.

2.2 Attacks

With the growing adoption of generative AI across domains, adversarial methods have emerged to extract sensitive information from trained models [25]. Among these, the Membership Inference Attacks (MIAs) are one of the most studied, as they directly target the confidentiality of training data. In this type of attack, an adversary aims to determine whether a specific data sample was part of a model's training data [37].

Existing surveys classify MIAs according to the type of target model and the nature of the data [31]. Early work focused on discriminative models [37], while subsequent work extended MIAs to generative models such as GANs [17]. This evolution led to increasingly sophisticated attack strategies designed to enhance attack performance [19, 6, 3].

MIAs can be broadly divided into two families: black box and white box settings. In the white-box scenario, the attacker can inspect model parameters or architectures, whereas in the black-box case, only the synthetic data is observable. While both settings are relevant and capture different threat models, this study focuses on the black-box scenario, which naturally arises in many realistic data-sharing and publishing use cases where only synthetic datasets are released and the underlying generative model remains inaccessible.

Formally, the MIA can be expressed as a function $A(x|G)$, estimating the probability that sample x belongs to the training set underlying the synthetic dataset G [6]. Assuming that the synthetic dataset closely resembles the distribution of the real data, this probability can be derived from the learned density, denoted as $P_G(x)$. The resulting formulation, where membership is inferred from the synthetic data distribution, is known as a distribution-based MIA.

In practice, this remains challenging because estimating the distribution may require many samples, which the attacker may not have. To address this, distance-based MIA infers membership by assuming that real samples close to synthetic records are more likely to originate from the training data [6]. Let

$d(x, x')$, denote a distance function between x and its nearest synthetic neighbor $x' \in G$, then:

$$A(x|G) = \min_{x' \in G} d(x, x') \tag{1}$$

The Monte Carlo-based MIA improves upon distance-based MIAs by considering multiple synthetic samples instead of just one [19]. Its key insight is that an overfitted generator tends to repeatedly produce samples near real training instances. Only distances below a threshold ϵ are considered, ensuring that outliers are excluded from the analysis.

A more recent variant, DOMIAS (Detecting Overfitting for Membership Inference Attacks against Synthetic Data) [3] assumes limited access to a subset of training data, denoted as R. The key idea is to compare the likelihood of a sample x under the synthetic data distribution G and the known training subset R. This is done by computing the ratio:

$$A(x|G) = \frac{p_G(x)}{p_R(x)} \tag{2}$$

A high ratio suggests that the synthetic data generator has overfitted to x, making it more likely that x was part of the training set.

Another line of work, shadow modeling, involves the attacker training a surrogate generator locally using the released synthetic data [17]. This surrogate allows the use of white-box inference methods: high discriminator confidence in the shadow model indicates a stronger likelihood of membership.

2.3 Privacy metrics

Synthetic data privacy metrics provide quantitative proxies to assess how effectively synthetic datasets balance privacy and utility. They aim to capture potential privacy leakage risks, especially under adversarial settings such as membership and attribute inference attacks [39]. We focus on three widely adopted metrics—Distance to Closest Record (DCR) [29], Nearest Neighbor Distance Ratio (NNDR) [26], and Privacy loss [42]—chosen for their prevalence in both academic and industrial practice [32, 43, 42]. Newly proposed metrics are outside the present scope and reserved for future work.

The DCR measures the Euclidean distance between each synthetic sample and its nearest neighbor in the original training data. A high concentration of synthetic samples with DCR values close to zero suggests potential information leakage, as these samples closely resemble individual training records. On the other hand, NNDR evaluates the ratio between the distances of a synthetic sample to its closest and second-closest real neighbors. A NNDR value close to zero indicates that the nearest neighbor is significantly closer than the second, implying that the synthetic sample may inadvertently reveal details about a specific training instance. In contrast, when the NNDR value is close to 1 both distances are similar, thus the sample likely resides in a densely populated region

of the feature space, thereby mitigating identifiability risks [26]. To prevent any feature from dominating the computed distances, all variables are normalized prior to measurement for both metrics.

A direct comparison between the synthetic and training sets alone fails to capture the broader distributional context in which samples are embedded. To address this limitation, metrics such as DCR and NNDR are computed using not only the training data but also a holdout set: a dataset drawn from the same distribution as the training data but excluded from the training process [33]. By comparing synthetic samples to both the training and holdout sets, one can assess whether proximity to the training data is exceptional or expected given the overall data distribution. If synthetic samples are similarly close to both sets, it suggests that they preserve general statistical properties without overfitting to specific training records. This adjustment makes DCR and NNDR more robust indicators of potential overfitting.

While these two metrics have been widely adopted, studies have highlighted their limitations, noting the lack of strong theoretical grounding and the challenges they face in capturing privacy risks for outliers or minority groups [13]. As they remain the most commonly used indicators in both academic and industrial settings, evaluating these limitations empirically is therefore crucial to understand how well such metrics reflect actual privacy risks.

The Privacy Loss metric quantifies how much a generative model exposes information about its training data [42]. It measures the difference in an attacker's ability to distinguish between samples that were part of the training set and those that were not. If the attacker cannot make this distinction, both accuracies are approximately 0.5, resulting in zero privacy loss. Conversely, if the model is overfitted, the accuracy on training samples will be considerably higher, leading to a high privacy loss. In our study, this attacker is instantiated through the Nearest Neighbor Adversarial Accuracy (NNAA) framework [42].

3 Related Work

Early research on applying MIAs to synthetic tabular data focused primarily on distance-based approaches [21, 24, 45]. These studies reported high attack success rates but also revealed the difficulty of achieving an optimal privacy–utility balance when using GAN-based or other generators. Parallel work explored shadow modeling techniques, where a local model is trained to replicate the target's behavior. While early attempts on synthetic image data achieved limited performance without auxiliary information [17], later studies using model predictions improved inference accuracy [37, 30]. Further evaluations combining distance-based and shadow-based attacks confirmed that naive MIAs were weak, but shadow modeling substantially increased inference success, particularly when no explicit defenses were applied [38]. These findings underscore the need for a unified empirical framework capable of evaluating multiple MIA families under consistent conditions.

Building on these findings, we quantitatively assess whether the use of synthetic data generators for sharing tabular data poses a risk to privacy. While prior studies have largely focused on distance-based MIAs or shadow modeling, our work expands this scope by incorporating newer black-box MIA attacks, aiming to provide a more comprehensive evaluation of privacy vulnerabilities. Our analysis takes into account a variety of synthetic data generators, three datasets, all containing personal data, and the aforementioned black-box MIA attacks.

Due to the high computational cost of training multiple shadow models and given that this attack has been extensively studied in the literature, we did not include shadow modeling in our main analysis. Its resource intensity makes it impractical in many real-world scenarios; however, we consider it a valuable direction for future work. Instead, we focus on the most representative black-box MIAs and complement our attack-based evaluation with a detailed assessment of widely adopted privacy metrics, analyzing their reliability in reflecting actual privacy robustness.

4 Research Questions and Metrics

The Research Questions that drive the study are:

- RQ1: How do different synthetic tabular data generators perform in terms of utility and resistance to a variety of black-box membership inference attacks: distribution-based MIA, distance-based MIA, Monte Carlo-based MIA, and DOMIAS?
- RQ2: Are the selected privacy metrics, Distance to Closest Record (DCR), Nearest Neighbors Distance Ratio (NNDR) and Privacy loss, reliable for evaluating synthetic data quality when faced with a variety of black-box MIAs?

To address RQ1, we evaluate two complementary aspects: utility and privacy.

Utility of a synthetic dataset can be evaluated using both statistical properties of the data, i.e. measuring the difference between the trained and synthetic data, and machine learning, which measures the difference in performance between training a model with the original data and the synthetic data [18]. Since it has been shown a weak correlation between statistical utility metrics and an overall absence of utility measurement [8], we focus on the machine learning utility: we measure the difference in accuracy, F1-score and AUC-ROC when training an XGBoost classifier [7] on synthetic data. A minimal drop in performance implies better synthetic data. This approach provides a consistent and reproducible basis for comparing generators using measurable downstream performance. In addition, we assess the coverage of the synthetic data by comparing, for each training record, its distance to the nearest synthetic sample and to the nearest holdout sample. This helps reveal whether the generator misses parts of the real distribution or places synthetic records too close to training data (a sign of overfitting) [33].

For privacy evaluation, we will use the black-box MIAs discussed in Section 2.2. Each attack was executed on balanced subsets consisting of 20% of training samples (expected to be classified as members of the training set) and an equal number of test samples (expected to be classified as non-members). The attack success was measured using the AUC-ROC score, where values near 0.5 indicate resistance (equivalent to random guessing) and values approaching 1.0 indicate a privacy risk [10].

While RQ1 measures empirical robustness, RQ2 examines whether existing privacy metrics can reliably predict this robustness. To this end, we compute DCR and NNDR for both training and holdout data, and Privacy Loss for the synthetic data, expecting a positive correlation between the metrics' values and the AUC-ROC of the attacks

Given a synthetic dataset S and the original dataset R, DCR is computed for all synthetic samples as the distance to the closest sample in R:

$$min_{r \in R} d(s, r) \tag{3}$$

On the other hand NNDR considers the ratio for all synthetic samples of the closest and the second-closest samples in R:

$$\frac{d(s, NN_1(R))}{d(s, NN_2(R))} \tag{4}$$

For measuring Privacy Loss, we adopt Nearest Neighbor Adversarial Accuracy (NNAA) to assess an attacker's performance on distinguishing real data from synthetic counterparts [42]. Real samples that remain distant from synthetic data are treated as true positives, while synthetic samples that remain distant from real data are considered true negatives. NNAA is computed as the balanced accuracy between the true positive rate and the false negative rate; an NNAA value of 0.5 corresponds to random guessing. Privacy Loss is then measured as the discrepancy in attacker performance between training and holdout samples: if NNAA is close to 0.5 on both sets, the resulting difference is near zero, indicating minimal membership leakage, while larger gaps indicate increased leakage due to overfitting.

5 Methodology

This section details the three main phases of our methodology: data pre-processing and splitting (Section 5.1), synthetic data generation (Section 5.2) and attack execution (Section 5.3). Figure 1 provides an overview of the experimental workflow, designed to address both RQ1 (utility and resistance of synthetic data generators to black-box MIAs) and RQ2 (reliability of privacy metrics). First, the datasets were preprocessed and split into training and test sets. The training set was used to generate synthetic data, while the test set served as an independent evaluation set. Finally, the synthetic data was evaluated using MIAs, privacy metrics, and utility measures to answer the research questions. Defensive

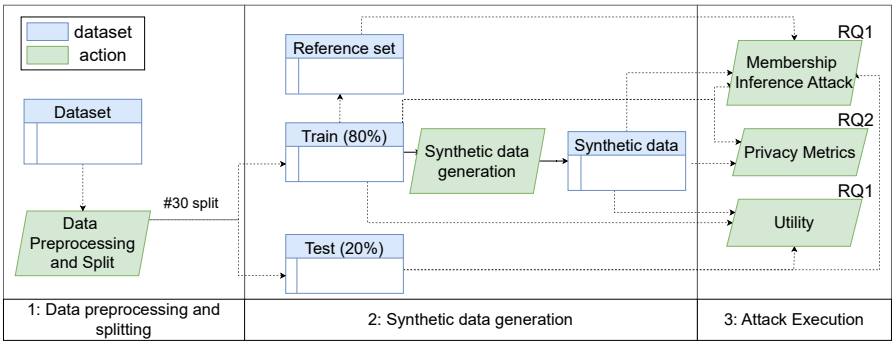

Fig. 1. Overview of the proposed methodology, divided into three main phases: data preprocessing and splitting, synthetic data generation, and attack execution.

mechanisms such as Differential Privacy (DP) were intentionally excluded. Preliminary exploratory tests using conservative privacy budgets for GAN-based generators indicated a reduction in attack performance, approaching random guessing, but at the cost of a substantial degradation in data utility. As differential privacy entails an explicit privacy–utility trade-off that depends critically on the choice of parameters, and a systematic evaluation of this trade-off falls outside the scope of this work, we focus on the inherent privacy–utility properties of existing generators without DP. The integration of DP-based defenses is therefore left for future work.

5.1 Data Preprocessing and Splitting Strategy

We used the Adult [2], COMPAS [34], and Southern German Credit (hereafter referred to as Credit) [1] datasets, which are widely adopted benchmarks in fairness research [9] and exhibit complementary structural characteristics, varying in sample size, feature dimensionality, and balance between categorical and numerical attributes (Table 1). Each dataset underwent preprocessing to ensure consistency and privacy compliance: duplicates and missing values were removed, and features typically excluded in standard machine learning pipelines, including identifiers that could link a record to an individual, were dropped. To enhance statistical robustness and reproducibility, we generated 30 independent train–test splits for each dataset, thereby mitigating the variability introduced by random partitioning. For each split, 80% of the data was used to train the generative models, while the remaining 20% served as a holdout set for evaluating both synthetic data utility and resistance to MIAs. To ensure comparability and simulate realistic data-sharing scenarios, the synthetic dataset generated in each run contained the same number of samples as the corresponding training set.

Dataset	#Samples	#Columns	#Categorical
Adult	26,904	13	12
COMPAS	2,294	9	9
Credit	1,000	21	21

Table 1. Dataset statistics after preprocessing

5.2 Synthetic Data Generation

We employed the four synthesizers introduced in Section 2.1: Mixup, TVAE, CTGAN, and CTAB-GAN+, using off-the-shelf implementations from publicly available libraries and repositories. Since TVAE and GAN-based generators require model training, we tuned hyperparameters—such as number of epochs, batch size, and discriminator steps for GANs—to ensure convergence and maximize fidelity of the generated samples. The tuning process prioritized data-utility metrics, reflecting common practice where generative models are optimized for downstream performance rather than explicit privacy preservation.

5.3 Attack Execution

The membership inference attacks were executed across all dataset splits to mitigate bias arising from random partitioning, as certain training subsets may inherently be more vulnerable to inference. Each attack received the synthetic dataset, a subset of the training set and the test set: the attack should infer membership with maximal accuracy. All attacks were re-implemented by the authors, closely following the threat models and parameter choices described in the respective original papers. Prior to execution, data were preprocessed to ensure comparability across attacks: Min–Max scaling was applied for distance-based MIAs to prevent feature dominance, Standard scaling for density-based attacks to stabilize density estimation, and PCA to remove low-variance features. The AUC-ROC metric was computed across splits to assess attack effectiveness. Although it may not capture all aspects of attack performance, it remains the most widely used metric in the literature for assessing MIA success, allowing direct comparison with previous work.

For the DOMIAS attack, a subset of the training data was used as a reference set. This reference set represents a portion of the leaked training data that the attacker can access to infer information about the overall training distribution. To ensure a realistic evaluation, this reference set was completely disjoint from the subset of training data used for assessing the attack, preventing trivial membership identification. We initially set the reference set to 10% of the training data and later examined the impact of increasing its size.

Complementary to the attack analysis, the utility and privacy metrics were computed for each split to quantify the trade-off between data utility and MIAs resistance. For the privacy metrics, we reused the test sets as holdout sets to account for the data distribution. After computing the distances, following the

methodology of Platzer and Reutter [33], for each synthetic sample we determined whether the closer sample was in the training or holdout set. We then calculated the ratio of samples closer to the training data: as the value gets closer to 1 the probability of information leakage increases.

Since the Monte Carlo–based MIA requires a value for the threshold parameter ϵ, we have used the median heuristic proposed by Hilprecht et al. [19]: it computes ϵ as the median of the minimum distances between the samples in the training and test sets and the synthetic samples, ensuring stable and reproducible attack behavior across splits.

Finally, Kernel Density Estimation (KDE) was used to approximate the distribution of both synthetic and real data. A grid search over multiple splits identified an exponential kernel with a bandwidth of 0.05 as optimal for modeling the distribution. The same configuration was used for DOMIAS, following prior evidence that it provides robust estimation under limited sample sizes.

6 Results and Discussion

In this section, we present the results addressing the research questions defined in Section 4. Specifically, in Section 6.1 we examine the outcomes of MIAs to assess the resilience of different synthetic data generators, considering both their privacy protection and the utility of the data produced. Subsequently, Section 6.2 investigates the relationship between the computed privacy metrics and the corresponding AUC-ROC values of the attacks, to determine whether such metrics can reliably predict a generator's resistance to black-box MIAs. All code and datasets are available for reproducibility.[1]

6.1 Utility and Resistance against black-box MIAs (RQ1)

This subsection addresses RQ1 by jointly analyzing the utility of the synthetic data and its resistance to black-box MIAs. We compare the AUC-ROC values of different attacks across generators and datasets, summarizing the distributions over 30 iterations of synthetic data generation.

Figure 2 shows the AUC-ROC distribution for each attack, with respect to the datasets and the generators. Mixup consistently yields the highest AUC-ROC values across attacks and datasets, reaching 0.8 on the Credit dataset, indicating strong vulnerability to MIAs likely due to its interpolation-based generation strategy. In contrast, TVAE, CTGAN, and CTAB-GAN+ exhibit near-optimal resistance, with AUC-ROC values below 0.6 across all attacks and datasets. Since an AUC-ROC of 0.5 corresponds to random guessing, values around 0.8 reflect strong separability between members and non-members, whereas values only moderately above 0.5 indicate a limited attacker advantage [10]. DOMIAS and Monte Carlo achieve only marginal gains against these generators, further confirming their robustness even against recent and more sophisticated MIAs.

[1] Source codes and data are available at: https://github.com/giacomofantino/Synthetic-Data-Privacy

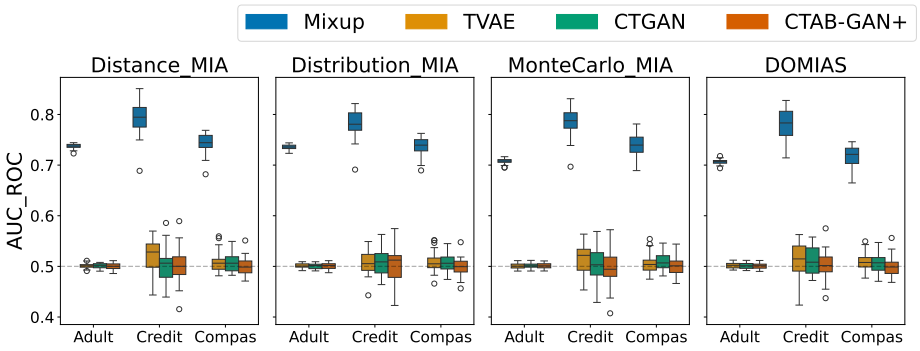

Fig. 2. AUC-ROC scores for each Membership Inference Attack across different datasets and synthetic data generators.

These findings are broadly consistent with prior observations on distance-based MIAs for tabular data [21, 45], reinforcing their validity across different settings. A direct comparison with [24] is less straightforward due to methodological differences (e.g., their use of a fixed synthetic sample size independent of the real dataset), but the overall trend of low attack performance against GAN-based generators is comparable. Our work extends these studies by systematically evaluating a wider range of black-box MIAs, including more recent attacks such as DOMIAS and Monte Carlo. Crucially, we show that generators previously regarded as strong (TVAE, CTGAN, CTAB-GAN+) retain their robustness even under these novel attacks. Finally, the consistently low AUC-ROC values across generators validate our decision to exclude differential privacy: given the already strong resilience of these models, adding DP would likely degrade utility without offering meaningful privacy gains.

Attack performance also varies with dataset characteristics. Mixup remains highly vulnerable across all datasets, though its AUC-ROC values are slightly lower for COMPAS and Adult than for Credit. CTGAN, CTAB-GAN+, and TVAE maintain low AUC-ROC values overall, but Credit and COMPAS display greater variance, consistent with prior findings that smaller training datasets increase susceptibility to information leakage.

The utility evaluation, shown in Figure 3, reveals distinct trade-offs across generators. Mixup consistently achieves the smallest differences in all three metrics, with values clustered just below zero, suggesting that classifiers trained on Mixup-generated data closely replicate the performance obtained on real data. However, this high utility comes at the cost of substantial privacy risk, as confirmed by the attack results. GAN-based generators show mild performance degradation, particularly on the Credit dataset, where CTGAN and CTAB-GAN+ display noticeable declines in AUC-ROC. Conversely, CTGAN performs well on COMPAS, achieving high utility with positive metric differences, while CTAB-GAN+ tends to underperform on this dataset. TVAE represents an intermediate case, showing moderate differences across metrics and datasets.

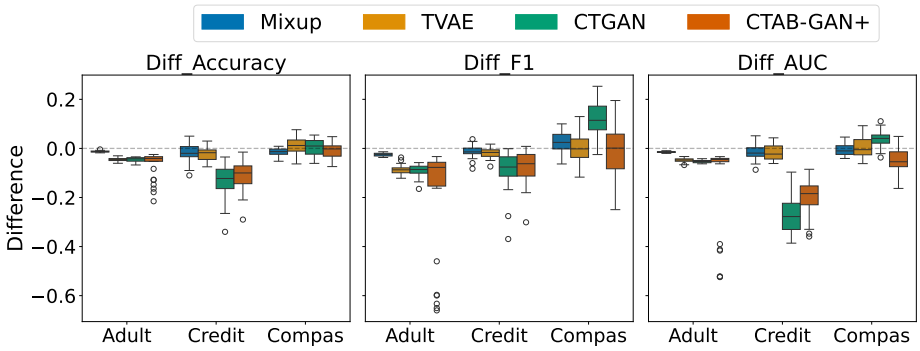

Fig. 3. Differences in utility metrics (Accuracy, F1-score, and AUC) across datasets and synthetic data generators.

Dataset complexity also affects utility stability. The Credit dataset exhibits the largest utility discrepancies across generators, indicating that its heterogeneous feature space challenges synthetic data generation. In contrast, the Adult dataset yields more consistent utility results, with smaller performance gaps, suggesting its structure is more conducive to accurate synthetic reproduction.

As outlined in Section 5.1, the DOMIAS attack was initially provided with 10% of the training data as a reference set. We then tested larger subsets to verify whether additional reference samples improved attack success. Consistent with the original findings [3], Figure 4 shows no significant improvement beyond 10%, apart from a minor increase in attack performance for Mixup on the COMPAS dataset.

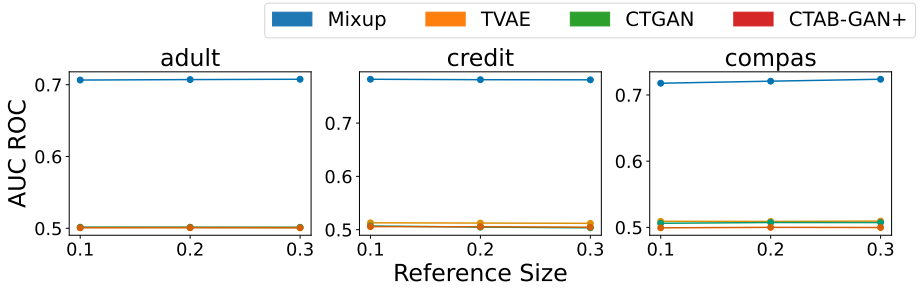

Fig. 4. DOMIAS performance as a function of the percentage of training data available, evaluated across different datasets and synthetic data generators.

To further analyze generator behavior, we examined the spatial relationship between real and synthetic samples by comparing, for each training record, the difference in distance to its closest synthetic sample versus its closest test sample.

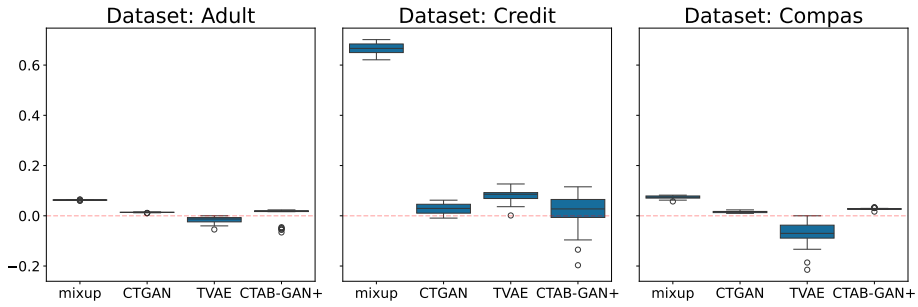

Fig. 5. Difference in distance to closest record between synthetic and test samples for each training data sample across datasets and synthetic data generators.

As shown in Figure 5, for the Credit dataset, Mixup produces synthetic samples that lie extremely close to the training data, confirming excessive proximity and potential leakage. In contrast, for COMPAS and partially for Adult, TVAE generates samples that remain consistently distant from the training distribution, indicating only partial coverage of the data space and a loss of information. This behavior reflects a known limitation of VAEs on tabular data: their tendency toward over-pruning (posterior collapse) [40]. The regularization term in the ELBO loss can dominate training, forcing the latent distribution to approximate a fixed prior (typically a unit Gaussian) and reducing dependence on the input data [16]. Consequently, TVAE maintains reasonable predictive utility but omits underrepresented data regions, reducing the fidelity and completeness of the generated samples.

Overall, CTGAN and CTAB-GAN+ achieve the best balance between utility and privacy, making them suitable for privacy-sensitive scenarios where minor utility degradation is acceptable. Mixup provides high utility but poor privacy protection, while TVAE, despite competitive performance, fails to fully reconstruct the training data distribution in COMPAS. Among the attacks, Monte Carlo and DOMIAS, though recently proposed, did not outperform traditional distance-based MIAs. In most cases, the distance-based attack remained the most effective, suggesting that direct overfitting exploitation continues to be the dominant threat to tabular synthetic data privacy.

> CTGAN and CTAB-GAN+ emerge as promising candidates for applications requiring both acceptable predictive utility and protection against diverse black-box MIAs. They demonstrate a balanced trade-off between utility and privacy, offering practical solutions for generating synthetic data that can be shared without compromising the confidentiality of personal information

6.2 Reliability of Privacy Metrics Under Black-Box MIAs (RQ2)

To answer RQ2, we assess whether the selected privacy metrics (DCR, NNDR, and Privacy Loss) correlate with the success of black-box MIAs, thereby evaluat-

ing their validity as indicators of privacy risk. We hypothesize that higher metric values should correspond to stronger MIA performance. Since each dataset split produces one value per metric and multiple AUC-ROC scores—one for each attack—we compute the average AUC-ROC across all attacks. This aggregation provides a holistic view of the relationship between privacy metrics and attack effectiveness, avoiding overemphasis on any single attack variant.

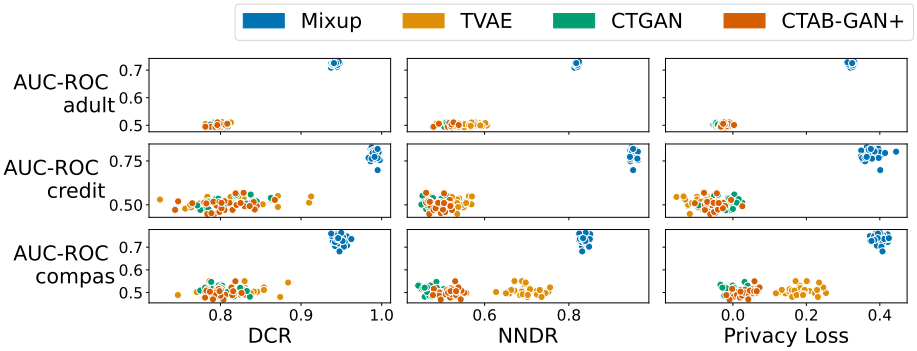

Fig. 6. Comparison of privacy metric values and MIA AUC-ROC scores across different datasets.

Figure 6 plots the metric values against the averaged AUC-ROC scores for all datasets. A general positive trend emerges: higher DCR, NNDR, and Privacy Loss values tend to coincide with stronger MIA performance, suggesting that these metrics partially capture underlying vulnerability patterns.

However, TVAE exhibits anomalous behavior on the COMPAS dataset, yielding comparatively high NNDR and Privacy Loss values while maintaining low AUC-ROC scores. This suggests that proximity-based metrics may not reliably quantify privacy risk under conditions such as latent-space underfitting, since they capture only geometric closeness without considering how an attacker could exploit this proximity to infer membership. In fact, the underfitting phenomenon discussed in Section 6.1—posterior collapse—led TVAE to generate synthetic samples that covered only a subset of the original feature space. This restricted coverage likely triggered proximity-based metrics to signal higher risk, even though it did not translate into stronger attack performance as reflected by AUC-ROC. By contrast, Mixup consistently records higher privacy metric values, aligned with its poor resistance to MIAs, whereas CTGAN and CTAB-GAN+ display lower metric outputs and reduced attack success. This alignment supports the expected relationship between metric magnitude and empirical vulnerability

Dataset-specific trends also emerge. In both the COMPAS and Credit datasets, the privacy metrics exhibit a wide spread of values across the metric axis, yet the corresponding AUC-ROC scores remain relatively stable, suggesting a re-

duced effectiveness in detecting privacy leakage. This pattern suggests that in datasets with fewer samples, proximity-based metrics may fluctuate more due to local density variations, without reflecting a proportional change in actual attack success.

Overall, the findings demonstrate that while DCR, NNDR, and Privacy Loss capture broad vulnerability trends, they fail to fully characterize leakage in models with uneven data coverage or latent-space collapse. Complementing these metrics with distributional or adversarially informed measures would improve the reliability and comprehensiveness of privacy validation frameworks for synthetic data.

> The analysis indicates that DCR, NNDR, and Privacy Loss provide a partial but incomplete picture of privacy robustness under black-box MIAs. While these metrics capture general vulnerability trends, their limitations—exemplified by the anomalous behavior of TVAE—highlight that proximity-based measures alone cannot fully represent the multifaceted nature of privacy leakage. This finding suggests the need for more comprehensive, empirically validated metrics that account for model architecture, data coverage, and attack strategy diversity.

7 Threats to validity

In this section we discuss the potential threats to validity that may influence the interpretation, reliability, or generalizability of our findings.

Internal validity: A potential threat lies in uncontrolled factors that might influence the outcomes, such as the choice of datasets or random initialization of models. To overcome this, we employed multiple, diverse datasets and repeated experiments across 30 random train/test splits to reduce bias. To ensure consistency and replicability of results, we have made the codebase fully replicable, allowing others to verify the experimental setup and results [1]. This reduces the likelihood of errors or biases affecting the results.

External validity: Although the datasets are diverse in structure and domain, they all represent socio-economic decision-making datasets and may not fully capture the variability of real-world applications. In addition, our study focuses on three widely-used deep tabular data generators, and the observed trends may not directly transfer to other synthesis approaches. Similarly, while the considered attacks and privacy metrics represent the current state of the art, they might not encompass all possible adversarial strategies or future privacy assessment techniques. Consequently, the generalizability of the results to unseen domains or emerging attack paradigms may be limited.

Construct validity: Our evaluation primarily relies on AUC-ROC, chosen for its widespread use in MIA literature and comparability with prior work, though it may not capture nuances such as class imbalance. Utility is evaluated using a single downstream model (XGBoost), reflecting a task-specific notion of utility; therefore, the observed privacy–utility trade-offs may differ for other downstream models. Conversely, privacy is assessed using record-level metrics and

MIAs, which do not explicitly account for domain-specific harm from disclosure. For RQ2, we average AUC-ROC across attacks to obtain an attack-agnostic view, potentially obscuring differences among attack types. Finally, we do not consider group-level disclosure risks, which may persist even when record-level membership inference is limited.

Conclusion Validity: The statistical significance of the observed differences among attacks and generators could be influenced by the limited number of experimental runs. To mitigate random variance and strengthen the robustness of conclusions, we conducted multiple independent repetitions and reported aggregate statistics across all runs.

8 Conclusion

This work establishes an empirical framework for quantifying privacy risks in synthetic tabular data, providing a systematic comparison of black-box membership inference attacks and privacy metrics, grounded in fairness-related tabular benchmarks and deep tabular generators, to better understand the privacy robustness of generative models and how effectively privacy metrics capture privacy leakage.

Among the evaluated generators, CTAB-GAN+ and CTGAN demonstrated the best balance between privacy and utility, achieving low AUC-ROC values across diverse MIAs while maintaining predictive performance. Conversely, Mixup exhibits the highest susceptibility to MIAs, making it unsuitable for privacy-sensitive applications, while TVAE offers strong resistance to MIA at the cost of loss of information.

Our evaluation of the privacy metrics, DCR, NNDR, and Privacy Loss, reveals that while they correlate with attack effectiveness, their reliance on proximity-based assumptions limits their sensitivity to distributional coverage. Anomalies such as TVAE's inflated privacy scores emphasize that current metrics do not fully capture attacker capabilities or structural gaps in the generated data. These findings underscore the need for improved measures that integrate both data coverage and exploitability perspectives. Finally, our study challenges the assumed superiority of newer MIAs, such as Monte Carlo-based attacks and DOMIAS, which failed to outperform traditional distance-based attacks in our experiment. Future work should focus on defining privacy metrics that incorporate data coverage and attacker modeling, extending this framework to other data modalities and attack families, and analyzing the conditions under which newer attacks may generalize more effectively.

Ultimately, this study contributes to the empirical validation of privacy properties in generative models—an essential step toward trustworthy and testable AI systems within the broader context of software engineering research.

Acknowledgments. This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU

(PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This publication is part of the project PNRR-NGEU which has received funding from the MUR – DM 630/2024. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. South German Credit. UCI Machine Learning Repository (2020), DOI: https://doi.org/10.24432/C5QG88
2. Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996), DOI: https://doi.org/10.24432/C5XW20
3. van Breugel, B., Hao, S., Zhaozhi, Q., van der Schaar, M.: Membership inference attacks against synthetic data through overfitting detection (2023), https://arxiv.org/abs/2302.12580
4. van Breugel, B., Kyono, T., Berrevoets, J., van der Schaar, M.: Decaf: Generating fair synthetic data using causally-aware generative networks (2021), https://arxiv.org/abs/2110.12884
5. Chakraborty, T., S, U.R.K., Naik, S.M., Panja, M., Manvitha, B.: Ten years of generative adversarial nets (gans): A survey of the state-of-the-art (2023), https://arxiv.org/abs/2308.16316
6. Chen, D., Yu, N., Zhang, Y., Fritz, M.: Gan-leaks: A taxonomy of membership inference attacks against generative models. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. pp. 343–362. ACM, Virtual Event, USA (Oct 2020). https://doi.org/10.1145/3372297.3417238, https://dl.acm.org/doi/10.1145/3372297.3417238
7. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 785–794. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785, https://doi.org/10.1145/2939672.2939785
8. Dankar, F.K., Ibrahim, M.K., Ismail, L.: A multi-dimensional evaluation of synthetic data generators. IEEE Access **10**, 11147–11158 (2022). https://doi.org/10.1109/ACCESS.2022.3144765, https://ieeexplore.ieee.org/document/9686689/
9. Fabris, A., Messina, S., Silvello, G., Susto, G.A.: Algorithmic fairness datasets: The story so far. Data Mining and Knowledge Discovery **36**, 2074–2152 (2022). https://doi.org/10.1007/s10618-022-00854-z, https://doi.org/10.1007/s10618-022-00854-z
10. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters **27**(8), 861–874 (2006). https://doi.org/10.1016/j.patrec.2005.10.010, https://www.sciencedirect.com/science/article/pii/S016786550500303X
11. Figueira, A., Vaz, B.: Survey on synthetic data generation, evaluation methods and gans. Mathematics **10**(15), 2733 (2022). https://doi.org/10.3390/math10152733, https://www.mdpi.com/2227-7390/10/15/2733

12. Fonseca, J., Bacao, F.: Tabular and latent space synthetic data generation: A literature review. Journal of Big Data **10**(1), 115 (Jul 2023). https://doi.org/10.1186/s40537-023-00792-7, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00792-7
13. Ganev, G.: Synthetic data, similarity-based privacy metrics, and regulatory (non-)compliance (2024), https://arxiv.org/abs/2407.16929
14. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. p. 2672–2680. NIPS'14, MIT Press, Cambridge, MA, USA (2014)
15. Gundersen, O.E., Gil, Y., Aha, D.W.: On reproducible ai: Towards reproducible research, open science, and digital scholarship in ai publications. AI Magazine **39**(3), 56–68 (Sep 2018). https://doi.org/10.1609/aimag.v39i3.2816, https://onlinelibrary.wiley.com/doi/10.1609/aimag.v39i3.2816
16. Guo, C., Zhou, J., Chen, H., Ying, N., Zhang, J., Zhou, D.: Variational autoencoder with optimizing gaussian mixture model priors. IEEE Access **8**, 43992–44005 (2020). https://doi.org/10.1109/ACCESS.2020.2977671
17. Hayes, J., Melis, L., Danezis, G., Cristofaro, E.D.: Logan: Membership inference attacks against generative models (2018), https://arxiv.org/abs/1705.07663
18. 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). https://doi.org/10.1016/j.neucom.2022.04.053, https://linkinghub.elsevier.com/retrieve/pii/S0925231222004349
19. Hilprecht, B., Härterich, M., Bernau, D.: Reconstruction and membership inference attacks against generative models (2019), https://arxiv.org/abs/1906.03006
20. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: A survey. ACM Computing Surveys **54**(11s), 1–37 (Jan 2022). https://doi.org/10.1145/3523273, https://dl.acm.org/doi/10.1145/3523273
21. Hyeong, J., Kim, J., Park, N., Jajodia, S.: An empirical study on the membership inference attack against tabular data synthesis models. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 4064–4068. ACM, A (Oct 2022)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2013), https://arxiv.org/abs/1312.6114
23. Latanya, S.: Simple demographics often identify people uniquely (2000)
24. Liu, Q., Khalil, M., Jovanovic, J., Shakya, R.: Scaling while privacy preserving: A comprehensive synthetic tabular data generation and evaluation in learning analytics. In: Proceedings of the 14th Learning Analytics and Knowledge Conference. pp. 620–631. ACM, Kyoto, Japan (Mar 2024). https://doi.org/10.1145/3636555.3636921, https://dl.acm.org/doi/10.1145/3636555.3636921
25. Liu, Y., Huang, J., Li, Y., Wang, D., Xiao, B.: Generative ai model privacy: A survey. Artificial Intelligence Review **58**(1), 33 (Dec 2024). https://doi.org/10.1007/s10462-024-11024-6, https://link.springer.com/10.1007/s10462-024-11024-6
26. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**, 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94, https://doi.org/10.1023/B:VISI.0000029664.99615.94

27. Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., Wei, W.: Machine learning for synthetic data generation: A review (2024), https://arxiv.org/abs/2302.04062
28. Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., Malossi, C.: Bagan: Data augmentation with balancing gan (2018), https://arxiv.org/abs/1803.09655
29. Mateo-Sanz, J.M., Sebé, F., Domingo-Ferrer, J.: Outlier protection in continuous microdata masking. In: Privacy in Statistical Databases. pp. 201–215. Springer Berlin Heidelberg, Berlin, Heidelberg (2004)
30. Niu, B., Sun, J., Chen, Y., Zhang, L., Cao, J., Geng, K., Li, F.: Evaluating the impact of adversarial factors on membership inference attacks. In: 2023 IEEE Smart World Congress (SWC). pp. 1–8. IEEE (2023). https://doi.org/10.1109/SWC57546.2023.10448806, https://ieeexplore.ieee.org/document/10448806/
31. Niu, J., Liu, P., Zhu, X., Shen, K., Wang, Y., Chi, H., Shen, Y., Jiang, X., Ma, J., Zhang, Y.: A survey on membership inference attacks and defenses in machine learning. Journal of Information and Intelligence **2**(5), 404–454 (Sep 2024). https://doi.org/10.1016/j.jiixd.2024.02.001, https://linkinghub.elsevier.com/retrieve/pii/S2949715924000064
32. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., Kim, Y.: Data synthesis based on generative adversarial networks. Proceedings of the VLDB Endowment **11**(10), 1071–1083 (2018). https://doi.org/10.14778/3231751.3231757, https://dl.acm.org/doi/10.14778/3231751.3231757
33. Platzer, M., Reutterer, T.: Holdout-based empirical assessment of mixed-type synthetic data. Frontiers in Big Data **4** (2021). https://doi.org/10.3389/fdata.2021.679939
34. ProPublica: Compas analysis (2016), https://github.com/propublica/compas-analysis
35. Raghavan, M., Barocas, S., Kleinberg, J., Levy, K.: Mitigating bias in algorithmic hiring: evaluating claims and practices. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 469–481. FAT* '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3372828, https://doi.org/10.1145/3351095.3372828
36. Razavi, A., van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
37. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 3–18. IEEE (2017). https://doi.org/10.1109/SP.2017.41
38. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data – anonymisation groundhog day. In: 31st USENIX Security Symposium (USENIX Security 22). pp. 1451–1468. USENIX Association (2022), https://www.usenix.org/conference/usenixsecurity22/presentation/stadler
39. Steier, A., Ramaswamy, L., Manoel, A., Haushalter, A.: Synthetic data privacy metrics (2025), https://arxiv.org/abs/2501.03941
40. Tazwar, S.M., Knobbout, M., Quesada, E.H., Popa, M.: Tab-vae: A novel VAE for generating synthetic tabular data. In: Proceedings of the 13th International Conference on Pattern Recognition Applications and Method. SCITEPRESS, Maastricht, The Netherlands (2024)

41. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN, pp. 659–669. Curran Associates Inc., Red Hook, NY, USA (2019)
42. Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., Bennett, K.: Generation and evaluation of privacy preserving synthetic health data. Neurocomputing **416** (04 2020). https://doi.org/10.1016/j.neucom.2019.12.136
43. Zhao, Z., Kunar, A., Birke, R., Chen, L.Y.: CTAB-GAN: Effective table data synthesizing. In: Balasubramanian, V.N., Tsang, I. (eds.) Proceedings of The 13th Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 157, pp. 97–112. PMLR (Nov 2021)
44. Zhao, Z., Kunar, A., Birke, R., Van Der Scheer, H., Chen, L.Y.: CTAB-GAN+: Enhancing tabular data synthesis **6**, 1296508. https://doi.org/10.3389/fdata.2023.1296508, https://www.frontiersin.org/articles/10.3389/fdata.2023.1296508/full
45. Zhu, C., Tang, J., Brouwer, H., Pérez, J.F., van Dijk, M., Chen, L.Y.: Quantifying and mitigating privacy risks for tabular generative models (2024), https://arxiv.org/abs/2403.07842