Privacy Potions for Production: Controlling Leakage in AI-Synthesized Data

White Paper

ACCENTURE, EY, HITACHI RAIL, UNILEVER, and VECTOR'S PETS TEAMS*

Executive Summary. We evaluate the privacy risks of AI-generated synthetic tabular data by analyzing the success of membership inference attacks (MIAs) under varying configurations and attacker profiles. Building on the MIDST challenge, which revealed vulnerabilities in state-of-the-art diffusion models, the experiments identify key factors influencing privacy leakage and provide actionable guidance for practitioners. The findings show that training setup and hyperparameter choices have a significant impact. For example, larger training datasets improve both privacy and utility, making data collection a more effective lever than oversynthesizing. In contrast, increasing model size or generating excessive synthetic data can amplify privacy leakage without meaningful gains in quality. Importantly, we demonstrate that MIAs remain effective even when attackers have imperfect knowledge of the data distribution or model parameters, providing a direct, regulator-aligned measure of privacy risk. We further show that commonly used proxies, such as distance to closest record (DCR), fail to reliably capture privacy risk. Collectively, these results underscore that: i) better privacy does not always imply worse utility, ii) smaller or moderately sized models are often more privacy-efficient, and iii) MIAs constitute an operationally meaningful assessment of privacy leakage. These insights provide organizations with concrete guidance for establishing defensible privacy thresholds in synthetic data deployment, even in the absence of industry-wide frameworks.

1 Introduction and Background

The rapid advancements in machine learning (ML) have created an unprecedented demand for large-scale datasets. However, in many settings, accessing large, real-world datasets remains challenging. Privacy regulations such as GDPR in Europe [6], HIPAA and CCPA [2] in the United States [21], and PIPEDA in Canada [15] impose strict limitations on how sensitive information may be collected, used, and shared. Beyond regulatory constraints, organizations face practical barriers including data scarcity, imbalanced datasets, and commercial sensitivities that prevent data sharing, even when it would be mutually beneficial. Synthetic data generation has emerged as a promising solution to this dilemma. The fundamental idea is to train a generative model, which learns the statistical patterns present in real data, then use that model to create entirely new records that never correspond to actual individuals. In theory, this synthetic dataset preserves the utility of the original data, maintaining its statistical properties, correlations, and predictive value, while eliminating the privacy risks associated with real records. However, while synthetic data offers significant advantages, it does not inherently guarantee privacy preservation in its classical form [20]. Recent research has revealed that generative models can leak information about the specific records used during their training process [24]. This mainly occurs because a model has inadvertently memorized aspects of its training data. The challenge lies in distinguishing between learning general patterns, which is desirable, and encoding specific training data points, which is a privacy risk.

1.1 Membership Inference Attacks

Membership Inference Attacks (MIAs) have emerged as a state-of-the-art approach for quantifying privacy risks [3, 7, 12, 18, 19, 26]. In an MIA, an adversary attempts to determine whether a particular individual's data was included in the training set used to create a synthetic dataset [8, 19]. For example, an attacker with access to synthetic financial transactions generated by a bank analyzes whether your actual transaction history was part of the training data. A successful attack reveals you as a customer of that bank. Such information could enable targeted phishing, identity theft, or financial fraud.

1.2 The MIDST Challenge

To rigorously evaluate the privacy resilience of synthetic tabular data generation, the Vector Institute organized the Membership Inference over Diffusion-models-based Synthetic Tabular data (MIDST) challenge at IEEE SaTML 2025. While diffusion models have demonstrated remarkable success in generating high-quality synthetic tabular data [9, 16, 29] among generative AI models, their vulnerability to privacy attacks over tabular data through systematic adversarial testing remained largely unexplored prior to the competition. The challenge was structured around the Czech Bank Berka [23] dataset, containing over one million financial transactions. For each of the 20 trained diffusion models, participants were presented with 200 data points, 100 that were part of the model's training data and 100 that were not. The goal was to predict which was which. If attackers could do better than random guessing (which would yield 10% true positive at a 10% false positive rate), it would indicate measurable privacy leakage. The competition included four tracks representing different threat scenarios. These included white-box settings where adversaries had access to model parameters and black-box settings where they could only examine synthetic outputs, and both single-table and multi-table generation models. Over three months, 71 participants submitted more than 700 attack strategies, revealing significant privacy vulnerabilities across all tracks. The winning white-box attack achieved a 46% success rate for the white-box single-table scenario, more than four times better than random guessing. Black-box attacks, which only had access to synthetic outputs, reached

*Team members' information provided in the document, Section 5

25% success rates. These findings demonstrated that state-of-the-art diffusion models, despite their sophisticated architectures, do leak information about their training data in ways that skilled adversaries can exploit. These results align with recent research showing that MIAs represent a persistent vulnerability in tabular synthetic data generation [22, 24].

1.3 From Competition to Research

The MIDST results raise critical questions for organizations considering the use of synthetic data. What models leak more information than others? What factors, such as dataset size, model architecture/size, training duration, or data characteristics, influence vulnerability to membership inference? This white paper builds on the foundations of the MIDST Challenge to provide systematic answers to these questions. The privacy risks of AI-generated synthetic data are evaluated by analyzing the success of MIAs under a variety of configurations and attacker profiles. Through extensive experiments manipulating factors such as training configurations, model capacity, size of synthetic data, and attacker knowledge, key drivers of privacy leakage are identified. The objective is to move beyond demonstrating that vulnerabilities exist and provide actionable guidance, helping organizations understand when synthetic data provides sufficient protection for their use case and which mitigation strategies are most effective.

2 Highlights of Experimental Findings

We ran an extensive set of experiment for both white-box and black-box settings, to investigate the influence of various factors on the success of the Tartan Federer MIA [25], which won the MIDST Challenge in all scenarios. As we provide a summary in the following, we report the full set of results in the appendix. In MIAs, the attacker has access to, or can train, a set of shadow models for developing attacks. These models are *similar* to the target model, simulated by the attacker. In many cases, such models are architecturally identical to the target model, but are not trained on the same data. In addition to MIA success rate, as measured by TPR@FPR=10%[1], distance to closest record (DCR) is also computed [11]. DCR is commonly used as a rule-of-thumb privacy metric in the synthetic data generation community. While DCR is easy to compute and relied upon in many settings [14, 28], existing work [27] and the experiments herein call into question its utility as a useful approximation of privacy in many scenarios.

2.1 Hyperparameter Choices

We investigate the effect of the number of diffusion (time) steps, the number of training steps, batch size, and model size. On one hand, as shown in Figure 1, increasing the number of diffusion steps adversely affects privacy in MIA success. Notably, DCR fails to capture this effect on privacy. On the other hand, increasing the diffusion steps slightly improves the synthetic data quality, to a point, as indicated in Table 1, pointing out the trade off between utility and privacy in selecting the parameter. A similar trend holds for increasing training steps. Increasing the batch size, increases MIA success without corresponding improvements in synthetic data quality. Varying the model size significantly affects MIA success, but it does not exhibit a clear relationship with quality metrics. Larger models are much more vulnerable to privacy leakage than smaller models. While this relationship is clear for MIA, DCR fails to reflect this pattern. Detailed results are presented in the appendix

2.2 Training and Synthetic Dataset Size

Increasing the training dataset size improves privacy significantly, as shown in Figure 1, and enhances synthetic data quality at the same time; see Table 1. However, increasing the input (training) to output (generated data) ratio from 1:1 to 1:10 and higher values, increases MIA success and degrades privacy. This effect is more visible for smaller training sets. Note that while DCR follows the MIA trend in the former experiment, it fails to capture the privacy differences in the latter. Industry participants also validated a subset of these results, available in the appendix, on other datasets, namely Madrid [4] and German Credit [17].

2.3 Attacker Knowledge of the Sensitive Data Distribution

In these experiments, four scenarios are investigated with the Berka dataset: i) Disjoint user accounts: the adversary and the target model trainer sample their data from entirely disjoint sets of accounts, ii) Disjoint time windows: the adversary observes data from a later time period, while the target model is trained on earlier records, iii) Noisy adversary data: half of the categorical and numerical features available to the attacker are replaced with samples from the uniform distribution of their respective ranges, and iv) Statistics-based synthesis: the adversary knows the distribution of each column, but the correlation between columns is completely distorted through independent sampling. Experimental results show that the attack success in all four scenarios changes only marginally in the black-box setting. In the white-box setting the maximum reduction in the attack success is to 41%, from the original of 46%, still yielding a powerful attack. This implies that the auxiliary data available to adversaries need not be identically distributed to the training data to produce highly successful attacks.

[1]True positive rate at 10% false positive rate

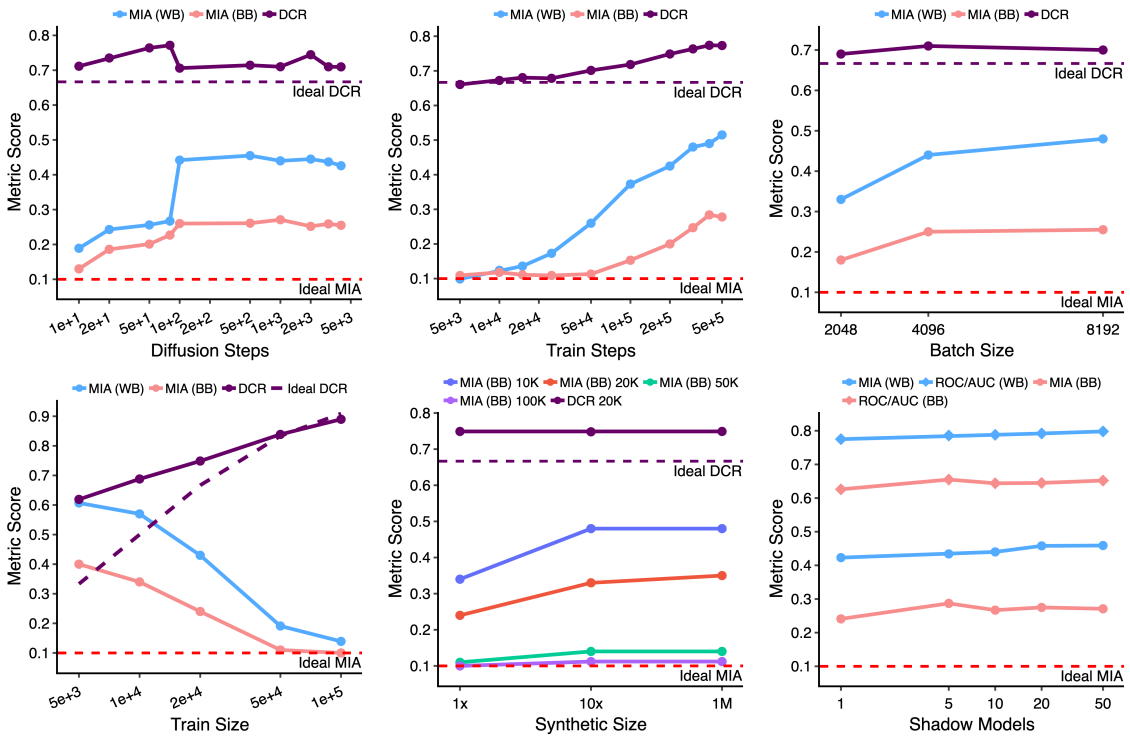

Fig. 1. Summary of the various levers that influence MIA success in white- and black-box settings, as well as DCR measures.

2.4 Attacker Knowledge of the Target Model Hyperparameters

In this set of experiments the shadow models used by the attack differ in significant ways from the target model being attacked. Shadow models differing in model size, the number of diffusion time steps, and the number of training steps are considered. Note that these experiments are limited to black-box scenarios where the target model is not provided to the adversary. The results indicate the effect of these discrepancies on MIA success varies. For example, large reductions in the diffusion time steps or training iterations significantly impact MIA success. While attacks are sensitive in certain respects, adversaries do not necessarily need perfect knowledge of target model hyperparameters to succeed.

Lever	MIA (WB) Success	MIA (BB) Success	DCR Impact	Quality
More Diffusion Steps	↑↑	↑	–	↑*
More Training Steps	↑↑	↑↑	↑	↑*
Larger Batch Size	↑	–	↑	–
More Training Data	↓↓	↓↓	↓↓	↑
More Synthetic Data	N/A	↑↑	–	↑
Data Misalignment	↓	–	N/A	N/A
Larger Models	↑↑	↑↑	–	–
More Shadow Models	–	–	N/A	N/A

Table 1. * indicates that quality increases only to a point, degrading thereafter. – indicates the metric is not impacted by changes to the lever.

2.5 Attacker Computing Power

It is commonly assumed that the more shadow models an adversary has access to, or the resources to train, the better their chances of producing successful MIAs. Surprisingly, however, the results show that increasing the the number of shadow models used for the attack development does not markedly affect MIA success, beyond a small handful of models in both the black- and white-box settings. In other words, the attackers do not necessarily need significant computing power to successfully apply MIAs.

3 Takeaways for Industry

Our industry collaborators highlight several practical insights following the experimental results review, reflecting that common intuitions about privacy, utility, and model scaling do not always hold in practice and that careful design choices are essential. To support practitioners in applying these insights, the Vector Institute has developed the MIDST Toolkit (https://github.com/vectorInstitute/midst-toolkit). The toolkit enables organizations to generate synthetic data, assess its quality through multiple metrics, and test its resilience to membership inference attacks.

3.1 Pick Your Utility Metrics: Better Privacy Does Not Always Mean Worse Utility

Privacy-utility trade-offs are highly dependent on how utility is measured. While certain hyperparameters, such as model size, increase MIA success, they often improve only specific quality metrics and only to a point. In the experiments, we examined three utility categories: i) general fidelity and diversity measures, ii) statistical similarity to real data in single-column and column-pairs, and iii) downstream classification/regression tasks. In several instances, meaningful reductions in privacy leakage can be achieved with negligible or no degradation in a certain utility category. This underscores the importance of selecting task-relevant and deployment-driven utility metrics when tuning synthetic data generators.

3.2 Larger Models Are Not Necessarily Better

Contrary to a common industry assumption, increasing model capacity worsens privacy without providing clear gains in synthetic data quality. Larger diffusion models are substantially more vulnerable to both white-box and black-box MIAs, while quality metrics remain largely unchanged. These results suggest that over-parameterization amplifies memorization risks without corresponding benefits, making smaller or moderately sized models a more privacy-efficient choice for tabular data synthesis.

3.3 Train on More, Synthesize Less

Increasing the size of the training dataset simultaneously improves privacy and utility, making it one of the most effective levers available to practitioners. In contrast, aggressively increasing the amount of synthetic data generated per real record significantly increases privacy leakage, particularly for smaller training sets. From a deployment perspective, this suggests that organizations should prioritize collecting more real data when feasible, rather than compensating for limited training data by oversynthesizing.

3.4 DCR Is a Very Limited Privacy Metric

DCR was originally developed as a model quality diagnostic to detect issues such as memorization in synthetic datasets. While widely used due to its simplicity, DCR fails to capture many of the privacy risks revealed by MIAs. This is easily seen in Figure 1. As such, despite its intuitive appeal as an indicator of closeness between real and synthetic records, the metric does not constitute a strong measure of privacy risk and cannot be relied upon as a standalone metric, especially in high-stakes or regulated settings.

3.5 MIAs as a Measure of Privacy Leakage and Regulations

MIAs emulate realistic adversarial behavior where attackers actively probe or exploit model outputs to gain a statistical advantage over random guessing in recognizing training members from non-members. They are robust across a wide range of adversarial assumptions, including mismatched data distributions and imperfect knowledge of model hyperparameters. Moreover, attackers do not require substantial computational resources or extensive shadow model training to mount effective attacks. Therefore, MIAs offer a quantifiable and replicable measure of privacy risk that aligns with regulatory definitions of re-identification and disclosure. MIAs can align closely with formal definitions of privacy harm, including singling-out, linkability, and inference risk in GDPR, NIST's definitions of re-identification risk (NIST SP-800-188), and OECD's privacy harm taxonomy.

4 Conclusions

Synthetic data governance currently lacks standardized frameworks to map privacy metrics to legal compliance, define minimum utility thresholds, or establish explicit tradeoffs between them, unlike cybersecurity (NIST), or risk management (ISO) standards. This gap forces organizations to improvise, often focusing on privacy without assessing utility, or optimizing utility without determining acceptable privacy risks. Our work addresses these challenges by investigating empirically grounded privacy metrics via membership inference attacks (MIAs), which offer a defensible, quantifiable basis for measuring privacy leakage. Complementing this, we present multiple quantifiable utility categories, capturing statistical fidelity, structural fidelity, and downstream task performance, enabling organizations to compare synthetic data methods and tailor them to diverse business use cases.

While MIAs provide the a direct, attack-aligned measure of privacy leakage, complementary privacy techniques such as differential privacy (DP), when applicable, can further limit adversarial advantage. DP remains the theoretical gold standard for privacy, yet utility is often heavily affected when DP is used in diffusion model training [5]. This work primarily investigated

privacy-utility tradeoffs in the non-DP setting. Our industry collaborators highlight the need for additional attention on the dimension of fairness through studying existing literature on optimizing these three aspects and providing preliminary results on their interaction with synthetic data. The privacy–utility–fairness tradeoff is multidimensional and nonlinear in nature, hence human intuition cannot reliably navigate it. Automated multi-metric optimization explores a variety of parameter combinations, and identifies viable privacy/utility configurations to illuminate Pareto optimal boundaries allowing organizations to optimally prioritize based on their particular goals and risk appetite. Work towards providing an automated optimization solution for these three features in synthetic data generation is the subject of ongoing work.

5 Participating Teams

- Accenture: Juan-Carlos Castañeda, Declan McClure, and Karthik Venkataraman.
- EY: Rasoul Shahsavarifar, Jean-Luc Rukundo, N'Golo Kone, Paulina Nouwou, and Yasmin Mokaberi.
- Hitachi Rail: Théo Pinardin and Safiya Kamal.
- Unilever: Colm Cleary and Marta Mischi.
- Vector Institute: Michael Joseph (Project Manager), Xi He (Faculty Advisor), Masoumeh Shafieinejad & David Emerson (Technical Leads), Behnoosh Zamanlooy, Elaheh Bassak, Fatemeh Tavakoli, Sara Kodeiri, and Marcelo Lotif (Research & Engineering).

References

[1] Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. 2022. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models. In Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162), Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 290–306. https://proceedings.mlr.press/v162/alaa22a.html

[2] California State Legislature. 2018. California Consumer Privacy Act of 2018. Cal. Civ. Code § 1798.100–1798.199 (2018). https://oag.ca.gov/privacy/ccpa Accessed: 2025-01-15; amended by CPRA.

[3] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. 2022. Membership Inference Attacks From First Principles. In 2022 IEEE Symposium on Security and Privacy (SP). 1897–1914. https://doi.org/10.1109/SP46214.2022.9833649

[4] DataGuapa. 2018. Madrid Public Transportation Data 2018. https://www.kaggle.com/datasets/dataguapa/madrid-public-transportation-data-2018. Accessed: 2025-12-10.

[5] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. 2023. Differentially Private Diffusion Models. Transactions on Machine Learning Research (2023). TMLR 2023.

[6] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. OJ L 119, 4.5.2016, p. 1–88 (2016). https://data.europa.eu/eli/reg/2016/679/oj

[7] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. LOGAN: Membership Inference Attacks Against Generative Models. Proceedings on Privacy Enhancing Technologies 2019, 1 (2019), 133–152.

[8] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership Inference Attacks on Machine Learning: A Survey. ACM Comput. Surv. 54, 11s, Article 235 (Sept. 2022), 37 pages. https://doi.org/10.1145/3523273

[9] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. 2020. Tabddpm: Modelling tabular data with diffusion models, In International Conference on Machine Learning. 2023 IEEE 36th Computer Security Foundations Symposium (CSF), 473–488. https://api.semanticscholar.org/CorpusID:245353931

[10] Hadrien Lautraite, Lorrie Herbault, , Yue Qi, Jean-François Rajotte, and Sébastien Gambs. 2025. Ensemble MIA: The 2nd place solution to the MIDST Black-box MIA on the single-table competition. https://github.com/CRCHUM-CITADEL/ensemble-mia. GitHub repository, accessed: 2025-12-10.

[11] Anton D. Lautrup, Tobias Hyrup, Arthur Zimek, and Peter Schneider-Kamp. 2024. Syntheval: a framework for detailed utility and privacy evaluation of tabular synthetic data. Data Mining and Knowledge Discovery 39, 1 (Dec. 2024), 6. https://doi.org/10.1007/s10618-024-01081-4

[12] Pei-Hsuan Lu, Pang-Chieh Wang, and Chia-Mu Yu. 2019. Empirical Evaluation on Synthetic Data Generation with Generative Adversarial Network. In Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics (Seoul, Republic of Korea) (WIMS2019). Association for Computing Machinery, New York, NY, USA, Article 16, 6 pages. https://doi.org/10.1145/3326467.3326474

[13] Matthieu Meeus, Lukas Wutschitz, Santiago Zanella-Béguelin, Shruti Tople, and Reza Shokri. 2025. The Canary's Echo: Auditing Privacy Risks of LLM-Generated Synthetic Text. In Proceedings of the 42nd International Conference on Machine Learning, Vol. 267. PMLR, 43557–43580. https://proceedings.mlr.press/v267/meeus25a.html

[14] Ofer Mendelevitch and Michael D. Lesh. 2021. Fidelity and Privacy of Synthetic Medical Data. arXiv:2101.08658 [cs.LG] https://arxiv.org/abs/2101.08658

[15] Government of Canada. 2000. Canada's Personal Information Protection and Electronic Documents Act (PIPEDA). https://laws-lois.justice.gc.ca/eng/acts/P-8.6/. https://laws-lois.justice.gc.ca/eng/acts/P-8.6/ Statutes of Canada 2000, c. 5.

[16] Wei Pang, Masoumeh Shafieinejad, Lucy Liu, Stephanie Hazlewood, and Xi He. 2025. ClavaDDPM: multi-relational data synthesis with cluster-guided diffusion models. In Proceedings of the 38th International Conference on Neural Information Processing Systems (Vancouver, BC, Canada) (NIPS '24). Curran Associates Inc., Red Hook, NY, USA, Article 2657, 27 pages.

[17] UCI Machine Learning Repository. 2019. German Credit Data. https://www.kaggle.com/datasets/uciml/german-credit. Accessed: 2025-12-10.

[18] Ahmad Salem, Yang Zhang, Martin Humbert, Michael Fritz, and Manuel Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In Network and Distributed Systems Security Symposium (NDSS) 2019. California, USA. https://doi.org/10.14722/ndss.2019.23119

[19] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017. IEEE Computer Society, 3–18. https://doi.org/10.1109/SP.2017.41

[20] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic Data–Anonymisation Groundhog Day. In 31st USENIX Security Symposium (USENIX Security 22). USENIX, USA, 1451–1468.

[21] U.S. Congress. 1996. Health Insurance Portability and Accountability Act of 1996 (HIPAA). Public Law 104–191, 110 Stat. 1936. https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf Enacted August 21, 1996.

[22] Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. 2023. Membership Inference Attacks against Synthetic Data through Overfitting Detection. In International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain (Proceedings of Machine Learning Research, Vol. 206), Francisco J. R. Ruiz, Jennifer G. Dy, and Jan-Willem van de Meent (Eds.). PMLR, 3493–3514. https://proceedings.mlr.press/v206/breugel23a.html

[23] Marcelo Ventura. 2020. The Berka Dataset. https://www.kaggle.com/datasets/marceloventura/the-berka-dataset. Accessed: 2025-12-10.

[24] Joshua Ward, Yuxuan Yang, Chi-Hua Wang, and Guang Cheng. 2025. Ensembling Membership Inference Attacks Against Tabular Generative Models. In Proceedings of the 2025 Workshop on Artificial Intelligence and Security (AISec '25). ACM, –. https://doi.org/10.1145/3733799.3762977

[25] Xiaoyu Wu, Yifei Pang, Terrance Liu, and Steven Wu. 2025. Winning the MIDST Challenge: New Membership Inference Attacks on Diffusion Models for Tabular Data Synthesis. arXiv preprint (2025). arXiv:2503.12008 [cs.LG] https://arxiv.org/abs/2503.12008

[26] Zexi Yao, Nataša Krčo, Georgi Ganev, and Yves-Alexandre de Montjoye. 2025. The DCR Delusion: Measuring the Privacy Risk of Synthetic Data. arXiv:2505.01524 [cs.CR] https://arxiv.org/abs/2505.01524

[27] Zexi Yao, Nataša Krčo, Georgi Ganev, and Yves-Alexandre de Montjoye. 2025. The DCR Delusion: Measuring the Privacy Risk of Synthetic Data. arXiv:2505.01524 [cs.CR] https://arxiv.org/abs/2505.01524

[28] YData. 2023. How to evaluate the re-identification risk in Synthetic Data? Technical Report.

[29] Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. 2024. Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space. In The twelfth International Conference on Learning Representations.

A Synthetic Data Evaluation Metrics

A.1 One-way Marginal Evaluation: KS and TVD

Single-column (one-way) marginal differences are used to measure how well individual feature distributions are preserved in synthetic data compared to a sample of real data. For numerical variables, the Kolmogorov-Smirnov (KS) statistic is computed, which quantifies the maximum difference between real and synthetic empirical cumulative density functions. For categorical variables, total variation distance (TVD), defined as $\text{TVD}(P, Q) = \frac{1}{2}\sum_i |P(i) - Q(i)|$, is measured. In both cases, lower values indicate closer alignment between the two distributions. We report average KS and TVD difference for all numerical and categorical columns.

A.2 Two-way Marginal Evaluation: Correlation Matrix and Mutual Information

To assess whether synthetic data preserves relational structure, mutual information (MI) is computed for all feature pairs, applying discretization for numerical variables to ensure consistent MI estimation. We compute the Frobenius norm of the MI difference matrix to quantify overall deviation. For purely numerical variables, correlation matrices are also computed and the Frobenius norm of their difference is reported. Smaller norms indicate better preservation of pairwise correlation in the synthetic data.

A.3 Fidelity and Diversity Metrics: α and β

We further examine synthetic data quality using the α-precision and β-recall framework [1], which together characterize synthetic data fidelity and diversity. The α-precision metric measures how well synthetic samples lie within the support of the real data distribution, thereby reflecting their realism and plausibility. In contrast, β-recall quantifies coverage of the real distribution's support by the synthetic data, capturing their diversity and measuring the extent of mode collapse. High scores on both metrics suggest that the synthetic dataset is simultaneously realistic and sufficiently diverse.

A.4 Downstream Utility Evaluation: Machine Learning Efficacy

As a measure of synthetic data utility preservation, machine learning efficacy (MLE) is estimated. Using both synthetic and real datasets, regression and classification models are trained to predict each numerical or categorical column, respectively. After training, all models are evaluated on a holdout dataset drawn from real data. The average performance difference of the models trained on synthetic and real data is reported. Negative values indicate that real data trains better-performing models and positive values indicate improved performance with synthetic data. A random forest regressor is used for numerical columns and the average difference in R^2 score across columns is reported. For categorical values, four distinct classification models are trained: Random Forest, Linear Regression, Multi-Layer Perceptron, and XGBoost. The difference in F_1 score is measured for each model and averaged. Thereafter, similar to the numerical columns, score differences for each categorical target column are averaged to produce the final metric. Note that for both F_1 and R^2, differences are reported as a percentage.

A.5 Distance to Closest Record Score: DCR and Delta-DCR

Several forms of DCR exist measuring similar but distinct quantities. In this work, DCR is computed using training data, the real data used to train the generative model, and holdout data, representing real data from the same distribution that was not part of model training. Given a set of generated synthetic data, DCR measures the proportion of synthetic data points whose nearest neighbor in the training set is closer than any point in the holdout set. An ideal score is equal to the ratio of the training set size to the combined size of the training and holdout sets. For example, when the two datasets, train and holdout, are the same size, an ideal score is simply 0.5. DCR values close to the ideal score are rough indicators that a model has not overfit or memorized training records. In addition to DCR, we report Delta-DCR which indicates how much the DCR score diverges from its ideal value.

A.6 Hitting Rate

Hitting rate (HR) describes a metric that measures how often real data points are "replicated" in synthetic data. Exact replication of numerical values is quite rare. Thus, for numerical variables, if a synthetic value falls within 3% of a numerical values, it is considered "replicated". For categorical values an exact match is required. Hitting rate, also often referred to as exact match rate, is the percentage of real points that meet these criteria. Lower rates may indicate better privacy protection.

B Attacks and Datasets

The results in this work are primarily derived by applying the Tartan Federer (TF) MIA attack [25], which won the MIDST challenge in both black- and white-box scenarios, to subsets of the public Berka dataset [23]. We also include limited supplementary results using the Ensemble MIA [10] and empirical metrics for generators trained on the Madrid [4] and German Credit [17] datasets.

B.1 MIDST Winner, Tartan Federer

B.1.1 White-Box. The core assumption of the attack is that a diffusion model will decode data points seen during training with higher fidelity than unseen ones. At a high level, the attack proceeds in three steps.

(1) **Shadow-model training**: Shadow models are trained to mimic the behavior of the target model. The adversary splits an auxiliary dataset, which is ideally drawn from the same distribution as the training data, into two disjoint subsets: *member* and *non-member*. The member subset is then used to train diffusion models with the same architecture and hyperparameters as the target, serving as shadow models.
(2) **Training the classifier**: For both member and non-member samples, the attacker generates features by running the reverse diffusion process with 300 different noise initializations at seven selected timesteps. At each timestep t, the attacker computes the difference between: i) the noise used to start the forward process in the diffusion model at step 0, and ii) the noise implied by reconstructing the sample by decoding it all the way back to timestep 0. This yields 7×300 features per data point. Using these features and the known membership labels, the attacker trains a binary classifier.
(3) **Attack**: With white-box access to the target model's weights, the attacker extracts features for the challenge points using the same procedure as in Step 2. These features are then fed into the trained classifier to determine membership.

B.1.2 Black-Box. This attack follows the same structure as the white-box attack, with several adjustments to account for the fact that the adversary only has access to synthetic data generated by the target model. Unlike the white-box setting, the black-box attack relies on synthetic data produced by a first shadow model, and a second shadow model is trained on this synthetic data to extract features for training the binary classifier. The final attack is then performed using the weights of a model trained solely on the synthetic data generated by the target model rather than using the target model's true weights.

B.2 MIDST Runner Up: Ensemble attack

The MIDST runner up in the black-box, single table track, trains a meta-classifier that takes the following inputs: i) all continuous features of the data (i.e., train or test data), ii) the minimum Gower distance between each data point and the synthetic dataset, iii) the mean Gower distance of each data point to the 5, 10, 20, 30, 40 and 50 nearest records in the synthetic dataset, iv) the nearest neighbor distance ratio of each data point for its neighbors in the synthetic dataset, v) the number of neighbors in the synthetic data for each data point, vi) the membership prediction from a DOMIAS model [22], and vii) the membership prediction from an RMIA model [13]. The meta-classifier outputs the probability that a challenge point was used to train the model that generated the synthetic data. Note that, similar to the TF attack, the meta-classifier is trained using shadow models, also trained by the adversary, where membership labels are known by construction.

C Results: Impact of Key Factors on MIA Success Rates

C.1 Utility and Privacy Trade-off in Hyperparameter Selection

We investigate the effects of four settings on the synthetic data utility and privacy: i) model training steps, ii) diffusion time steps, iii) model size, and iv) batch size. The results highlight tradeoffs associated with these properties, or lack thereof in certain scenarios.

C.1.1 Training steps. Table 2 shows the effect of changing the number of model training steps. The following quality metrics improve with an increasing number of training steps to a point and then degrade: correlation matrix difference, mutual information difference, average TVD, and α-precision. Average KS and β-recall trend in the opposite direction. The MLE metrics do not indicate a significant difference between the real and synthetic data across different training steps. On the other hand, the success rate of the white-box and black-box attacks solidly increases with increase in training steps. This implies that higher step counts for diffusion model training reduce privacy robustness in terms of MIAs. The DCR-delta also climbs with growth in iteration values, following the MIA trend.

C.1.2 Diffusion (time) steps. Table 3 summarize the quality metrics. β-recall achieves the best diversity at lower time steps, while average KS, average TVD, correlation matrix difference, and mutual information difference all show the best quality at around 2000 time steps. On the other hand, for privacy metrics, MIA success rates for both black-box and white-box attacks are minimal at around 10 time steps. This also means that the best privacy, in terms of MIA, is achieved when the timestep is very small. Alternatively, utility metrics, such as correlation and mutual information differences, are quite low in this timestep range. The DCR-delta privacy metric does not follow the same trend as the MIA success rates and, therefore, provides a misleading quantification of the rising privacy risks.

Metric	5K	15K	50K	100K	200K	300K	500K
MIA (TF White) (%)	9.9	13.6	26.0	37.3	42.5	48.0	51.5
MIA (TF Black) (%)	10.9	11.1	11.3	15.3	20.0	24.7	27.8
DCR	0.66	0.68	0.71	0.72	0.75	0.76	0.77
DCR-delta	0.006	0.014	0.034	0.051	0.081	0.101	0.106
β-recall	0.59	0.59	0.45	0.46	0.48	0.49	0.51
α-precision	0.931	0.933	0.994	0.996	0.993	0.995	0.994
Avg KS	7.0e-3	8.4e-3	4.0e-2	4.0e-2	3.9e-2	3.9e-2	3.8e-2
Avg TVD	1.3e-2	1.4e-2	9.3e-3	8.3e-3	5.8e-3	6.8e-3	7.2e-3
Correlation Diff.	0.05	0.04	0.03	0.03	0.04	0.02	0.04
Mutual Inf. Diff.	4.2e-2	1.2e-2	4.1e-3	5.2e-3	3.6e-3	4.2e-3	6.3e-3
MLE: $\overline{\Delta R^2}$ (%)	0.62	0.58	-0.63	-0.10	0.24	-0.25	0.01
MLE: $\overline{\Delta F_1}$ (%)	-0.25	-0.84	-0.67	-0.52	0.17	-0.01	0.01

Table 2. Quality and privacy metrics across different training steps.

Metric	10	50	100	500	2000	3000	4000
MIA (TF White) (%)	18.9	25.6	44.2	45.5	44.5	43.7	42.6
MIA (TF Black) (%)	13.0	20.1	26.0	26.1	25.2	25.9	25.5
DCR	0.71	0.76	0.71	0.71	0.74	0.71	0.71
DCR-delta	0.044	0.104	0.039	0.047	0.077	0.043	0.042
β-recall	0.52	0.61	0.49	0.49	0.48	0.50	0.50
α-precision	0.930	0.930	0.990	0.990	0.990	0.990	0.990
Avg KS	2.6e-2	8.5e-3	7.4e-3	8.0e-3	4.0e-3	7.7e-3	7.8e-3
Avg TVD	5.9e-2	1.1e-2	6.9e-3	7.5e-3	4.4e-3	5.4e-3	7.1e-3
Correlation Diff.	1.5e-1	3.2e-2	4.6e-2	3.9e-2	5.6e-2	3.4e-2	4.0e-2
Mutual Inf. Diff.	1.2e-1	1.3e-2	7.7e-3	9.1e-3	4.2e-3	6.2e-3	8.2e-3
MLE: $\overline{\Delta R^2}$ (%)	-1.48	-0.64	-0.13	-0.63	0.24	0.89	0.79
MLE: $\overline{\Delta F_1}$ (%)	-2.71	0.38	0.46	0.32	0.17	0.08	-0.39

Table 3. Quality and privacy metrics across diffusion (time) steps.

C.1.3 Model size. We vary the model size, namely the number of middle layers and the width of the neural network of the diffusion model in the following scenarios, while keeping other hyperparameters at their default values unless stated. The results are presented in Table 4.

(1) Narrow and shallow model with `d_layers = [256, 512, 512, 256]`. We decrease the number of training iterations to 100,000 because the model is lighter.
(2) Narrow and shallow model with `d_layers = [256, 512, 512, 256]`. We keep the number of training iterations at the default value of 200,000.
(3) Narrow model with `d_layers = [256, 512, 512, 512, 512, 256]`. We keep the number of training iterations at the default value of 200,000.
(4) Default case with `d_layers = [512, 1024, 1024, 1024, 1024, 512]`. We keep the number of training iterations at the default value of 200,000.
(5) Wider and deeper model with `d_layers = [1024, 2048, 2048, 2048 2048, 2048, 2048, 1024]`. We keep the number of training iterations at the default value of 200,000.

(6) Wider and deeper model with `d_layers = [1024, 2048, 2048, 2048 2048, 2048, 2048, 1024]`. We increase the number of training iterations to 300,000, since the model is heavier.

Metric	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
MIA (TF White) (%)	18.9	24.5	38.0	43.0	44.7	48.1
MIA (TF Black) (%)	12.0	13.4	17.6	25.2	25.4	26.2
DCR	0.70	0.74	0.70	0.74	0.71	0.71
DCR-delta	0.044	0.078	0.038	0.080	0.054	0.054
β-recall	0.60	0.58	0.61	0.47	0.61	0.59
α-precision	0.930	0.947	0.930	0.990	0.930	0.933
Avg KS	7.5e-03	7.1e-03	9e-03	3.9e-03	6.2e-03	8.8e-03
Avg TVD	8.6e-03	7.5e-03	7.2e-03	4.3e-03	5.9e-03	8.5e-03
Correlation Diff.	2.9e-02	2.5e-02	4.6e-02	5.7e-02	3.2e-02	3.3e-02
Mutual Inf. Diff.	1.3e-02	4.8e-03	5.9e-03	4.2e-03	1.1e-02	4.5e-03
MLE: $\overline{\Delta R^2}$ (%)	-0.89	-0.43	0.17	0.24	-0.34	0.40
MLE: $\overline{\Delta F_1}$ (%)	-0.74	1.45	0.39	0.17	0.46	0.34

Table 4. Quality and privacy metrics across different setups for the diffusion model.

Changing the default values from Setup (4) to smaller or larger models, causes a slight improvement in β-recall , but slight degradation in α-precision. Single column metrics are slightly worse for both variations compared to the default setup. In contrast to the quality metrics, there is a clear relationship between model size and MIA success in both the white-box and black-box scenarios. As shown in Table 4, larger models, or those of equivalent size trained for longer, appear to be more vulnerable to privacy leakage. While this relationship is clear for MIA, DCR fails to reflect this pattern.

C.1.4 Batch size. The experiment results are shown in Table 5. These results show that most quality metrics remain stable across varying training batch sizes. However, correlation matrix difference shows slight degradation at higher-end and lower-end batch sizes, indicating that moderate values, such as 4096, best preserve the correlation structure. Unlike quality metrics, MIA success rates for both black-box and white-box are fairly sensitive to batch size, with privacy leakage significantly minimized at smaller batch sizes. The DCR-delta metric does not follow this trend and therefore provides a naive evaluation of privacy risks. These results indicate that smaller batch sizes increase privacy without substantially degrading quality.

Metric	**2048**	**4096**	**4092**
MIA (TF White) (%)	33.1	44.5	48.4
MIA (TF Black) (%)	18.6	25.2	25.5
DCR	0.69	0.71	0.70
DCR-delta	0.035	0.056	0.049
β-recall	0.495	0.494	0.489
α-precision	0.99	0.99	0.99
Avg KS	7.4e-3	7.6e-3	8.1e-3
Avg TVD	4.9e-3	4.7e-3	6.8e-3
Correlation Diff.	4.4e-2	2.9e-2	4.0e-2
Mutual Inf. Diff.	4.3e-3	3.9e-3	6.1e-3
MLE: $\overline{\Delta R^2}$ (%)	0.23	0.24	-0.54
MLE: $\overline{\Delta F_1}$ (%)	-0.05	0.17	-0.08

Table 5. Quality and privacy metrics across batch sizes.

C.2 Size of the Training Data

The results presented in Table 6 indicate that with increase in training size, the following quality metrics improve: average KS, average TVD test, and correlation matrix difference. We also observe that the white-box and black-box attack success rates are both very high for smaller training sizes and monotonically improve with increasing training size. The Ensemble black-box attack also confirms the results, as shown in Table 7.

Metric	**5K**	**10K**	**20K**	**50K**	**100K**
MIA (TF White) (%)	60.7	57.0	43.0	19.1	13.9
MIA (TF Black) (%)	36.0	34.0	24.0	11.0	10.0
DCR	0.62	0.69	0.75	0.84	0.89
Delta DCR	0.290	0.180	0.088	0.008	0.020
β-recall	0.61	0.62	0.54	0.53	0.34
α-precision	0.93	0.930	0.990	0.930	0.996
Avg KS	0.042	0.042	0.039	0.039	0.038
Avg TVD	8.9e-3	8.6e-3	5.8e-3	5.8e-3	2.8e-3
Correlation Diff.	6.9e-2	5.0e-2	3.8e-2	1.3e-2	1.8e-2
Mutual Inf. Diff.	7.0e-3	7.0e-3	3.6e-3	5.0e-3	8.9e-3
MLE: $\overline{\Delta R^2}$ (%)	-0.039	-3.270	-0.021	-11.100	-12.084
MLE: $\overline{\Delta F_1}$ (%)	1.57	0.27	0.69	-3.28	-3.56

Table 6. Quality and Privacy metrics for train sizes: 5K to 100K.

Metric	**10K**	**20K**	**50K**
Black-Box MIA (%)	23.0	22.0	12.0

Table 7. Ensemble MIA over training dataset sizes: 10K to 50K.

C.3 Ratio of Training Data Size to Generated Dataset Size

We evaluate this by considering training dataset sizes of 10K, 20K, 50K, and 100K. For each training size, we vary the synthetic data provided to the adversary to be 1× the training dataset size, 10× the training dataset size, and a fixed size of 1M synthetic samples. The results presented in Table 8, indicate that the marginal distributions of individual columns and the correlation structure across columns are well preserved across all synthetic data sizes. The experiments indicate consistently high α-precision over different synthetic data sizes, which suggests that most synthetic samples lie within the support of the real data distribution and exhibit fidelity to the underlying real data. The quality measure that is improved by increasing the synthetic data size is β-recall. As shown in Figure 2, β-recall increases substantially as the synthetic dataset grows, reaching values close to 1 (the maximum and ideal value) when using 1M synthetic data points across all training data sizes. This indicates a strong improvement in the diversity of the generated samples. Table 9 shows the effect of changing the ratio on the machine learning efficacy. The results indicate that while regression on numerical columns degrades by increasing the synthetic data size, classification over categorical columns generally improves. These results show that not all quality metrics behave similarly, and therefore the choice of synthetic data size should depend on which properties of the synthetic data are most important for a given application.

When examining the MIA success rate in the black-box setting, we observe that increasing the amount of synthetic data available to the adversary generally leads to higher attack success rates. However, the magnitude of this improvement depends strongly on the training dataset size. Specifically, smaller training datasets exhibit larger gains in black-box MIA success, while larger training datasets show only modest improvements. This reinforces the hypothesis that when the training dataset is small and the training hyperparameters are fixed, there is a greater risk of overfitting. Generating increasingly large volumes of synthetic data with such models makes an adversary's task even easier. This effect is clearly visible when comparing the results for the 20k and 50k

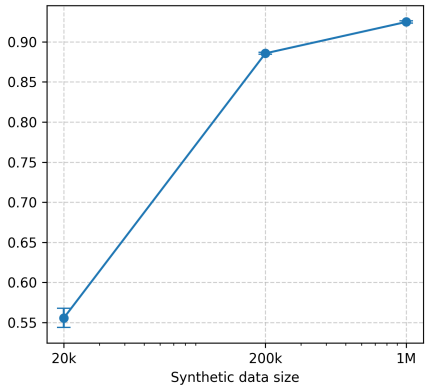

Fig. 2. β-recall improving with increasing output size.

Training Size	Synthetic 0.1×	Synthetic 1×	Synthetic 10×	MLE
10k	-4.75	-4.43	-9.97	$\overline{\Delta R^2}$ (%)
	-1.42	-0.76	0.67	$\overline{\Delta F_1}$ (%)
20k	-4.32	-2.89	-8.89	$\overline{\Delta R^2}$ (%)
	-1.93	-1.79	0.12	$\overline{\Delta F_1}$ (%)
50k	-4.00	-2.80	-8.30	$\overline{\Delta R^2}$ (%)
	-4.35	-2.76	-2.6-	$\overline{\Delta F_1}$ (%)

Table 8. MLE results across training sizes and synthetic data sizes.

training-size settings. As shown in Table 10, increasing the synthetic data size compared to the training size, also increases the Ensemble MIA success.

Training Size	Synthetic 1×	Synthetic 10×	Synthetic 1M	WB Success Rate
10k	34%	48%	48%	57%
20k	24%	33%	35%	46%
50k	11%	14%	14%	19%
100k	10%	11%	12%	14%

Table 9. MIA success rates across training sizes and synthetic data sizes in the black-box (BB) setting, with the white-box (WB) success rate shown as a baseline indicating the upper bound on BB attack performance.

Training Size	Synthetic 1×	Synthetic 10×
10k	23%	34%
20k	22%	28%

Table 10. Ensemble MIA success rates across various synthetic data size to training data size ratios.

C.3.1 Other datasets. Industry participants also validated some of the results of C.2 and C.3 on other datasets, namely the Madrid [4] and German Credit [17] datasets, as shown in Figure 3. In agreement with other findings DCR does not show much divergence from its ideal values for different training dataset sizes. The hitting rate rises with larger synthetic datasets (except for training

dataset size of 1K in the German Credit dataset), consistent with the MIA pattern. The quality metric, KS score, suffers most when the training data size takes its minimum at 1K. The metric improves by the increase in the output size (except for the train size of 1K) until the synthetic data size reaches the training data set size. The KS score remains mostly unaffected by increases thereafter.

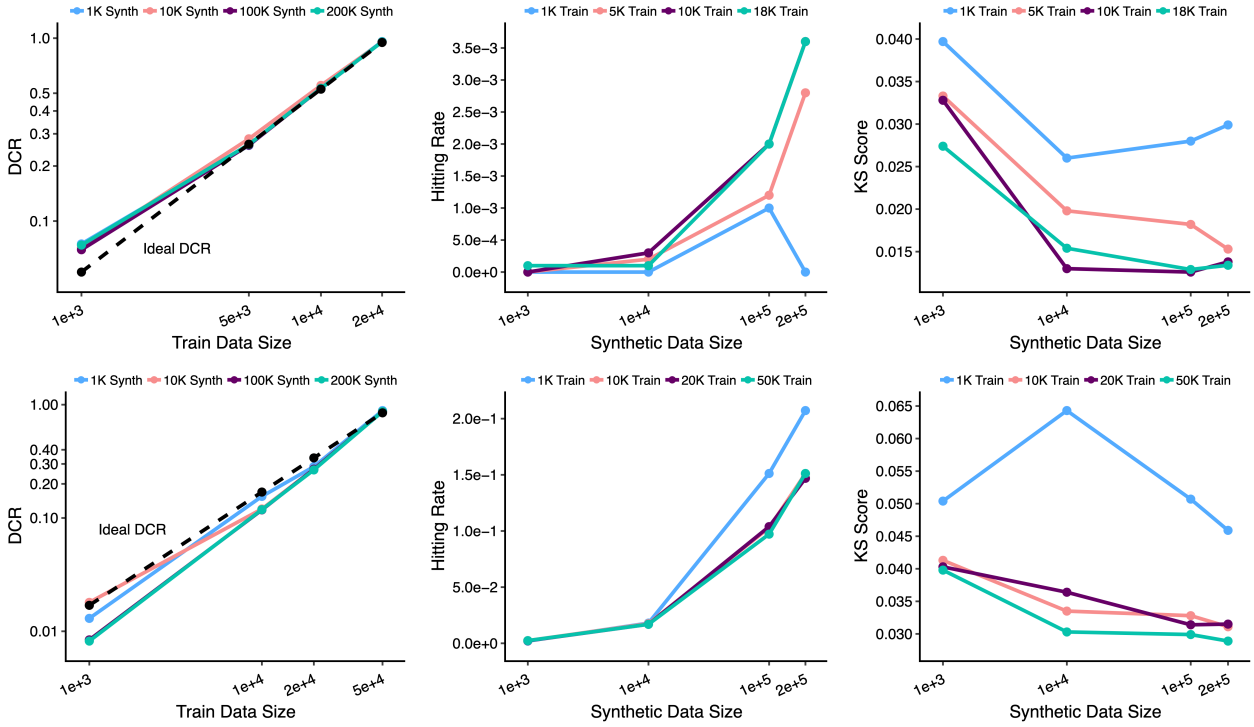

Fig. 3. Top: Credit dataset, bottom: Madrid dataset; the effect of train and synthetic data size on privacy and quality.

C.4 Attacker Knowledge of the Sensitive Data Distribution

We study scenarios where the adversary lacks access to the true training distribution and instead uses a related one to train shadow models. The aim is to understand how similarity between an attacker's data and the true training data affects MIA success.

C.4.1 Access to different marginal distributions. We examine two realistic cases:

(1) **Disjoint user accounts:** The adversary and trainer use transactions from different accounts in the Berka dataset.
(2) **Disjoint time windows:** The adversary observes later transactions; trainer observes earlier ones.

In both, the categorical and numerical feature marginals differ significantly at the 0.05 confidence level (KS and TVD tests), while correlation matrices and MIA remain similar, as shown in Table 11. This isolates the effect of mismatched marginals on MIA success. With disjoint accounts, attack success remains around 43. With disjoint time windows, it drops slightly from 43 to 41, indicating resilience to marginal shifts.

C.4.2 Perturbed correlation structures: To evaluate the effect of altered dependencies, we corrupt the adversary's auxiliary data by randomly permuting half of the categorical columns and replacing half of the numerical columns with uniform samples over their feature ranges. These produce substantial dependency changes and significant marginal differences at the 0.05 confidence level. Despite these large perturbations, MIA performance only decreases from around 46 to 41. Overall, the attack remains robust even when the adversary's data has markedly different marginal and joint structures.

C.5 Attacker Knowledge of the Target Model Hyperparameters

In this experiment, we focus on varying the model size, namely the number of middle layers and the width of the neural network of the diffusion model, while keeping all other hyperparameters at their default values. We keep the target model parameters set to default in all scenarios explored in this section, namely diffusion time steps = 2,000, training steps = 200,000, and model size = `[512, 1024, 1024, 1024, 1024, 512]`. We consider the following scenarios for the shadow models:

Exp.	White-Box TF (%)	Black-Box TF (%)	Corr. Δ	MI Δ	% Diff.
Disjoint Accounts	44.4 (43.2)	26.0 (26.8)	0.010	0.080	25.0
Disjoint Time Windows	41.6 (41.3)	24.4 (24.6)	0.142	0.029	87.5
Statistics-Based Synthesis	41.1 (46.2)	23.5 (24.6)	2.100	1.518	0.0
Noisy Adversary Data	41.6 (45.6)	24.0 (25.4)	2.000	1.449	100.0

Table 11. MIA performance under distribution mismatch: WB and BB denote white-box and black-box attack success, with baseline (matched-distribution) performance in parentheses. "Corr. Δ" is the Frobenius-norm difference between correlation matrices. "MI Δ" is the mutual-information matrix difference (categorical features only). "% Diff." is the fraction of feature columns whose empirical marginals differ significantly ($\alpha = 0.05$) between the adversary's data and the true training data.

(1) Diffusion training iterations set to a higher value of 300,000, while the diffusion time steps are default (2,000) and the model size is default (`[512, 1024, 1024, 1024, 1024, 512]`).
(2) Diffusion training iterations set to a lower value of 5,000, while the diffusion time steps are default (2,000) and the model size is default (`[512, 1024, 1024, 1024, 1024, 512]`).
(3) Diffusion time steps set to a middle value of 100, while the number of diffusion training iterations is default (200,000) and the model size is default (`[512, 1024, 1024, 1024, 1024, 512]`).
(4) Diffusion time steps set to the lowest value of 10, while the number of diffusion training iterations is default (200,000) and the model size is default (`[512, 1024, 1024, 1024, 1024, 512]`).
(5) Model size set to the heaviest configuration of `[1024, 2048, 2048, 2048, 2048, 2048, 2048, 1024]`, while the number of diffusion training iterations is default (200,000) and the number of diffusion time steps is default (2,000).
(6) Model size set to the lightest configuration of `[256, 512, 512, 256]`, while the number of diffusion training iterations is default (200,000) and the number of diffusion time steps is default (2,000).

The experiment results presented in Table 12 show that the attacker's success highly depends on the shadow model architecture. However, there are settings where success is not significantly reduced. For example, additional training steps or moderate changes in the number of diffusion steps do not significantly reduce attack success.

Metric	(1)	(2)	(3)	(4)	(5)	(6)
MIA (TF Black) (%)	25.3 (+0.1)	10.6 (-14.6)	24.6 (-0.6)	17.8 (-7.4)	25.4 (+0.2)	13.5 (-11.7)

Table 12. Quality and privacy metrics across different hyperparameters for shadow models; the original black-box success rate is 25.2.

C.6 Attacker Computing Power

We study how an attacker's computing power, measured by the number of shadow models used in attack construction, can affect MIA performance. Since the size of the attacker's training data used to train the MIA classifier grows linearly with the number of shadow models, one might expect attack success to improve as more models are trained. However, the results in Table 13 show that such improvement is limited at best. For both white-box and black-box attacks, the MIA success rate only slightly improves as the number of shadow models increases. This indicates that additional shadow models do not meaningfully strengthen the attack. Instead, the TF attack already extracts highly informative features even when constructed using a single shadow model.

# of Shadow Models	1	5	20	50
MIA TF Black-box (%)	24.1	28.7	27.5	27.1
MIA TF White-box (%)	42.3	43.4	45.8	45.9

Table 13. Increasing the number of shadow models yields subtle gains on *BB* and *WB* attack performance.