# ON PRIVACY IN DATA-SPACE TABULAR DIFFUSION MODELS: INFLUENTIAL FACTORS, ATTACKER KNOWLEDGE, AND METRICS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Tabular data plays an important role in many fields and industries, including those with elevated privacy considerations and risks. As such, there is a rising interest in generating high-quality synthetic proxies for real tabular data as a means of reducing privacy risk and proprietary data exposure. With data-space tabular diffusion models (TDMs) demonstrating leading performance in synthesizing such data, understanding and measuring the privacy risks associated with these models is imperative. Leveraging state-of-the-art membership inference attacks for such TDMs in both black- and white-box settings, this work quantifies the impact of training setup, synthesis choices, and attacker knowledge on privacy leakage. Moreover, the results demonstrate that adversaries need not have perfect knowledge of the training setup, identical data distributions, or massive compute resources to construct successful attacks. Finally, the substantial pitfalls associated with heuristic privacy metrics, such as distance-to-closest record, are identified.

## 1 INTRODUCTION

With the accelerating importance of machine learning (ML) models in all aspects of society, there has been a corresponding rise in interest in using advanced models to generate structured synthetic data (Hernandez et al., 2025; Stoian et al., 2026; Fang et al., 2024; Shi et al., 2025b). One of the most prominent modalities is tabular data, given its ubiquity across industries. High-fidelity data synthesis has many applications, including training data augmentation (Fajardo et al., 2021; Du et al., 2025; Ye-Bin et al., 2025; Ismail Fawaz et al., 2018; Dankar & Ibrahim, 2021), improving model robustness (Bartolo et al., 2021; Yang et al., 2020; Ruiz-Torrubiano et al., 2024; Moroianu et al., 2025), and exploring hypothetical or novel circumstances. However, a primary driver of interest in synthetic tabular data is its potential as a proxy for real data with reduced privacy risks. In many settings, existing and emerging privacy legislation, such as GDPR, HIPAA, and PIPEDA (European Parliament & Council of the European Union, 2016; Government of Canada, 2000; U.S. Congress, 1996), regulate the collection, use, and release of sensitive information and datasets. Moreover, many institutions strictly manage internal risk associated with privacy breach or data misuse. As such, synthetic tabular data is regarded as a prospective solution to privacy risks. While synthetic data has many advantages and can increase barriers to privacy breach, it is not intrinsically privacy preserving (Stadler et al., 2022), motivating significant research in privacy-preserving synthetic generation and privacy auditing techniques (Shokri et al., 2017; Kazmi et al., 2024; Wen et al., 2025; Sanchez-Serrano et al., 2025; Ghatak & Sakurai, 2022).

Recent work has shown that data-space tabular diffusion models (TDMs) achieve leading performance in tabular synthesis along a number of metrics (Pang et al., 2025; Kotelnikov et al., 2023; Shi et al., 2025a). While some research suggests that diffusion models are robust to overfitting (Bonnaire et al., 2025), other work has shown that TDMs still leak information (Ward et al., 2026), sometimes at higher rates than GAN-based counterparts (Zhu et al., 2025). Differentially private (DP) training can reduce privacy risks (Jayaraman & Evans, 2019; Annamalai et al., 2024b) but often comes with steep utility and efficiency tradeoffs, especially without large-scale public pretraining (Dockhorn et al., 2023; Liu et al., 2024; Zhu et al., 2025). As such, it is important to identify and quantify the mechanisms that impact privacy leakage in TDMs to guide practical and safe use of such models.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, state-of-the-art membership inference attacks (MIAs) for data-space TDMs are used as a lens to quantify privacy leakage. Such attacks are a fundamental privacy auditing technique for ML models (Carlini et al., 2022; Hayes et al., 2019; Salem et al., 2019; Shokri et al., 2017). Moreover, successful MIAs can be used to build deeper privacy attacks (Salem et al., 2023; Annamalai et al., 2024a). Using leading white-box (WB) and black-box (BB) MIAs, we investigate the impact of training choices, architecture configurations, and generation decisions on privacy leakage. Further, the influence of attacker knowledge, compute power, and data access on MIA success is examined. Finally, the efficacy of distance-to-closest-record (DCR) (Lautrup et al., 2024), and other widely used pseudo-privacy metrics, in identifying rising or falling privacy risk is compared to MIAs.

The results show that some traditional levers, such as training steps or dataset size, remain important for privacy in data-space TDMs. Others depend on the MIA technique or model, reinforcing the need to model a diversity of approaches, and some are surprisingly unimportant. Moreover, TDMs remain vulnerable to attacks even when adversaries have limited compute, imperfect knowledge, or corrupted data. Lastly, a nuanced and largely unreliable relationship between pseudo-privacy metrics, such as DCR, and MIA success is demonstrated. In some settings, DCR is well positioned to estimate rising MIA risk, while in others it fails entirely. These findings have broad implications for the practical use of TDMs and reinforce the importance of MIA auditing in synthetic data generation.

## 2 Related Work

MIAs have been established as a core technique for quantifying privacy risks in ML model training data (Shokri et al., 2017; Hayes et al., 2019; Salem et al., 2019; Chen et al., 2020; Carlini et al., 2022). The success of diffusion models has driven advances in MIAs tailored to these architectures, including TDMs. Model-agnostic, BB methods include DOMIAS (van Breugel et al., 2023), RMIA (Zarifzadeh et al., 2024), EPT (German et al., 2025), and others (Ward et al., 2026; Platzer & Reutterer, 2021). For WB settings, the SecMI approach (Duan et al., 2023) has been adapted for tabular models (Cheng & Bahmani, 2025; Wu et al., 2025; Yao et al., 2025; Mahyar et al., 2026). In this work, we focus on state-of-the-art approaches from the MIDST Challenge (Shafieinejad et al., 2026), ensuring the experimental framework leverages strong MIAs in both WB and BB settings.

Some existing work has considered factors that impact MIA success rates and utility in generative models for tabular data in controlled settings and for ML models more broadly. Salem et al. (2019) propose dropout and model stacking as defenses against MIAs for classification models. For the simple set of models studied, it is shown that shadow models need not have identical architectures or data distributions for successful MIAs. In (Gobbo, 2025), the effect of input-output ratio on the statistical similarity and predictive utility of synthetic data is studied. Overfitting is examined in (Hu et al., 2022) as a key driver of MIA success across different models, including GANs and VAEs. A taxonomy of existing defenses is provided, with the majority of generative model defenses falling under DP techniques. They also exhibit that traditional methods that combat overfitting are not necessarily enough. Finally, for classification models, studies have shown that training data diversity plays an important role, with MIA success decreasing with growing training data (Shokri et al., 2017). The present work focuses on a wide range of mechanisms and attack regimes that influence privacy leakage in data-space TDMs, going well beyond previous work.

Several studies investigate the interplay of DP and MIA success in the context of general ML models. Theoretical results demonstrate that small-$\epsilon$ DP provides an upper bound on MIA success (Úlfar Erlingsson et al., 2020; Yeom et al., 2017), while empirical studies suggest that larger $\epsilon$ values offer practical MIA protection (Lowy et al., 2024; Jayaraman et al., 2020). Humphries et al. (2020) investigate the impact of data dependencies on DP protections against MIAs for classifiers. The results indicate that DP does not fully protect against MIAs due to the correlative nature of the data distribution. While DP can provide privacy leakage protection, it can also significantly degrade synthetic data quality in diffusion models (Zhu et al., 2025). Moreover, experiments probing the mechanisms that reduce leakage in TDMs, as undertaken herein, provide a path to combining the benefits of high-$\epsilon$ DP with such levers to minimize privacy risk while maintaining maximum utility.

The impact of training decisions and attacker configurations on proxy metrics for privacy risks in synthetic data, especially distance-based measures, have been considered for TDMs. Experiments in (Fang et al., 2025) investigate the influence of factors such as training epochs, dataset size, and model type on a thresholded form of NNDR (Steier et al., 2025). Yao et al. (2025) examine DCR as

a statistical indicator of privacy risk for such models. Therein, for a limited setup, DCR-based hypothesis tests fail to indicate leakage uncovered by MIAs. The experiments in this work significantly expand these studies in several directions. The relationship of multiple heuristic privacy metrics to MIA success across a variety of setups is investigated, revealing an unreliable relationship where, in some settings, these metrics reflect rising privacy risk and dramatically fail in others. Further, the impact of model training, attack configuration, and adversary knowledge are explored relative to MIA success. Finally, other BB MIAs that do not require internal parameters are considered.

## 3 METHODOLOGY

This paper focuses on single-table synthesis using two frontier, data-space TDMs, ClavaDDPM (Pang et al., 2025), an improved form of TabDDPM (Kotelnikov et al., 2023), and TabDiff (Shi et al., 2025a). The main lens through which privacy is assessed is that of MIAs. Provided a set of data points and a trained target model, the goal of an MIA is to distinguish between points that were members of the model's training dataset and those that were not. In the domain of synthetic data, such methods are commonly broken into two categories, WB and BB attacks. BB attacks strictly operate on the synthetic outputs of the target model, potentially with access to related public data. Alternatively, WB attacks are afforded full-model access, including weights. Experiments are conducted with the Tartan Federer (TF) (Wu et al., 2025) and Ensemble (Lautraite et al., 2025) approaches from the MIDST Challenge (Shafieinejad et al., 2026). The TF method won the competition in both WB and BB settings, while the Ensemble technique placed second for BB attacks.

The Ensemble and TF attacks are fundamentally different in kind. The Ensemble method combines several distance-based approaches, probability mass estimations in DOMIAS, and pairwise likelihood ratio tests in RMIA. The approach only requires access to target-model generated synthetic data and real, similarly distributed background data as a reference. As such, it is not directly tied to diffusion architectures. However, its success can rely more heavily on auxiliary data availability. This is explored in Appendix D. Alternatively, whether applied as a WB or BB attack, the TF approach is tailored to TDMs, adapting the SecMI algorithm. Briefly, given a data point $x_0$, a forward process producing noisy samples at time $t$ such that $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon$ for some sampled Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, and a backward process model, $m_\theta(x_t, t)$ that predicts the noise in $x_t$, the TF attack extracts loss values for a candidate $x_0$ as $\|m_\theta(x_t, t) - \epsilon\|_2^2$ across a wide range of $\epsilon$ and $t$ values. Employing shadow models, these losses are used to train a dense neural network (DNN) for membership classification. In WB settings, target models are directly available to produce such losses at inference time. For the BB attack, proxies are constructed by training additional, target shadow models on the target's synthetic data from which the loss values are then extracted. For ClavaDDPM, the TF attack from (Shafieinejad et al., 2026) is applied, while the modified form in (Mahyar et al., 2026) is used for TabDiff. Additional details are provided in Appendix B.

### 3.1 ATTACK SETUP, METRICS, AND DATASETS

For a dataset $\mathcal{D}$, points are split into two subsets. The first, denoted $\mathcal{D}_t = \bigcup_{i=1}^{n_t}\{\mathcal{D}_t^i\}$, is used to train a set of target models, $\{m_\theta^i\}_{i=1}^{n_t}$. This dataset is strictly unavailable to an attacker. The second, $\mathcal{D}_h$, is a set of reference data available, in some form, to the attacker for the development of shadow models or calculating statistics between synthetic and real data. In both WB and BB setups, attackers are provided with synthetic datasets, $\mathcal{D}_s^i$, generated by $m_\theta^i$. Adversaries are also assumed to have knowledge of target-model architectures and training procedures, including hyperparameters. Traditionally, oracle knowledge of training and architecture elements is assumed. Below, however, settings with imperfect knowledge are also extensively explored. In the WB framework, all internal parameters for a target model $i$ are also revealed to attackers.

For each target model, membership challenge points, $\mathcal{C}_i$, are drawn from $\mathcal{D}$. Half are from of $\mathcal{D}_t^i$ and half are not. An MIA aims to provide an accurate distinction between such points across all target models. Following Carlini et al. (2022), MIA success is measured as the true-positive rate (TPR) at a false-positive rate (FPR) of $0.1$. Throughout, this is also simply referred to as attack success rate and privacy risk. Expected random classifier success in this metric is $0.1$. Several pseudo-privacy metrics are also measured, including DCR, NNDR, Hitting Rate, and EIR. DCR is reported in the main results as a representative metric, while definitions and results for others appear in Appendix E. Define $\mathcal{D}_h^i \subset \mathcal{D}_h$, a holdout datasets for model $i$. Let $x_s \in \mathcal{D}_s^i$. Denote by $x_t \in \mathcal{D}_t^i$ and $x_h \in \mathcal{D}_h^i$

points in the respective subsets closest, in some measure $d(\cdot, \cdot)$, to $x_s$. DCR is computed as the proportion of $x_s$ where $d(x_s, x_t) < d(x_s, x_h)$. An ideal DCR is given as $|\mathcal{D}_t^i|/|\mathcal{D}_t^i \cup \mathcal{D}_h^i|$.

Several synthetic tabular-data quality metrics are also computed, including $\alpha$-precision and $\beta$-recall (Alaa et al., 2022), KS statistics and TVD (Dankar et al., 2022), correlation and MI differences (Zhu et al., 2022; Ping et al., 2017), and machine-learning efficacy (Maheshwari et al., 2024). Definitions and values appear in Appendix F. The principle function of these metrics is to demonstrate that synthetic data quality does not collapse, decoupling MIA success from significant utility degradation.

Two large-scale tabular datasets are used in the experiments. The first is the Transactions table from the Berka dataset (Ventura, 2020) consisting of financial records from a Czech Bank, collected in 1999. The table has four numerical and four categorical columns with over $10^6$ rows of account transactions. The second table is the Diabetes dataset (Clore et al., 2014) with records derived from 130 US hospitals from 1999-2008. Each row represents a patient with diabetes and includes both numerical and categorical columns associated with demographics, clinical tests, and other health data. The dataset is smaller than Berka, containing just over $10^5$ records, but incorporates $47$ total features. For each dataset, row-level MIAs are conducted. Meaningfully, these datasets are larger than many in previous work. Preprocessing and MIA construction details appear in Appendix G.

### 3.2 Experiment Configurations: Training, Synthesis, and Attacker Knowledge

The influence of training setup, generation choices, and attacker knowledge on privacy risk is quantified through an extensive set of experiments. To start, a set of default configurations for ClavaDDPM, adapted from (Pang et al., 2025), and TabDiff, based on (Mahyar et al., 2026), are fixed for each dataset and MIA. These defaults are outlined in Appendices A–C. Sections 4.1 and 4.2 consider the impact of varying configurations on the success of WB and BB TF attacks across models. In Section 4.1, target-model training steps, training dataset size, and model composition are independently perturbed from the defaults. Details on model architecture variations appear in Appendix A. Experiments in Section 4.2 vary training batch size and diffusion timesteps. The influence of adjusting the ratio of synthetic data generated to training set size with respect to BB attack success is also evaluated. Target models generate synthetic data at various multiples of their training data, providing growing pools of synthetic data for attackers relative to training set size. Section 4.3 extends these results to the Ensemble attack setting, demonstrating generalization of the findings to other attack regimes. Several aspects specific to the Ensemble attack are also analyzed. Attack components are ablated to understand which contribute most to success. Different RMIA shadow model configurations are considered to understand the sensitivity of the algorithm to such choices. Finally, robustness to reference data volume is quantified. Results are reported in Appendix D.

Each setup above considers a standard MIA framework where attackers are assumed to be able to train many shadow models, perfect knowledge of the target-model training setup, and access to independently and identically distributed (IID) reference data. In Section 4.4, these assumptions are weakened in separate ways. First, MIA success as a function of the number of shadow models is considered, simulating resource constraints. Next, attackers train shadow models with configurations that diverge from the target model. Shadow models are trained for fewer or more iterations, use fewer diffusion timesteps, differ in size, or have different training set sizes compared to the target models. Finally, we consider settings where the data available to the adversary differs statistically from target-model training data. For Berka, four scenarios are explored. In the first, adversary and target model data is drawn from disjoint accounts. Second, the adversary observes data from a later time period, while target models are trained on earlier transactions. Third, adversary data is corrupted such that half of the categorical and numerical features are replaced with uniform draws across their respective ranges. Finally, independent sampling from one-way marginal distributions is simulated to constructed attacker data. For Diabetes, the final two scenarios are again used to corrupt adversary data. Appendix A.3 provides more details about the shadow model and dataset mismatch experiments, including quantifying the statistical divergence of the adversary datasets.

## 4 Main Results

The experiments in this section are structured to systematically probe the relationship between privacy leakage and training setup, attacker knowledge, and compute power. Where appropriate, the

connection, or lack thereof, between MIA success and DCR is also exhibited. Unless otherwise specified, when varying different aspects, such as training steps, all other settings remained fixed to the default values discussed in Section 3 and Appendices A–C.

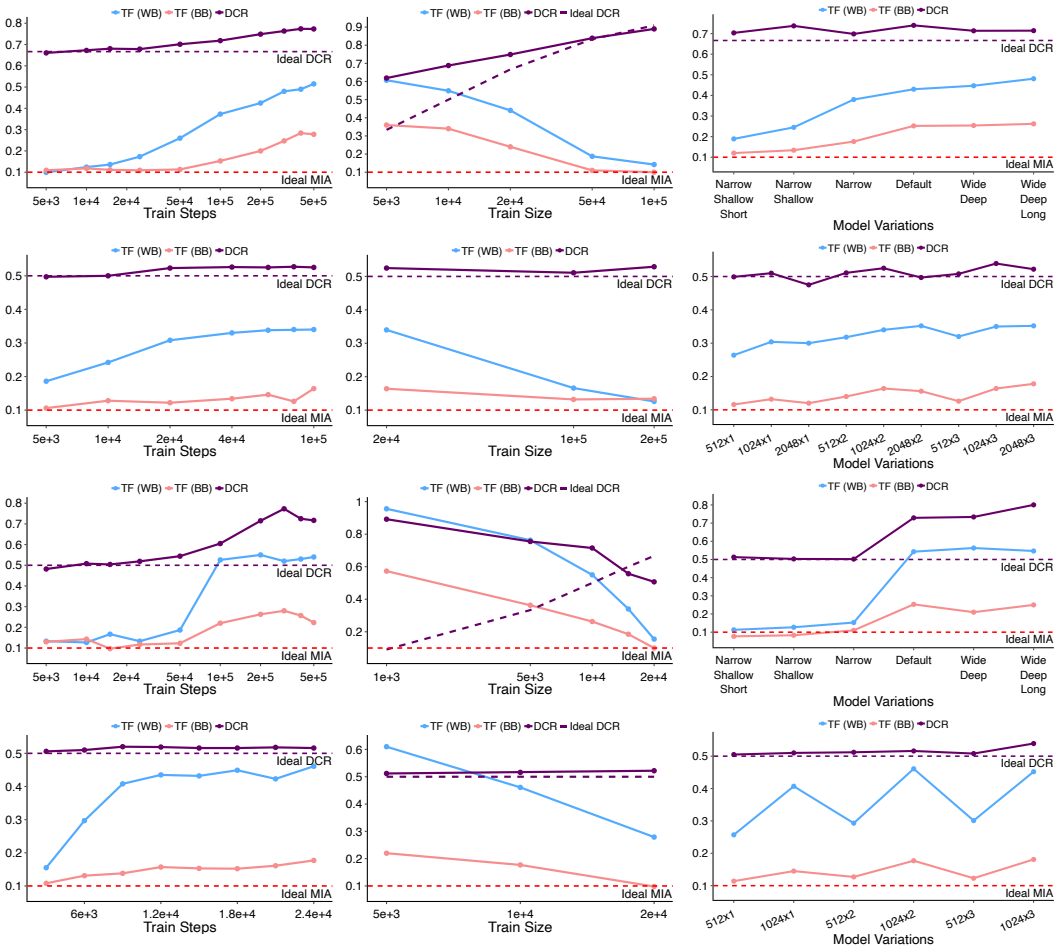


Figure 1: Variations in training steps, training dataset size, and architecture vs. TF MIA success and DCR in WB and BB settings for Berka (Rows 1, 2) and Diabetes (Rows 3, 4) across ClavaDDPM (Rows 1, 3) and TabDiff (Rows 2, 4) target models.

### 4.1 Training Steps, Training Size, and Model Variations

In this section, the number of training steps, training dataset size, and diffusion architecture are modified for both ClavaDDPM and TabDiff models. The influence of such variations on privacy leakage, measured via WB and BB TF attacks, is quantified. Despite previous research suggesting that diffusion models are more robust to memorization (Bonnaire et al., 2025), the number of training steps and amount of training data have a stark impact on privacy leakage. Increasing training iterations coincide with a nearly monotonic increase in MIA success for both TDMs in Figure 1 (left). Figure 1 (center) shows that larger training datasets dramatically reduce MIA success across all models and datasets. For ClavaDDPM, these relationships are generally captured by DCR, where it diverges from and converges to ideal values when increasing training steps or dataset sizes, respectively. However, it fails to capture the scale of rising leakage for the WB TF attack with increasing training steps for both datasets. On the other hand, DCR is quite poorly calibrated for TabDiff, hovering near the ideal value in all settings, despite significant privacy leakage and dramatic changes.

Figure 1 (right) exhibits attack success as a function of architecture changes. In line with previous work on simpler ML models, TDMs incorporating more parameters and longer training runs induce

rising MIA success. Furthermore, the DCR metric demonstrates significant unreliability for both TDMs. With ClavaDDPM, DCR largely reflects attack success for Diabetes but materially fails to do so in the case of Berka. A similar pattern is seen for other pseudo-privacy metrics in Appendix E. The larger number of training points and smaller set of columns in Berka potentially prevents simple distance-based output-space collapse in the generation process of ClavaDDPM, despite the potential for leakage clearly being present. The ablation studies in Appendix D, where none of the three Ensemble modules on their own produce significant MIA success, also suggests this. For TabDiff, DCR captures almost none of the observed leakage variations for either dataset, in spite of rising MIA success. In nearly all TabDiff experiments, DCR demonstrates poor sensitivity, signaling that pseudo-privacy measurements carry even less utility for this model compared to ClavaDDPM.

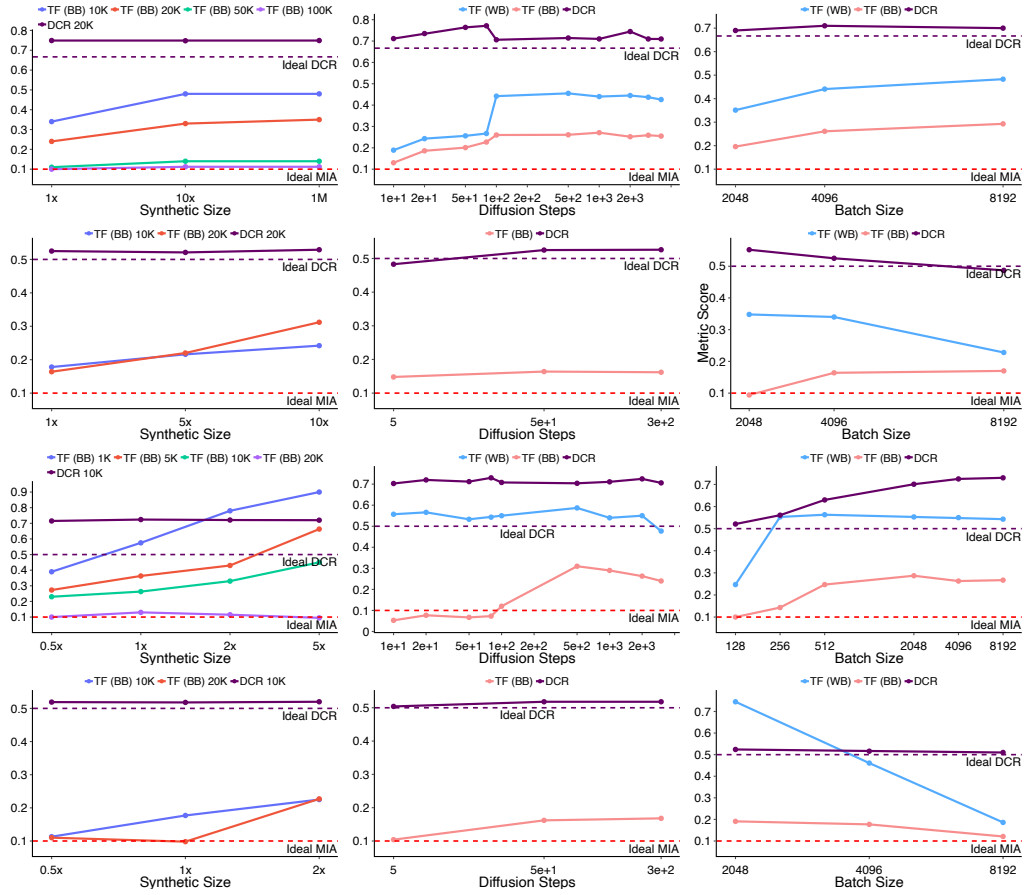


Figure 2: Variations in synthesis ratios, batch size, and diffusion steps vs. TF MIA success and DCR in WB and BB settings for Berka (Rows 1, 2) and Diabetes (Rows 3, 4) across ClavaDDPM (Rows 1, 3) and TabDiff (Rows 2, 4) target models.

### 4.2 Synthetic Ratio, Diffusion Steps, and Batch Size

Experiments in this section vary the number of diffusion timesteps, training batch size, and ratios of synthetic data to training data for both ClavaDDPM and TabDiff. Figure 2 (left) shows privacy leakage as the ratio of generated synthetic data to training data rises for the BB TF attack. In all cases, MIA success is strongly tied to this ratio, but the scaling also depends on the training set size and TDM. For example, in smaller data regimes, ClavaDDPM exhibits a drastic increase in privacy risk when generating $10\times$ more synthetic data than the size of its training set, whereas models trained on larger datasets do not experience an increase of comparable magnitude. This relationship is less clear for TabDiff, but generating more synthetic data uniformly increases leakage, regardless of training set size. In this setting, DCR remains deficient in capturing the marked increases in privacy leakage when synthesizing increasingly large quantities of data. This is likely because the

models' learned data-space distribution for generating synthetic samples is static, regardless of how many points either model is asked to produce after training, hampering DCR sensitivity.

As seen in Figure 2 (center), increasing the number of diffusion timesteps also typically increases leakage. Note that, as a continuous-time diffusion model, diffusion step discretization only occurs at inference time for TabDiff and, thus, only alters BB attack success. The effect is more pronounced for ClavaDDPM, likely due to it influencing both training dynamics and data generation. However, using more steps does moderately elevate leakage for TabDiff. For both datasets and TDMs, DCR is misleading, though it is worse for ClavaDDPM. For example, with Berka and ClavaDDPM, it does rise for the initial set of values, but then drops, suggesting that any risk may have modulated, though it did not. For Diabetes and TabDiff, MIA success rises, despite DCR hovering near ideal values.

The only lever for which the consequences of adjustment significantly diverge between the two models is batch size. In Figure 2 (right), MIA success climbs with increasing batch size for ClavaDDPM, intuitively matching previous work suggesting regularizing effects from small batches (Smith et al., 2020). In this regime, DCR is essentially flat for Berka, missing rising privacy risk. For the Diabetes dataset, DCR does rise and appears well-calibrated to BB TF success rate. However, there is an early jump in WB success rate for this dataset, which is maintained thereafter. This step change is not reflected by DCR or any other pseudo-privacy metric in Appendix E. On the other hand, for TabDiff, TF attack success generally drops with increasing batch size, especially for Diabetes. Some work does suggest that well-calibrated training can elicit good generalization with larger batch sizes (Do et al., 2025), but the contrast with ClavaDDPM is interesting. While the phenomenon is consistent between datasets, the mechanism behind this divergence is an important focus of future work.

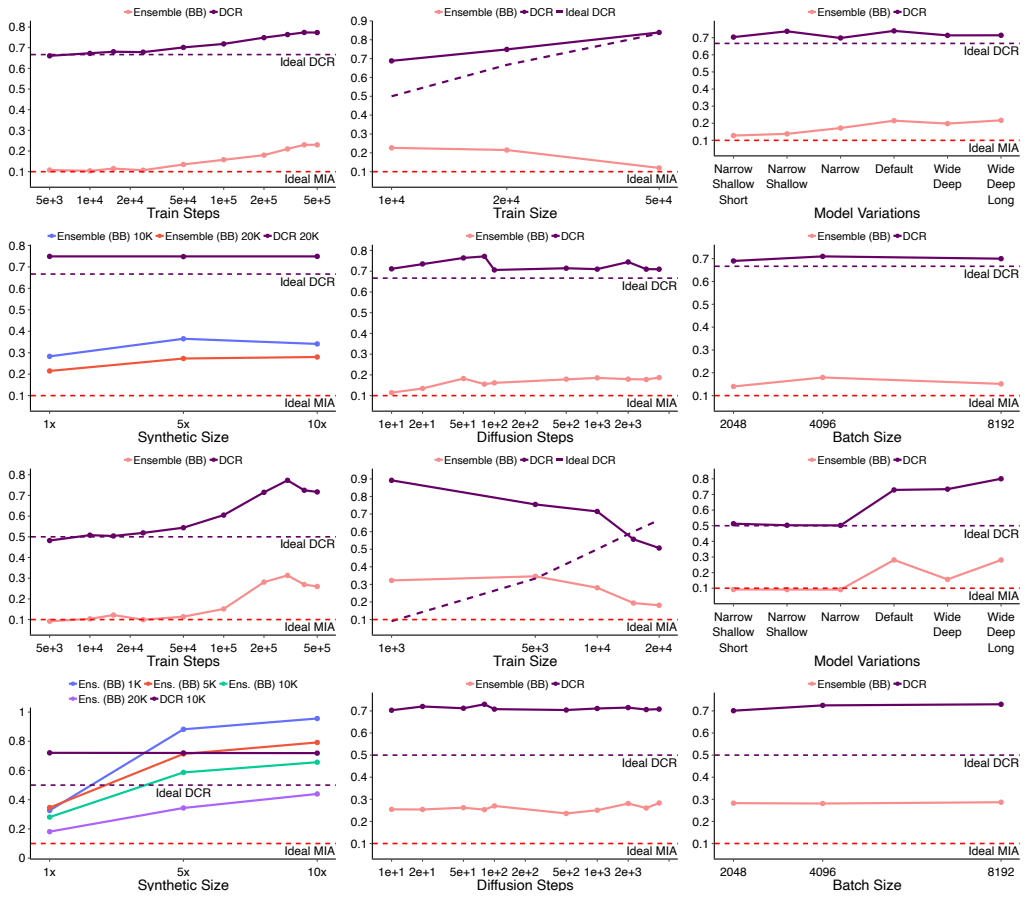


Figure 3: Variations in training steps, train size, architecture, synthesis ratios, diffusion steps, and batch size vs. Ensemble MIA success and DCR for Berka (Rows 1, 2) and Diabetes (Rows 3, 4).

### 4.3 Generalization to Other Attacks

A natural question is whether the patterns observed in the previous sections are also visible for designs beyond the TF attack. To test this, we apply the Ensemble attack to ClavaDDPM and explore the same levers. The results in Figure 3 demonstrate that the relationship of training steps, dataset size, and model variations to privacy leakage is also present in the Ensemble attack results. Here, DCR is more correlated with MIA success. However, it still fails for the Berka model variation experiments. The synthetic-to-train dataset ratios experiments also replicate the relationships of the TF attacks and further underscore the unreliability of DCR leakage estimates in this setting. Batch size has limited impact on Ensemble attack success, though both datasets maintain non-trivial success rates. This disconnect is potentially due to batch size more strongly influencing model loss landscapes and internal representations, on which the TF attack relies, compared to output-space structures, where the Ensemble approach operates (Keskar et al., 2017; Granziol et al., 2022).

The growth in Ensemble MIA success with respect to changes in diffusion steps is more tempered than the TF attack, in somewhat better alignment with DCR, but does modestly rise for both datasets while DCR does not. This moderation is likely due to the mechanism of loss reconstruction in the TF attack on ClavaDDPM, which is heavily influenced by the timestep scale of the diffusion model, whereas the Ensemble method operates solely on properties of the generated data. Recall that the Ensemble technique is a BB and data-driven approach with a non-trivial reliance on distance-, density-, and canary-based measures. As such, the results demonstrate that, while DCR and other pseudo-privacy measures provides a flawed estimate of privacy risk, see Appendix E, more sophisticated use of distance-based metrics can produce more meaningful and reliable privacy leakage evaluations. Regardless, many of the drivers of privacy leakage are consistent between the attacks, and the results emphasize the benefit of modeling distinct attack vectors.

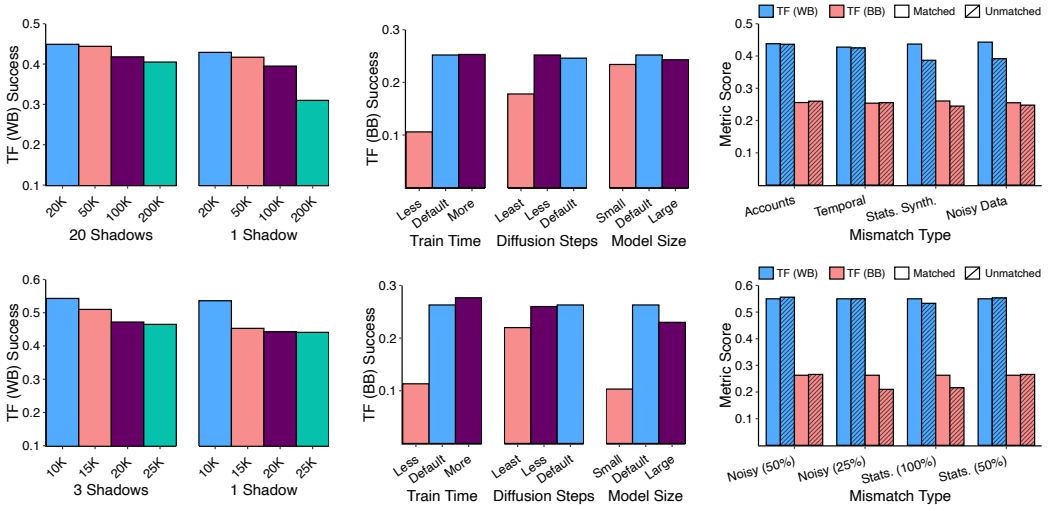


Figure 4: Variations in attacker compute (left), shadow model mismatch (center) and shadow-model data mismatch (right) vs. WB and BB TF MIA success for Berka (Row 1) and Diabetes (Row 2).

### 4.4 Attacker Resources, Knowledge, and Data

The privacy risks observed in previous sections accrue under the assumption that attackers have access to significant compute, perfect knowledge of training procedures, and a pool of IID reference data. Here, we consider whether such risks persist for TDMs beyond these ideal conditions using ClavaDDPM as the target model. Figure 4 displays the impact of variations in attacker compute power and imperfect knowledge or data access for the TF attack. The attack is surprisingly robust to reductions in the number of trained shadow models. For example, training a single such model reduces WB MIA success for both datasets by less than 0.03, contrasting previous work (Salem et al., 2019; Hu et al., 2022). This implies that adversaries need not apply heavy compute to construct successful TF attacks. MIA success is also resilient to shadow-model training mismatches. When trained on more data (Figure 4, left) than the target, attack success remains high. Using larger

shadow models than the target also preserves MIA success. When using a smaller shadow model, the attack remains effective for Berka, but does collapse for Diabetes. Training shadow models for significantly fewer iterations or with many fewer timesteps does noticeably reduce attack effectiveness. However, training such models for more iterations or with only marginally misaligned timesteps has little impact. The final set of experiments show that the TF attack is effective even when an attacker's data distribution differs in substantial ways. The BB version of the attack shows, at most, only small degradations across all scenarios. The WB attack is marginally impacted by the statistical synthesis and noisy data modifications, but remains quite strong.

The results for similar variations in attacker compute, target training knowledge, and reference data access for the Ensemble attack are shown in Figure 5. The default configuration for the attack constructs a single shadow model for meta-classifier training. The results demonstrate that this setup is sufficient and attackers need not train large collections of shadow models to produce successful MIAs. Similar to the TF attack, perfect knowledge of target-model training and architecture are not critical for Ensemble attack success. Longer training, fewer timesteps, or larger shadow models minimally degrade MIA success. Finally, for the Berka dataset, the Ensemble attack is robust to certain kinds of distribution divergence, especially disjoint accounts and statistically synthesized data. However, as a data-driven approach, it is more sensitive to temporal shifts and data noise. For Diabetes, the attack is resilient to certain kinds of data disagreement, but is impacted by heavily noised data. Additional results are discussed in Appendix D. Taken together, these results imply that successful attacks on TDMs may be constructed well beyond idealized scenarios.

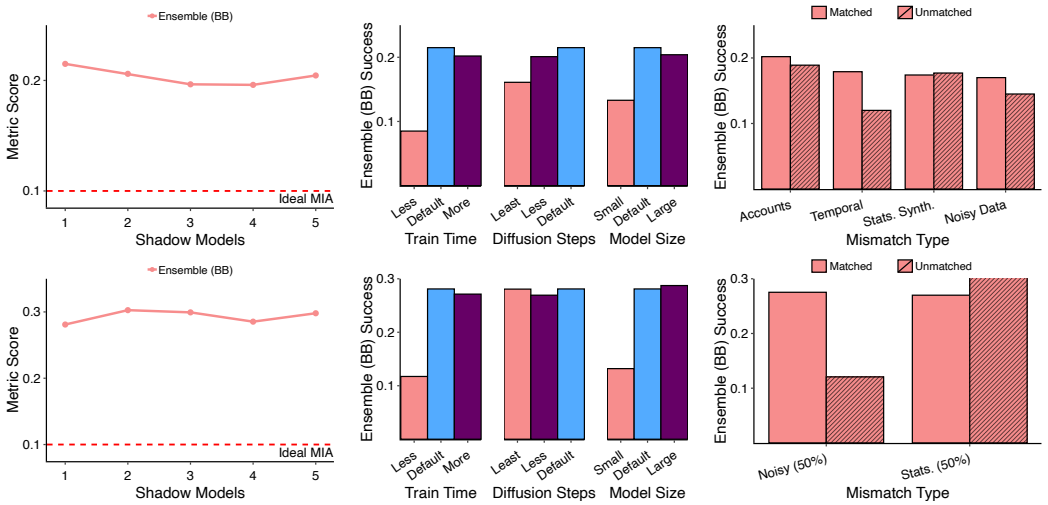


Figure 5: Variations in attacker compute (left), shadow model mismatch (center) and shadow-model data mismatch (right) vs. Ensemble MIA success for Berka (Row 1) and Diabetes (Row 2).

## 5 Conclusions, Limitations, and Future Work

In this work, factors that influence privacy leakage in data-space TDMs are quantified through controlled experimentation. Such factors include training choices, synthesis volume, and limitations on attacker compute, knowledge, and data access. Through the lens of leading WB and BB MIAs specifically designed for TDMs, we show that access to large compute resources, perfect knowledge of hyperparameters and model architecture, and access to IID data are not prerequisites to constructing successful attacks. Moreover, DCR, and other widely used pseudo-privacy risk metrics, are shown to be ineffective gauges of privacy risk in many scenarios, going well beyond previous studies. A potential limitation of this work is the specific focus on data-space TDMs. The development of strong MIAs for latent-space TDMs like TabSyn remains an open problem (Cheng & Bahmani, 2025; Zhang et al., 2024; Shafieinejad et al., 2026), precluding evaluation of the same levers considered here for such models. Future work will focus on developing such attacks and conducting a similarly extensive analysis of privacy leakage for these models. Further, we plan to apply the insights from this work to improve the utility-privacy tradeoffs in DP-training of diffusion models.

### AI USE STATEMENT

Generative AI tools were not used in place of human researchers for any core tasks in this research, including those requiring disclosure as part of the ICLR 2027 guidelines. LLMs were sparsely used in the course of writing the manuscript to improve readability by exploring potential rephrasings and grammatical structures. No central conclusions, takeaways, or ideas are derived with the help of Generative AI tools. All writing elements have been thoroughly reviewed by the authors. We take full responsibility for the final content of this work, including text, claims, or artifacts.

### ETHICS STATEMENT

All authors have read and respected the ICLR Code of Ethics through the course of this research. While the topic of this work concerns privacy leakage in ML models, no real privacy breach occurs in the experiments. We work with existing MIAs and use publicly available datasets, highlighting the impact of various training choices, attacker knowledge, and data access on privacy leakage. These experiments aim to inform practitioners of the risks associated with synthetic tabular data and tabular diffusion models. This includes how to better assess and control privacy leakage risks.

### REPRODUCIBILITY STATEMENT

The paper includes extensive details to ensure that the datasets, preprocessing, training setup, and experimental framework are fully understandable and reproducible for interested readers. The Methodology section in the main body and Appendices A–D, and G document the details associated with the main results and conclusions of this work. The code associated with the main results, including model training, attack construction, and evaluation will be made public along with this work. An anonymized version of the code used for the ClavaDDPM experiments and utility evaluation is here: `https://anonymous.4open.science/r/Privacy-TDMs-C3D3`. For the TabDiff experiments, the repository associated with (Mahyar et al., 2026) was kindly shared with us by the authors. The datasets used in this work, namely the Berka and Diabetes datasets, are publicly accessible through Kaggle and the UCI Machine Learning Repository, respectively. The experiments in this paper produce a large quantity of synthetic data and model artifacts, including many shadow models in various forms. Due to the volume of these assets, they will not be publicly available, but will be made available upon request. Finally, the computational resources required to run the experiments reported in the main results of this paper are discussed in detail in Appendix H.

## REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework, 2019. URL `https://arxiv.org/abs/1907.10902`.

Ahmed M. Alaa, Floris Van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. 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 (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 290–306. PMLR, 2022.

Meenatchi Sundaram Muthu Selva Annamalai, Andrea Gadotti, and Luc Rocher. A linear reconstruction approach for attribute inference attacks against synthetic data. In *Proceedings of the 33rd USENIX Conference on Security Symposium*, SEC '24, USA, 2024a. USENIX Association. ISBN 978-1-939133-44-1.

Meenatchi Sundaram Muthu Selva Annamalai, Georgi Ganev, and Emiliano De Cristofaro. "What do you want from theory alone?" Experimenting with tight auditing of differentially private synthetic data generation. In *Proceedings of the 33rd USENIX Conference on Security Symposium*, SEC '24, USA, 2024b. USENIX Association.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pp. 8830–8848, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Tony Bonnaire, Raphael Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2025. Curran Associates Inc.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022.

Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, pp. 343–362, New York, NY, USA, 2020. Association for Computing Machinery.

Peini Cheng and Amir Bahmani. Membership inference over diffusion-models-based synthetic tabular data, 2025. URL https://arxiv.org/abs/2510.16037.

John Clore, Krzysztof Cios, Jon DeShazo, and Beata Strack. Diabetes 130-US Hospitals for Years 1999-2008. UCI Machine Learning Repository, 2014.

Fida K. Dankar and Mahmoud Ibrahim. Fake it till you make it: Guidelines for effective synthetic data generation. *Applied Sciences*, 11(5), 2021.

Fida K. Dankar, Mahmoud K. Ibrahim, and Leila Ismail. A multi-dimensional evaluation of synthetic data generators. *IEEE Access*, 10:11147–11158, 2022.

Khoi Do, Minh-Duong Nguyen, Nguyen Tien Hoa, Long Tran-Thanh, Nguyen H. Tran, and Quoc-Viet Pham. Revisiting lars for large batch training generalization of neural networks. *IEEE Transactions on Artificial Intelligence*, 6(5):1321–1333, 2025.

Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially Private Diffusion Models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=ZPpQk7FJXF.

Mengze Du, Fei Wang, Weibing Yan, Jie Guo, Lu Liu, Ping Lv, Yong He, Xuping Feng, and Yuwei Wang. Improving food safety: Synthetic data augmentation for accurate mushroom species identification in complex environments. *Applied Food Research*, 5(1):101039, 2025.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

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

Val Andrei Fajardo, David Findlay, Charu Jaiswal, Xinshang Yin, Roshanak Houmanfar, Honglei Xie, Jiaxi Liang, Xichen She, and D.B. Emerson. On oversampling imbalanced data with deep conditional generative models. *Expert Systems with Applications*, 169:114463, 2021.

Xi Fang, Weijie Xu, Fiona Anting Tan, Ziqing Hu, Jiani Zhang, Yanjun Qi, Srinivasan H. Sengamedu, and Christos Faloutsos. Large language models (LLMs) on tabular data: Prediction, generation, and understanding - a survey. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=IZnrCGF9WI.

Zhengyu Fang, Zhimeng Jiang, Huiyuan Chen, Xiao Li, and Jing Li. Understanding and mitigating memorization in diffusion models for tabular data. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, 2025.

Eyal German, Daniel Samira, Yuval Elovici, and Asaf Shabtai. MIA-EPT: Membership inference attack via error prediction for tabular data, 2025. URL https://arxiv.org/abs/2509.13046.

Debolina Ghatak and Kouichi Sakurai. A survey on privacy preserving synthetic data generation and a discussion on a privacy-utility trade-off problem. In Chunhua Su and Kouichi Sakurai (eds.), *Science of Cyber Security - SciSec 2022 Workshops*, pp. 167–180, Singapore, 2022. Springer Nature Singapore. ISBN 978-981-19-7769-5.

Cristian Del Gobbo. A comparative study of open-source libraries for synthetic tabular data generation: SDV vs. SynthCity, 2025.

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

Diego Granziol, Stefan Zohren, and Stephen Roberts. Learning rates as a function of batch size: A random matrix theory approach to neural network training. *Journal of Machine Learning Research*, 23(173):1–65, 2022. URL http://jmlr.org/papers/v23/20-1258.html.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(1):133–152, 2019.

Mikel Hernandez, Pablo A Osorio-Marulanda, Mikel Catalina, Lorea Loinaz, Gorka Epelde, and Naiara Aginako. Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees. *Front Digit Health*, 7:1576290, 2025.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s), September 2022.

Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, U. Hengartner, and Florian Kerschbaum. Investigating membership inference attacks under data dependencies. *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*, pp. 473–488, 2020.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Data augmentation using synthetic data for time series classification with deep residual networks. In *International Workshop on Advanced Analytics and Learning on Temporal Data, ECML PKDD*, 2018.

Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, pp. 1895–1912, USA, 2019. USENIX Association.

Bargav Jayaraman, Lingxiao Wang, David E. Evans, and Quanquan Gu. Revisiting membership inference under realistic assumptions. *Proceedings on Privacy Enhancing Technologies*, 2021: 348–368, 2020.

Mishaal Kazmi, Hadrien Lautraite, Alireza Akbari, Qiaoyue Tang, Mauricio Soroco, Tao Wang, Sébastien Gambs, and Mathias Lécuyer. PANORAMIA: Privacy auditing of machine learning models without retraining. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=5atraF1tbg.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pp. 473–488, 2023.

Hadrien Lautraite, Lorrie Herbault, , Yue Qi, Jean-François Rajotte, and Sébastien Gambs. ensemble-mia. https://github.com/CRCHUM-CITADEL/ensemble-mia, 2025.

Anton D. Lautrup, Tobias Hyrup, Arthur Zimek, and Peter Schneider-Kamp. Syntheval: A framework for detailed utility and privacy evaluation of tabular synthetic data. https://arxiv.org/abs/2404.15821, 2024.

Bochao Liu, Pengju Wang, and Shiming Ge. Learning differentially private diffusion models via stochastic adversarial distillation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VII*, pp. 55–71, Berlin, Heidelberg, 2024. Springer-Verlag.

Andrew Lowy, Zhuohang Li, Jing Liu, Toshiaki Koike-Akino, Kieran Parsons, and Ye Wang. Why does differential privacy with large epsilon defend against practical membership inference attacks? *ArXiv*, abs/2402.09540, 2024.

Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. Efficacy of synthetic data as a benchmark, 2024. URL https://arxiv.org/abs/2409.11968.

Abtin Mahyar, Masoumeh Shafieinejad, Yuhan Liu, and Xi He. Fermi: Exploiting relations for membership inference against tabular diffusion models, 2026. URL https://arxiv.org/abs/2605.11527.

Stefania L. Moroianu, Christian Bluethgen, Pierre Chambon, Mehdi Cherti, Jean-Benoit Delbrouck, Magdalini Paschali, Brandon Price, Judy Gichoya, Jenia Jitsev, Curtis P. Langlotz, and Akshay S. Chaudhari. Improving performance, robustness, and fairness of radiographic ai models with finely-controllable synthetic data, 2025. URL https://arxiv.org/abs/2508.16783.

Wei Pang, Masoumeh Shafieinejad, Lucy Liu, Stephanie Hazlewood, and Xi He. ClavaDDPM: multi-relational data synthesis with cluster-guided diffusion models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

Haoyue Ping, Julia Stoyanovich, and Bill Howe. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, SSDBM '17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450352826.

Michael Platzer and Thomas Reutterer. Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in Big Data*, 4:679939, 2021.

Ruben Ruiz-Torrubiano, Gerhard Kormann-Hainzl, and Sarita Paudel. *Using Synthetic Data for Improving Robustness and Resilience in ML-Based Smart Services*, volume None of *Progress in IS*, chapter None, pp. 3–13. Springer, none edition, June 2024.

Ahmad Salem, Yang Zhang, Martin Humbert, Michael Fritz, and Manuel Backes. 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, February 24–27 2019.

Ahmed Salem, Giovanni Cherubin, David Evans, Boris Kopf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Beguelin. SoK: Let the Privacy Games Begin! A Unified Treatment of Data Inference Privacy in Machine Learning . In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 327–345, Los Alamitos, CA, USA, May 2023. IEEE Computer Society.

Pablo Sanchez-Serrano, Ruben Rios, and Isaac Agudo. A decision framework for privacy-preserving synthetic data generation. *Computers and Electrical Engineering*, 126:110468, 2025.

Masoumeh Shafieinejad, Xi He, Mahshid Alinoori, John Jewell, Sana Ayromlou, Wei Pang, Veronica Chatrath, Gauri Sharma, and Deval Pandya. MIDST Challenge at SaTML 2025: Membership Inference over Diffusion-models-based Synthetic Tabular data. *arXiv:2603.19185*, 2026.

Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. TabDiff: a mixed-type diffusion model for tabular data generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 37353–37375, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5c882988ce5fac487974ee4f415b96a9-Paper-Conference.pdf.

Ruxue Shi, Yili Wang, Mengnan Du, Xu Shen, Yi Chang, and Xin Wang. A comprehensive survey of synthetic tabular data generation, 2025b. URL https://arxiv.org/abs/2504.16506.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 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*, pp. 3–18. IEEE Computer Society, 2017.

Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9058–9067. PMLR, 13–18 Jul 2020.

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

Amy Steier, Lipika Ramaswamy, Andre Manoel, and Alexa Haushalter. Synthetic data privacy metrics, 2025. URL https://arxiv.org/abs/2501.03941.

Mihaela C. Stoian, Eleonora Giunchiglia, and Thomas Lukasiewicz. A survey on deep learning approaches for tabular data generation: Utility, alignment, fidelity, privacy, diversity, and beyond. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=RoShSRQQ67.

U.S. Congress. Health insurance portability and accountability act of 1996 (hipaa). Public Law 104–191, 110 Stat. 1936, 1996. URL https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf. Enacted August 21, 1996.

Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership inference attacks against synthetic data through overfitting detection. In Francisco J. R. Ruiz, Jennifer G. Dy, and Jan-Willem van de Meent (eds.), *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pp. 3493–3514. PMLR, 2023.

Marcelo Ventura. The berka dataset. https://www.kaggle.com/datasets/marcelov entura/the-berka-dataset, 2020. Accessed: 2025-12-10.

Joshua Ward, Yuxuan Yang, Chi-Hua Wang, and Guang Cheng. Ensembling membership inference attacks against tabular generative models. In *Proceedings of the 18th ACM Workshop on Artificial Intelligence and Security*, AISec '25, pp. 182–193, New York, NY, USA, 2026. Association for Computing Machinery.

Chengjiang Wen, Yang Yue, and Zhixiang Wang. The application of membership inference in privacy auditing of large language models based on fine-tuning method. In *Proceedings of the 2025 2nd International Conference on Generative Artificial Intelligence and Information Security*, GAIIS '25, pp. 473–479, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713453.

Xiaoyu Wu, Yifei Pang, Terrance Liu, and Steven Wu. Winning the MIDST challenge: New membership inference attacks on diffusion models for tabular data synthesis. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2503.12008.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative data augmentation for commonsense reasoning. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1008–1025, Online, November 2020. Association for Computational Linguistics.

Zexi Yao, Nataša Krčo, Georgi Ganev, and Yves-Alexandre de Montjoye. The DCR delusion: Measuring the privacy risk of synthetic data, 2025. URL https://arxiv.org/abs/2505.01524.

Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, Nayeong Kim, Suha Kwak, and Tae-Hyun Oh. Synaug: Exploiting synthetic data for data imbalance problems. *Pattern Recognition Letters*, 193:115–121, 2025.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2017.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58244–58282. PMLR, 21–27 Jul 2024.

Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *The twelfth International Conference on Learning Representations*, 2024.

Chaoyi Zhu, Jiayi Tang, Juan F. Pérez, Marten van Dijk, and Lydia Y. Chen. DP-TLDM: Differentially private tabular latent diffusion model, 2025. URL https://arxiv.org/abs/2403.07842.

Yujin Zhu, Zilong Zhao, Robert Birke, and Lydia Y. Chen. Permutation-invariant tabular data synthesis. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 5855–5864, 2022. doi: 10.1109/BigData55660.2022.10020639.

Úlfar Erlingsson, Ilya Mironov, Ananth Raghunathan, and Shuang Song. That which we call private, 2020. URL https://arxiv.org/abs/1908.03566.

# A Training and Attacker Configuration Experiment Details

As noted in Section 3.2, the default configurations for ClavaDDPM are adapted from (Pang et al., 2025), while the default settings for TabDiff are drawn from (Mahyar et al., 2026).

## A.1 ClavaDDPM Default Configuration and Variations

For ClavaDDPM, the default settings are as follows. The target models are DNN-based diffusion architectures with six layers and hidden dimension sequence [512, 1024, 1024, 1024, 1024, 512]. Each model is trained for 200000 steps with a batch size of 4096. The number of diffusion timesteps is 2000. For Berka, each target model is trained on 20000 distinct data points and generates an equal number of synthetic points such that $|\mathcal{D}_t^i| = |\mathcal{D}_s^i| = 20000$. A total of $n_t = 10$ target models are trained, each with a unique challenge set of size $|\mathcal{C}_i| = 200$. Alternatively, for the Diabetes dataset, target-model training and synthesized datasets are comprised of 10000 samples. There are three target models, each with $|\mathcal{C}_i| = 1000$. The focus of this work is record-level membership inference. As such, in constructing all datasets for the Berka and Diabetes experiments, any identifier columns are excluded and records are randomly sampled without replacement, including $\mathcal{D}_t^i$, $\mathcal{C}_i$, and others.

In the TF attack default setup, adversaries train 20 shadow models using 20000 points drawn without replacement from $\mathcal{D}_h$ for the Berka dataset. For Diabetes, seven shadow models are trained using 10000 points each. When applying the Ensemble attack for ClavaDDPM, a single target shadow model is trained by default, and 16 RMIA shadow models. The attack draws samples of 20000 points without replacement from $\mathcal{D}_h$ for Berka and 10000 samples for Diabetes to train the attack classifier.

In experiments modifying the diffusion model architecture, changes for each setting are as follows. The "narrow, shallow, short" setting uses a DNN with layers and dimensions [256, 512, 512, 256]

and trains for 100000 iterations. The "narrow, shallow" setting uses the aforementioned DNN but trained for 200000 iterations. For the "wider, deeper, long" configuration, the DNN is increased to have layers and dimensions [1024, 2048, 2048, 2048, 2048, 2048, 2048, 1024] and trains for 300000 steps. The "wider, deeper" experiment uses the larger DNN, but returns to 200000 training steps.

### A.2 TabDiff Default Configuration and Variations

The TabDiff target models are composed of transformer-based diffusion models. In the default setting, these models incorporate two transformer blocks, each with a DNN hidden dimension of 1024. Each model uses a training batch size of 4096. While TabDiff is a continuous-time diffusion model during training, inference is performed with discrete timesteps. The default number of diffusion steps is 50. For Berka, each target model is trained on 20000 points for 20000 epochs and generates an equal number of synthetic points. In each experiment, $n_t = 5$ target models are trained, each with a unique challenge set of size $|\mathcal{C}_i| = 2000$. For the Diabetes dataset, target-model training and synthetically generated datasets have 10000 samples. By default, these models are trained for 8000 epochs. There are 3 target models, each with $|\mathcal{C}_i| = 1000$. For the default TF attack setup, adversaries train 10 shadow models using 20000 points drawn without replacement from $\mathcal{D}_h$ for Berka. For Diabetes, seven shadow models are trained using 10000 points each. Similar to the ClavaDDPM experiments, all membership inference is record-level identification. For the Diabetes experiments, all datasets are constructed by randomly sampling records without replacement, as in the ClavaDDPM setup. For the Berka dataset, the partitions, challenge point compositions, and shadow model training sets are identical to that used in (Mahyar et al., 2026). This includes preprocessing procedures.

When modifying the TabDiff model architecture, the number of training steps is held constant for each dataset. Two dimensions of variation are considered. The number of repeated transformer blocks (depth) is varied from $D = 1$ to $D = 3$. For each depth, the size of the DNN latent dimension (width) is varied from $W = 512$ to $W = 2048$ for Berka. Due to the smaller dataset size of Diabetes, only $W = 512$ and $W = 1024$ are considered. In the figures, these variations are denoted as $W \times D$ along the x-axis.

### A.3 Attack Resources, Knowledge, and Data Experimental Details: ClavaDDPM

For experiments incorporating mismatches between target and shadow model setups for ClavaDDPM, shorter training times correspond to just 5000 steps, while longer training consumes 300000 steps. When varying diffusion timesteps, the model with the "least" steps uses only 10 and the "less" setting has 100. The small shadow models are DNN-based architectures with layers and dimensions [256, 512, 512, 256], and the large models are [1024, 2048, 2048, 2048, 2048, 2048, 2048, 1024]. For Figure 4 (left), where experiments induce mismatches between the training data volume of the shadow and target models, the default setting corresponds to the left-hand bars with a shadow-model training dataset size of 20000 for Berka and 10000 for Diabetes. The training dataset size for the shadow models is increased, but the target model dataset is held constant, along with all other settings and hyperparameters.

For the Berka dataset, when simulating imperfect adversary data access via disjoint accounts, accounts, and their corresponding transactions, are randomly assigned to the attacker or target model datasets such that there is no overlap. When segmenting transactions temporally, target models are trained on transactions occurring on or before 1997-04-10 and adversary data is drawn from transactions occurring thereafter. The scenario of statistics-based synthesis considers the setting where the adversary only possesses knowledge of population-level marginal feature distributions and can sample from them. To simulate this, we randomly permute the rows of all columns independently. This preserves the marginal distribution of each column but removes all inter-feature dependencies.

Finally, the noisy adversary data scenario assumes the adversary has significantly less information. That is, they know the unique values in each column but have no knowledge of the population-level marginal feature distributions. To simulate this, we replace each numerical column with draws from a uniform distribution over that column's observed range, and each categorical column with uniform draws over its observed categories. These perturbations induce substantial changes to the correlation structure, as reflected in the large correlation and MI matrix differences in Table 1.

Table 1: Comparison of the statistical distributions of the training data for target model and the adversary data used for training the shadow models under different distribution mismatch scenarios for the Berka dataset. "Corr. $\Delta$" is the Frobenius-norm difference between correlation matrices. "MI $\Delta$" is the Frobenius-norm difference between mutual-information matrices (categorical features only). "% Diff." is the fraction of feature columns whose empirical marginals differ significantly ($\alpha = 0.05$). For numerical columns, distributions are compared using a Kolmogorov-Smirnov test, while categorical columns are compared using Total Variation Distance with significance established via a permutation test (1000 permutations).

| Mismatch Type | Corr. $\Delta$ | MI $\Delta$ | % Diff. |
|---|---|---|---|
| Disjoint Accounts | 0.010 | 0.080 | 25.0 |
| Disjoint Time Windows | 0.142 | 0.029 | 87.5 |
| Statistics-Based Synthesis | 2.100 | 1.518 | 0.0 |
| Noisy Adversary Data | 2.000 | 1.449 | 100.0 |

For the Diabetes dataset, adversary data mismatches are induced via the same noisy and statistics-based synthesis strategies used for Berka. For both attacks, adversary data is modified by randomly selecting half of the numerical columns and half of the categorical columns and either applying the noisy or statistics-based synthesis strategies. Changes to the correlation structure for these manipulations are reported in Table 2. In addition, for the TF attack, we also consider the setting of noising 25% of each column type or statistics-based synthesis for 100% of the columns.

Table 2: Comparison of the statistical distributions of the training data for target model and the adversary data used for training the shadow models under different distribution mismatch scenarios for the Diabetes dataset. "Corr. $\Delta$" is the Frobenius-norm difference between correlation matrices. "MI $\Delta$" is the Frobenius-norm difference between mutual-information matrices (categorical features only). "% Diff." is the fraction of feature columns whose empirical marginals differ significantly ($\alpha = 0.05$). For numerical columns, distributions are compared using a Kolmogorov-Smirnov test, while categorical columns are compared using Total Variation Distance with significance established via a permutation test (1000 permutations).

| Mismatch Type | Corr. $\Delta$ | MI $\Delta$ | % Diff. |
|---|---|---|---|
| Statistics-Based Synthesis (50%) | 1.87 | 3.31 | 0.0 |
| Statistics-Based Synthesis (100%) | 1.90 | 3.33 | 0.0 |
| Noisy Adversary Data (25%) | 1.16 | 3.18 | 22.9 |
| Noisy Adversary Data (50%) | 1.78 | 2.21 | 47.9 |

In experiments varying training size for the Diabetes dataset, there is an important caveat to the experiments when the training sizes were 15000 and 20000. Because the size of the overall Diabetes set is limited, compared to Berka, only four and three shadow models are trained for attack development, while two models are still present as target models. This does imply that, in these settings, the attack has access to fewer shadow models. However, the results varying the number of shadow models for TF attack training in Figure 4 demonstrate that this does not dramatically undermine attack success.

# B Tartan Federer Attack: Default Settings and Other Details

As described in Section 3, the TF attack constructs features for each input, $x_0$, by computing the loss $\|(m_\theta(x_t, t) - \epsilon\|_2^2$ for a fixed set of $\epsilon$ and $t$ values. Note that the attack loss is modified for TabDiff following the description in (Mahyar et al., 2026).

## B.1 ClavaDDPM Setup

For the default number of diffusion steps, 2000, following (Wu et al., 2025), a static set of 300 initial noise values, $\epsilon$, are sampled from a standard Gaussian and the timesteps are $t \in$

$\{5, 10, 20, 30, 40, 50, 100\}$. This yields $2100$ features in total to be processed by the DNN for membership inference. When varying the number of timesteps in model architectures, models with timesteps in $\{500, 1000, 2000, 3000, 4000\}$ use the aforementioned set of timesteps. For models with $\{10, 20, 50, 80, 100\}$ timesteps, $t \in \{3, 4, 5, 6, 7, 8, 9\}$. These settings are used for both datasets. In all setups, the DNN used for membership classification has two hidden layers, each with dimension $200$. Hidden layers have $\tanh$ activations and the output layer is a sigmoid. Each shadow model contributes $3000$ member and $3000$ non-member samples. As such, in the Berka experiments with $20$ shadow models, this yields $120000$ training points for the network, whereas the Diabetes setting has seven shadow models, producing $42000$ training samples. The classifier is trained with standard binary cross-entropy loss, an Adam optimizer with default parameters, and a learning rate of 1e-04. Training proceeds for $5000$ steps with batch sizes of $6000$.

### B.2 TabDiff Setup

The TF attack for TabDiff replicates the settings of Mahyar et al. (2026). The TF loss features are constructed at time values $t \in \{0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.9\}$. Similar to ClavaDDPM, a fixed set of $500$ noise values are sampled. Numerical and categorical column loss features are constructed in distinct ways, consistent with Mahyar et al. (2026). As such, each row yields a feature vector of dimensionality $7 \cdot 500 \cdot 2 = 7000$ for both datasets in the experiments. The DNN used for membership classification has four hidden layers, each with dimension $128$. Hidden layers have $\mathrm{ReLU}$ activations and the output layer is a sigmoid. Batch normalization and dropout, with a rate of $0.2$, follow each dense layer. Each shadow model contributes $3000$ member and $3000$ non-member samples. As such, in the Berka experiments with $10$ shadow models, this yields $60000$ training points for the network, whereas the Diabetes setting has seven shadow models, producing $42000$ training samples. The classifier is trained with standard binary cross-entropy loss, an Adam optimizer with default parameters, and a learning rate of 5e-03. Training proceeds for $5000$ steps with batch sizes of $1024$.

## C Ensemble Attack: Default Settings and Other Details

By default, the attack classifier uses $20000$ samples from the Berka dataset or $10000$ samples from the Diabetes dataset drawn from $\mathcal{D}_h$ to which the attacker has access. The entire sampled subset is used to train the target shadow model. Further, half of the sample set is then used to train each of the RMIA shadow models. More specifically, each sample is randomly assigned to be used as training data for half of the RMIA shadow models. To construct the DOMIAS features, the full population data, $\mathcal{D}_h$, available to the attacker is used as the reference dataset, in line with the attack setup for the MIDST competition (Shafieinejad et al., 2026). For Berka, this consists of approximately $800000$ samples and $70000$ samples for the Diabetes dataset. The use of the entire reference dataset is relaxed in the ablations of Appendix D.

In the default settings, 16 different RMIA shadow models are trained, in agreement with the algorithm from (Shafieinejad et al., 2026). These are split into models with pre-train and fine-tune phases and those strictly with a single training phase. The first category of models are pre-trained on a subset of $\mathcal{D}_h$ not used to train the primary target shadow model constructed for the Ensemble attack. For Berka, this set is of size $60000$, and for Diabetes it has size $50000$. Thereafter, they are fine-tuned on the constructed subset of member points, as described above, from the training data of the target shadow model. For all experiments, except those in the results of Appendix D, two models are pre-trained and four models are fine-tuned from each of these base models, for a total of eight. In the second category, eight RMIA shadow models are also simply trained directly on a subset of data points drawn from the training data of the primary target shadow model with no pre-training phase.

The Ensemble attack meta-classifier is an XGBoost model trained on the features constructed through distance-based, DOMIAS-based, and RMIA-based extraction. The hyperparameters of the XGBoost model are optimized with 5-fold validation through Optuna (Akiba et al., 2019). The parameters optimized are

- `eta=trial.suggest_float("eta", 0.0001, 0.1, log=True)`,
- `max_depth=trial.suggest_int("max_depth", 3, 10)`,

- `subsample=trial.suggest_float("subsample", 0.1, 1)`,
- `colsample_bytree=trial.suggest_float("colsample_bylevel", 0.5, 1)`,
- `reg_alpha=trial.suggest_categorical("reg_alpha", [0, 0.1, 0.5, 1, 5, 10])`,
- `reg_lambda=trial.suggest_categorical("reg_lambda", [0, 0.1, 0.5, 1, 5, 10, 100])`.

For the Berka dataset, the meta-classifier training set is a random selection of 10000 points from the target shadow model training as member points and an equal number of samples drawn from $\mathcal{D}_h$ constituting non-members. The same is true for the Diabetes dataset, but with 5000 points for each.

## D TF AND ENSEMBLE ATTACK ABLATION RESULTS

In this section, additional experimental results associated with ablations of the TF and Ensemble MIAs are shared. The results in Figure 6 extend the experiments of Figure 4 (left) in terms of the number of shadow models trained for the WB and BB TF attacks across both datasets. While the success of the attack for both settings and across both datasets does improve with more models, the improvements are quite modest, implying that attacks need not have the compute or data resources required to train many shadow models in order to construct successful attacks.

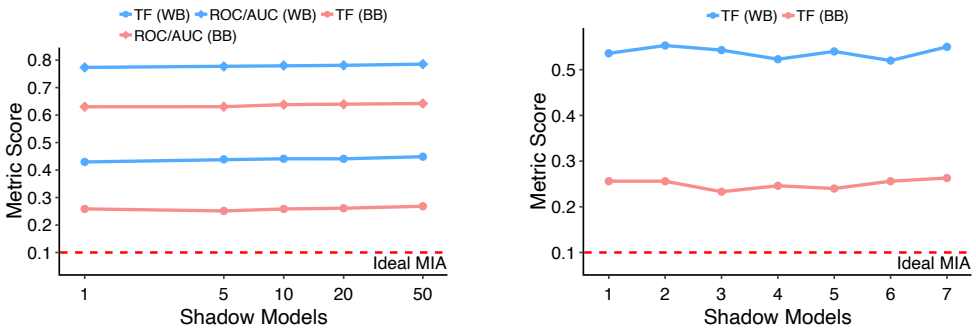


Figure 6: Variations in the number of shadow models used to construct the WB and BB TF attacks against ClavaDDPM target models for the Berka (left) and Diabetes (right) datasets.

As previously discussed, the Ensemble attack combines three components in RMIA, DOMIAS, and Distance-based features through a trained meta-classifier. To understand the contributions of each component and their interactions, attacks are constructed excluding features associated with different combinations of the ensemble. The results are shown in the top row of Figure 7. For the Berka dataset, there is a clear synergy associated with the components comprising the ensemble, as each one in isolation is insufficient to produce an MIA with significantly better than random performance. Interestingly, for the Diabetes dataset, the distance-based features are all that is necessary to produce a successful attack, whereas the RMIA and DOMIAS features alone are not effective. These results highlight that dataset composition plays a role in attack success and the strength of leveraging an ensemble-based approach to capture diverse inputs.

As detailed in Appendix C, the default construction of the RMIA approach is training 16 shadow models, eight single-stage and eight two-stage models. We experiment with alternative configurations to demonstrate robustness to this choice. The results appear in the bottom row of Figure 7. While the most consistent arrangement for both datasets is the 8+4+4 split, other RMIA shadow-model compositions perform well above random chance even with half the number of models and a single stage of training.

In the final set of ablations, the size of the reference dataset used to construct the DOMIAS features for both datasets is considered. By default, the attack uses all available reference points to construct such features. The results presented in Figure 8 examine using decreasing amounts of data for these computations. For both datasets, using smaller pools of reference data for the DOMIAS calculations

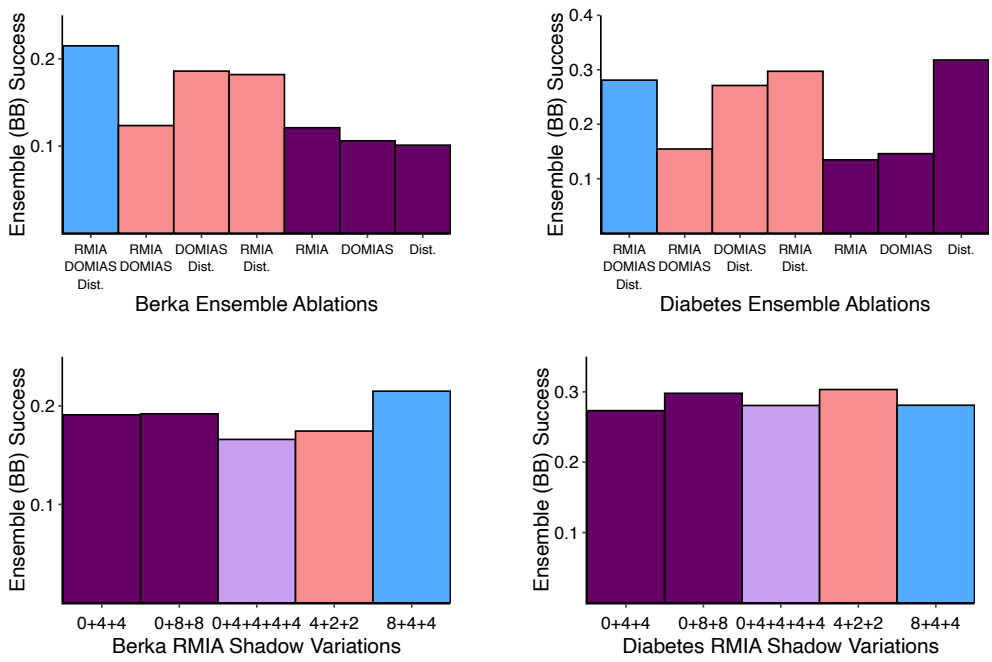


Figure 7: Ensemble attack ablations (top) and RMIA shadow model variations (bottom) for the Berka (left) and Diabetes (right) datasets. In the RMIA shadow experiments, the first number is single-phase training models, and the remaining are counts of fine-tuned models for each pre-trained base model. For example, 4+2+2 denotes four single-phase models and two sets of two models, each set trained on a distinct pre-trained base.

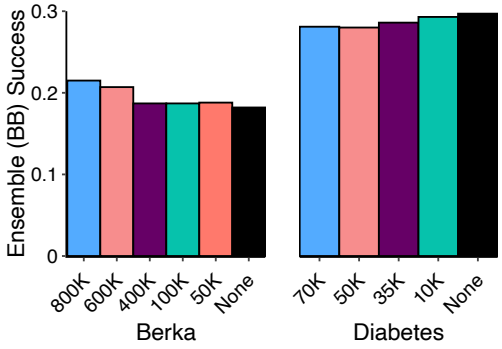


Figure 8: Ensemble attack success with variations in the amount of reference data available for the construction of DOMIAS features for both datasets. The size of the reference data used for the attack is along the x-axis. None indicates that the DOMIAS features are not used in MIA classification.

does not lead to attack collapse, implying that massive pools of reference data are not required for the Ensemble attack to be effective.

## E Privacy Heuristics: DCR, NNDR, HR, and EIR

In this appendix, definitions of the auxiliary pseudo-privacy risk metrics and computation details are provided. When computing DCR metrics for ClavaDDPM, $|\mathcal{D}_h^i| = 10000$ for both the Berka and Diabetes datasets. This implies an idea DCR of $0.\overline{66}$ and $0.5$ for these datasets, respectively, except for settings when the training dataset size is modified. On the other hand, for TabDiff, $|\mathcal{D}_h^i|$ is

augmented for both datasets such that the ideal DCR always remains 0.5, even when training dataset size is varied. The average DCR across all target models is reported. The distance measure used for DCR is the Euclidean norm with categorical variables one-hot encoded. Beyond DCR, the metrics described below are only computed for ClavaDDPM.

**Nearest Neighbor Distance Ratio (NNDR)** is the ratio of the $\ell_2$ distance between each synthetic data point and its two nearest neighbors in the training data (closest vs. second-closest). This measure is intended to capture whether a synthetic sample lies in a sparsely populated region of the training distribution, with values closer to 0 indicating this. Using a holdout dataset, comprising real data points not used to train the generative model, we recompute the NNDR and compare the two ratios. The difference between the train-based and holdout-based ratios yields a "privacy loss" score. In our experiments we report 1-NNDR values so that, similar to the other metrics, the larger the value, the more the synthetic data may reveal about the original training set.

**Hitting Rate (HR)** measures how often real data points are "replicated" in synthetic data, based on whether numerical values fall within a specified percentage (3% in these experiments) of a real training data point's value range and categorical values match exactly; the hitting (exact match) rate is the percentage of real points that meet these criteria, and lower rates indicate better privacy protection.

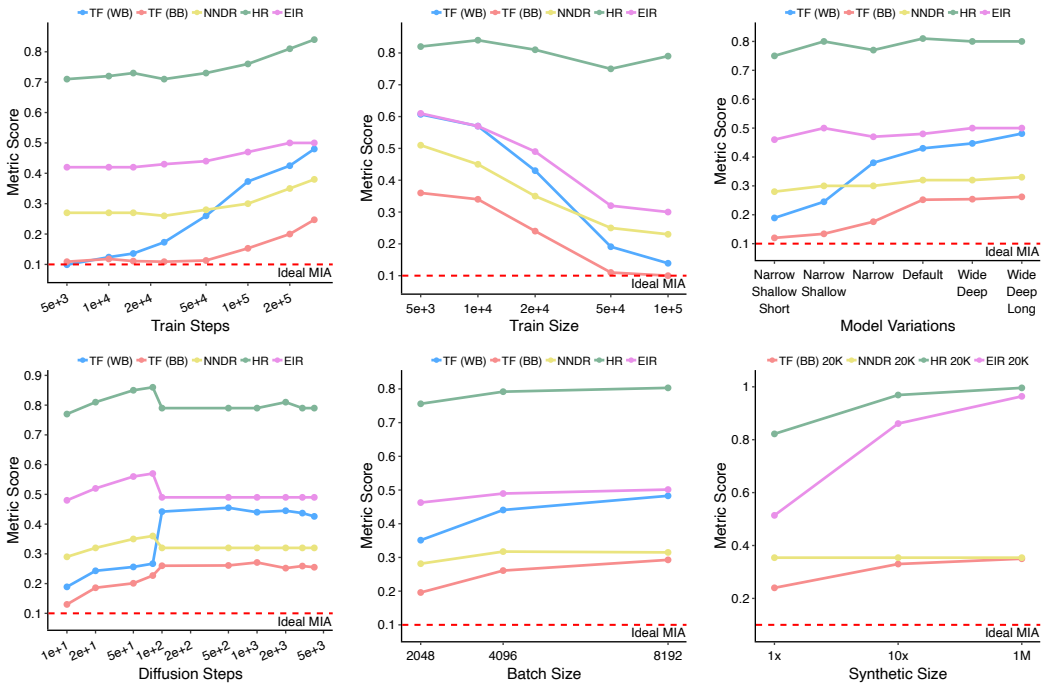


Figure 9: Training and synthesis levers that influence NNDR, HR, and EIR in comparison with TF MIA success for the Berka dataset.

**Epsilon Identifiability Risk (EIR)** measures the percentage of training points for which a synthetic data point is closer than any other real point, indicating potential overfitting of the model that produced the synthetic data. Values closer to zero are meant to reflect lower privacy risk. When a holdout set is available, the same ratio is computed for holdout points, and the difference between the training and holdout ratios is reported, ideally near zero or negative showing that synthetic data is not memorizing training examples. The metric also weights features by their inverse entropy, giving greater influence to rare attributes.

Figures 9 and 10 show the effects of training and synthesis levers on NNDR, HR, and EIR privacy heuristics compared to ClavaDDPM TF MIA for the Berka and Diabetes datasets, respectively. Note that for, all metrics, lower values are meant to indicate lower privacy risk. The plots indicate that, similar to DCR, none of these privacy heuristics are reliable as a standalone metric for privacy leakage. Several observations support this argument. Some heuristics report high privacy risk even

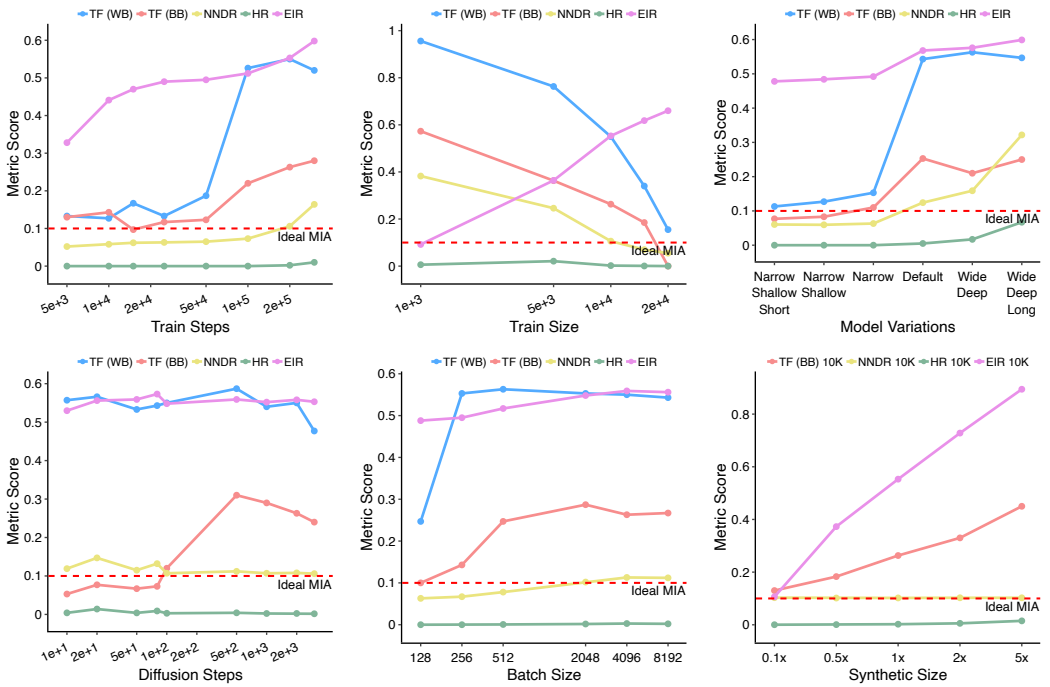


Figure 10: Training and synthesis levers that influence NNDR, HR, and EIR in comparison with TF MIA success for the Diabetes dataset.

when the WB MIA indicates no privacy risk, as HR does in most of the experiments of Figure 9. Some follow a completely different pattern than the WB and BB MIA, as EIR does in the *Train Size* experiment on the Diabetes dataset in Figure 9. Furthermore, some do not show any changes in privacy leakage with parameters variations unlike WB and BB MIA, as NNDR does in *Batch Size* and *Synthetic Size* experiments over the Diabetes dataset as well as HR in almost all experiments of Figure 10.

## F Quality Metrics and Results

In this appendix, more information is provided on the quality metrics used in this work. Further, the quality metrics associated with the target models considered are reported. The metrics offer insight into the experimental results of Section 4 in several ways. First, they showcase the relationship of synthetic-data quality to the various configuration changes and privacy risk. For example, increasing training iterations improves quality with diminishing returns while privacy risk continues to rise. Second, they demonstrate that the observed changes in MIA success are generally not tied to precipitous falloffs in synthetic quality. In most cases, generated data quality shows, at most, moderate and smooth variation with changes in the experimental configurations.

When computing quality metrics, for model $i$, evaluation compares the synthetic data, $\mathcal{D}_s^i$, to the training data, $\mathcal{D}_t^i$, to measure how closely the synthetic data resembles the training data it is meant to mimic. For each setting below, five target models are chosen at random for quality measurement and the resulting metric values averaged. Note that, depending on the experimental configurations, $|\mathcal{D}_s^i|$ is not always equal to $|\mathcal{D}_t^i|$. As fidelity and diversity measures of synthetic tabular data, $\alpha$-precision and $\beta$-coverage (-recall) metrics from (Alaa et al., 2022) are computed. The former measures how frequently generated samples fall within high-density regions of the real-data distribution. The latter measures the extent to which the synthetic distribution spans the diversity of the real data. To assess the extent to which generated data captures the statistical structure of the real data, differences in the one- and two-way marginals are computed. For numerical columns, a Kolmogorov-Smirnov (KS) test statistic is computed, comparing the synthetic and real dataset distributions for each column separately. Test statistics are then averaged over all columns. To compare real and synthetic

categorical column distributions, total variation distance (TVD) is computed, also yielding a test statistics, which is averaged over all columns. In both cases, larger statistics imply wider differences in column-wise distributions. As a measure of column dependency preservation, pair-wise correlation and mutual information (MI) matrices are computed for synthetic and real tables. The Frobenius norm of the difference between these matrices is calculated. Correlation matrices are computed only for numerical column pairs, while MI matrices include both column types and are computed using SynthEval (Lautrup et al., 2024). Metrics are computed against holdout sets of size 20000 and 10000 for the Berka and Diabetes datasets, respectively, where required.

The final two metrics computed are so called machine learning efficacy (MLE) measures and only calculated for ClavaDDPM. These measures estimate how well synthetic data preserves dataset utility by training models on both synthetic and real data and comparing their task performance. Following (Pang et al., 2025), separate models are trained to predict each column value from the others. For numerical columns, a random forest (RF) regressor is trained and an RF classifier is constructed for categorical columns. For model $m_\theta^i$, MLE models are trained on either synthetic data generated by $m_\theta^i$ or the training data, $\mathcal{D}_t^i$. Regressor performance is measured via $R^2$ scores and classification with macro $F_1$. The $R^2$ or $F_1$ metrics are averaged across all columns in the respective categories and the difference in the averages is reported. For example, $\Delta F_1 = \text{avg}(F_1)_{\text{synthetic}} - \text{avg}(F_1)_{\text{real}}$. Larger values imply closer, or even improved performance, compared to real data. Reported values are for synthetic and training data from a single target model and RF performance is evaluated on training data from a separate target model as holdout data.

Table 3: Quality metrics across training steps for ClavaDDPM models trained on Berka.

| Train Steps | 5e03 | 1e04 | 1.5e04 | 2.5e04 | 5e04 | 1e05 | 2e05 | 3e05 | 4e05 | 5e05 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\beta$ Cov. ↑ | 5.91e-01 | 5.78-e-01 | 5.91e-01 | 4.29e-01 | 4.51e-01 | 4.58e-01 | 4.79e-01 | 4.89e-01 | 5.04e-01 | 5.05e-01 |
| $\alpha$ Prec. ↑ | 9.31e-01 | 9.37e-01 | 9.33e-01 | 9.97e-01 | 9.94e-01 | 9.96e-01 | 9.93e-01 | 9.95e-01 | 9.96e-01 | 9.94e-01 |
| Avg. KS ↓ | 7.02e-03 | 9.53e-03 | 8.48e-03 | 3.99e-02 | 4.03e-02 | 3.96e-02 | 3.90e-02 | 3.91e-02 | 3.91e-02 | 3.82e-02 |
| Avg. TVD ↓ | 1.28e-02 | 1.25e-02 | 1.35e-02 | 7.30e-03 | 9.30e-03 | 8.32e-03 | 5.82e-03 | 6.84e-03 | 4.85e-03 | 7.17e-03 |
| Corr. Diff. ↓ | 5.25e-02 | 3.81e-02 | 3.81e-02 | 3.09e-02 | 3.24e-02 | 3.35e-02 | 3.77e-02 | 2.36e-02 | 2.79e-02 | 4.32e-02 |
| MI Diff. ↓ | 4.18e-02 | 2.43e-02 | 1.19e-02 | 8.02e-03 | 4.11e-03 | 5.19e-03 | 3.58e-03 | 4.19e-03 | 4.44e-03 | 6.32e-03 |
| $\Delta R^2$ ↑ | 6.20e-03 | 6.90e-03 | 5.80e-03 | 1.28e-02 | -6.29e-03 | -1.00e-03 | 2.40e-03 | -2.50e-03 | 8.00e-05 | 5.50e-05 |
| $\Delta F_1$ ↑ | -1.22e-02 | -1.64e-02 | -1.05e-02 | -1.53e-02 | -1.84e-02 | -2.07e-02 | -1.92e-02 | -1.72e-02 | -1.72e-02 | -1.46e-02 |

Table 4: Quality metrics across training dataset sizes for ClavaDDPM models trained on Berka.

| Training Size | 5e03 | 1e04 | 2e04 | 5e04 | 1e05 |
|---|---|---|---|---|---|
| $\beta$ Cov. ↑ | 6.06e-01 | 6.16e-01 | 5.44e-01 | 5.32e-01 | 3.37e-01 |
| $\alpha$ Prec. ↑ | 9.29e-01 | 9.29e-01 | 9.93e-01 | 9.31e-01 | 9.95e-01 |
| Avg. KS ↓ | 4.24e-02 | 4.17e-02 | 3.90e-02 | 3.91e-02 | 3.81e-02 |
| Avg. TVD ↓ | 8.90e-03 | 8.55e-03 | 5.82e-03 | 5.75e-03 | 2.75e-03 |
| Corr. Diff. ↓ | 6.92e-02 | 5.01e-02 | 3.77e-02 | 1.32e-02 | 1.75e-02 |
| MI Diff. ↓ | 7.01e-03 | 7.06e-03 | 3.58e-03 | 4.98e-03 | 8.86e-03 |
| $\Delta R^2$ ↑ | -3.90e-04 | -3.27e-02 | -2.10e-04 | -1.11e-01 | -1.21e-01 |
| $\Delta F_1$ ↑ | -2.20e-03 | -8.60e-02 | -1.01e-02 | -1.25e-02 | -1.68e-02 |

### F.1 ClavaDDPM: Berka Quality Results

Synthetic data quality measures for target models trained on the Berka dataset are reported in Tables 3–7. The primary take away from these tables is that synthetic data quality does vary with training configuration changes, but it does not sharply degrade in any of the scenarios. As such, the main privacy risk results presented in Section 4 are based on models with reasonable utility and reflect legitimate modeling-privacy trade-offs.

For the Berka dataset, synthetic quality is impacted by training choices in different ways. In the ranges tested, some parameter changes have fairly limited impact on quality, such as batch size or model architecture changes. Others, like training iterations and diffusion steps, generally improve the quality metrics with larger values. Finally, training dataset size has a more nuanced impact on quality, where some metrics improve with more data and some improve to a point and then degrade.

Table 5: Quality metrics across different ClavaDDPM model configurations trained on the Berka dataset. Scenario (1) is "narrow, shallow, short," (2) is "narrow, shallow," (3) is "narrow," (4) is default, (5) is "wide, deep," (6) is "wide, deep, long."

| Scenario | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| $\beta$ Cov. ↑ | 5.95e-01 | 5.75e-01 | 6.06e-01 | 4.70e-01 | 6.13e-01 | 5.87e-01 |
| $\alpha$ Prec. ↑ | 9.30e-01 | 9.47e-01 | 9.30e-01 | 9.90e-01 | 9.30e-01 | 9.33e-01 |
| Avg. KS ↓ | 7.47e-03 | 7.05e-03 | 8.99e-03 | 3.90e-03 | 6.21e-03 | 8.80e-03 |
| Avg. TVD ↓ | 8.60e-03 | 7.47e-03 | 7.15e-03 | 4.30e-03 | 5.94e-03 | 8.47e-03 |
| Corr. Diff. ↓ | 2.92e-02 | 2.49e-02 | 4.56e-02 | 5.64e-02 | 3.21e-02 | 3.27e-02 |
| MI Diff. ↓ | 1.32e-02 | 4.80e-03 | 5.92e-03 | 4.23e-03 | 1.12e-02 | 4.53e-03 |
| $\Delta R^2$ ↑ | -8.90e-03 | -4.32e-03 | 1.70e-03 | 2.40e-03 | -3.40e-03 | 4.00e-03 |
| $\Delta F_1$ ↑ | -1.16e-02 | -5.40e-03 | 2.10e-03 | -3.70e-03 | 2.70e-03 | 4.10e-03 |

Table 6: Quality metrics across diffusion timesteps for ClavaDDPM models trained on Berka.

| Diffusion Steps | 10 | 20 | 50 | 80 | 100 | 500 | 1000 | 2000 | 3000 | 4000 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\beta$ Cov. ↑ | 5.20e-01 | 5.90e-01 | 6.13e-01 | 5.17e-01 | 4.94e-01 | 4.94e-01 | 4.97e-01 | 4.79e-01 | 4.98e-01 | 4.95e-01 |
| $\alpha$ Prec. ↑ | 9.30e-01 | 9.22e-01 | 9.28e-01 | 9.87e-01 | 9.90e-01 | 9.90e-01 | 9.90e-01 | 9.97e-01 | 9.90e-01 | 9.90e-01 |
| Avg. KS ↓ | 2.63e-02 | 1.23e-02 | 8.55e-03 | 7.73e-03 | 7.38e-03 | 8.05e-01 | 7.80e-03 | 3.96e-03 | 7.67e-03 | 7.82e-03 |
| Avg. TVD ↓ | 5.90e-02 | 2.55e-02 | 1.07e-02 | 9.52e-03 | 6.92e-03 | 7.47e-03 | 4.77e-03 | 4.35e-03 | 5.45e-03 | 7.12e-03 |
| Corr. Diff. ↓ | 1.46e-01 | 7.12e-02 | 3.23e-02 | 2.98e-02 | 4.64e-02 | 3.87e-02 | 3.32e-02 | 5.64e-02 | 3.44e-02 | 4.02e-02 |
| MI Diff. ↓ | 1.24e-01 | 4.32e-02 | 1.33e-02 | 7.29e-03 | 7.66e-03 | 9.13e-03 | 1.64e-02 | 4.23e-03 | 6.24e-03 | 8.20e-03 |
| $\Delta R^2$ ↑ | -1.48e-02 | -1.23e-02 | -6.40e-03 | -1.51e-03 | -1.30e-03 | -6.30e-03 | 7.10e-03 | 2.40e-03 | 8.90e-03 | 7.90e-03 |
| $\Delta F_1$ ↑ | -2.33e-02 | -1.57e-02 | -1.49e-02 | -1.82e-02 | -5.30e-03 | -6.60e-03 | -8.00e-03 | -4.80e-03 | 2.20e-03 | -4.40e-03 |

Table 7: Quality metrics across batch sizes for ClavaDDPM models trained on Berka.

| Batch Size | 2048 | 4096 | 8192 |
|---|---|---|---|
| $\beta$ Cov. ↑ | 4.90e-01 | 4.90e-01 | 4.80e-01 |
| $\alpha$ Prec. ↑ | 9.90e-01 | 9.90e-01 | 9.90e-01 |
| Avg. KS ↓ | 7.40e-03 | 7.60e-03 | 8.10e-03 |
| Avg. TVD ↓ | 4.90e-03 | 4.70e-03 | 6.80e-03 |
| Corr. Diff. ↓ | 4.44e-02 | 2.90e-02 | 4.02e-02 |
| MI Diff. ↓ | 4.30e-03 | 3.90e-03 | 6.10e-03 |
| $\Delta R^2$ ↑ | 2.30e-03 | 2.40e-03 | -5.40e-03 |
| $\Delta F_1$ ↑ | -1.22e-02 | -1.17e-02 | -1.43e-02 |

Table 8: Quality metrics across training steps for ClavaDDPM models trained on Diabetes.

| Train Steps | 5e03 | 1e04 | 1.5e04 | 2.5e04 | 5e04 | 1e05 | 2e05 | 3e05 | 4e05 | 5e05 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\beta$ Cov. ↑ | 3.19e-01 | 4.28e-01 | 4.61e-01 | 4.89e-01 | 5.30e-01 | 6.01e-01 | 6.60e-01 | 6.93e-01 | 6.61e-01 | 6.73e-01 |
| $\alpha$ Prec. ↑ | 8.00e-01 | 8.09e-01 | 7.77e-01 | 8.19e-01 | 8.33e-01 | 7.81e-01 | 7.53e-01 | 7.77e-01 | 7.58e-01 | 7.63e-01 |
| Avg. KS ↓ | 1.93e-01 | 8.00e-02 | 4.30e-02 | 2.10e-02 | 2.00e-02 | 3.60e-02 | 3.31e-02 | 3.10e-02 | 3.50e-02 | 3.20e-02 |
| Avg. TVD ↓ | 1.75e-01 | 8.10e-02 | 5.10e-02 | 3.20e-02 | 2.60e-02 | 3.30e-02 | 3.64e-02 | 3.20e-02 | 3.60e-02 | 3.50e-02 |
| Corr. Diff. ↓ | 2.11e-00 | 1.66e+00 | 1.25e+00 | 5.33e-01 | 3.44e-01 | 3.25e-01 | 1.42e-01 | 9.40e-02 | 1.41e-01 | 1.23e-01 |
| MI Diff. ↓ | 7.27e-00 | 5.78e+00 | 4.40e+00 | 2.81e+00 | 5.85e+00 | 4.31e+00 | 5.84e+00 | 5.84e+00 | 7.35e+00 | 4.07e+00 |
| $\Delta R^2$ ↑ | -5.77e-01 | -5.40e-02 | -2.70e-02 | -3.10e-02 | 1.90e-02 | -1.60e-02 | -1.78e-01 | -1.30e-02 | -2.70e-02 | 1.60e-02 |
| $\Delta F_1$ ↑ | -4.00e-02 | -2.80e-02 | -2.40e-02 | -1.80e-02 | -1.80e-02 | -1.60e-02 | -4.10e-02 | -1.00e-02 | -3.00e-03 | -9.00e-03 |

### F.2 CLAVADDPM: DIABETES QUALITY RESULTS

Tables 8–12 report the quality metrics for variations in the target model training setups on the Diabetes dataset. As discussed in Section 3.1, the Diabetes dataset is smaller, in total size, than Berka, but has a larger number of columns. As such, there is more variation in the target model quality results with changes to the experimental configurations. It should be noted that, with the increased

number of columns, the correlation and MI matrices are also much larger, thereby inducing larger Froebenius norms for similar quality statistics.

Table 9: Quality metrics across training data sizes for ClavaDDPM models trained on Diabetes.

| Training Size | 1e03 | 5e03 | 1e04 | 1.5e04 | 2e04 |
|---|---|---|---|---|---|
| $\beta$ Cov. ↑ | 1.53e-01 | 4.95e-01 | 6.60e-01 | 6.93e-01 | 7.12e-01 |
| $\alpha$ Prec. ↑ | 7.22e-01 | 7.16e-01 | 7.53e-01 | 7.81e-01 | 8.29e-01 |
| Avg. KS ↓ | 4.80e-02 | 4.29e-02 | 3.31e-02 | 3.20e-02 | 2.90e-02 |
| Avg. TVD ↓ | 5.10e-02 | 4.40e-02 | 3.64e-02 | 3.20e-02 | 2.90e-02 |
| Corr. Diff. ↓ | 4.27e-01 | 2.71e-01 | 1.42e-01 | 1.54e-01 | 1.43e-01 |
| MI Diff. ↓ | 8.84e+00 | 8.72e+00 | 5.84e+00 | 5.84e+00 | 5.31e+00 |
| $\Delta R^2$ ↑ | -8.80e-02 | -3.60e-02 | -1.78e-01 | 1.00e-02 | 4.10e-02 |
| $\Delta F_1$ ↑ | -2.60e-02 | -1.25e-02 | -4.10e-02 | -5.00e-03 | 1.00e-03 |

Table 10: Quality metrics across different ClavaDDPM model configurations trained on the Diabetes dataset. Scenario (1) is "narrow, shallow, short," (2) is "narrow, shallow," (3) is "narrow," (4) is default, (5) is "wide, deep," (6) is "wide, deep, long."

| Scenario | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| $\beta$ Cov. ↑ | 4.73e-01 | 4.78e-01 | 4.96e-01 | 6.85e-01 | 6.81e-01 | 7.42e-01 |
| $\alpha$ Prec. ↑ | 8.18e-01 | 8.51e-01 | 8.55e-01 | 7.66e-01 | 7.63e-01 | 7.74e-01 |
| Avg. KS ↓ | 3.80e-02 | 4.00e-02 | 2.60e-02 | 3.50e-02 | 3.60e-02 | 3.20e-02 |
| Avg. TVD ↓ | 4.10e-02 | 4.30e-02 | 2.70e-02 | 3.50e-02 | 3.60e-02 | 3.30e-02 |
| Corr. Diff. ↓ | 1.14e+00 | 1.54e+00 | 8.87e-01 | 3.27e-01 | 3.11e-01 | 1.43e-01 |
| MI Diff. ↓ | 4.95e+00 | 5.16e+00 | 3.20e+00 | 4.01e+00 | 4.01e+00 | 5.84e+00 |
| $\Delta R^2$ ↑ | -3.10e-02 | -6.40e-02 | -1.90e-02 | -2.30e-02 | -1.90e-02 | -1.08e-01 |
| $\Delta F_1$ ↑ | -3.10e-02 | -2.10e-02 | -1.50e-02 | -8.00e-03 | -1.30e-02 | -1.60e-02 |

Table 11: Quality metrics across diffusion timesteps for ClavaDDPM models trained on Diabetes.

| Diffusion Steps | 10 | 20 | 50 | 80 | 100 | 500 | 1000 | 2000 | 3000 |
|---|---|---|---|---|---|---|---|---|---|
| $\beta$ Cov. ↑ | 6.55e-01 | 6.79e-01 | 6.67e-01 | 6.63e-01 | 6.62e-01 | 6.55e-01 | 6.59e-01 | 6.68e-01 | 6.66e-01 |
| $\alpha$ Prec. ↑ | 6.52e-01 | 7.04e-01 | 7.44e-01 | 7.45e-01 | 7.25e-01 | 7.58e-01 | 7.59e-01 | 7.52e-01 | 7.72e-01 |
| Avg. KS ↓ | 4.90e-02 | 4.20e-02 | 3.80e-02 | 3.60e-02 | 3.70e-02 | 3.60e-02 | 3.70e-02 | 3.50e-02 | 3.40e-02 |
| Avg. TVD ↓ | 4.90e-02 | 4.20e-02 | 3.80e-02 | 3.60e-02 | 3.80e-02 | 3.70e-02 | 3.70e-02 | 3.60e-02 | 3.40e-02 |
| Corr. Diff. ↓ | 1.52e-01 | 1.18e-01 | 1.35e-01 | 1.18e-01 | 1.86e-01 | 1.27e-01 | 1.38e-01 | 1.49e-01 | 1.24e-01 |
| MI Diff. ↓ | 7.36e+00 | 7.35e+00 | 5.84e+00 | 4.01e+00 | 7.00e-03 | 7.35e+00 | 5.84e+00 | 4.01e+00 | 5.84e+00 |
| $\Delta R^2$ ↑ | -1.50e-02 | -3.40e-02 | -4.60e-02 | -1.70e-02 | -1.80e-02 | -1.60e-02 | -2.90e-02 | -2.60e-02 | -2.50e-02 |
| $\Delta F_1$ ↑ | -9.00e-03 | -9.00e-03 | -1.10e-02 | -8.00e-03 | -1.70e-02 | -1.10e-02 | -1.60e-02 | -9.00e-03 | -1.00e-02 |

Similar to the Berka results, generated data quality does not collapse across a wide variety of training setups. However, there are a few instances, at the extreme ends of the hyperparameter spectra, where metrics indicate larger deterioration. For instance, when training for only a few thousand steps in Table 8, many of the metrics are significantly worse. Similarly very small training sizes also produce low $\beta$-coverage and large correlation differences.

The dependencies of the quality metrics for target models trained on the Diabetes dataset have some similarities and notable differences compared to the Berka-trained models. Quality metrics are inconsistently affected by changes in batch size and increasing training steps generally improves such measurements. On the other hand, for the Diabetes dataset, larger models and training dataset sizes mostly improve synthetic quality, while the impact of changing the number of diffusion steps is varied.

Table 12: Quality metrics across batch sizes for ClavaDDPM models trained on Diabetes.

| Batch Size | 128 | 256 | 512 | 2048 | 4096 | 8192 |
|---|---|---|---|---|---|---|
| $\beta$ Cov. ↑ | 4.87e-01 | 5.65e-01 | 6.02e-01 | 6.62e-01 | 6.71e-01 | 6.64e-01 |
| $\alpha$ Prec. ↑ | 8.67e-01 | 8.06e-01 | 7.68e-01 | 7.49e-01 | 7.70e-01 | 7.64e-01 |
| Avg. KS ↓ | 3.60e-02 | 2.90e-02 | 3.62e-02 | 3.60e-02 | 3.40e-02 | 3.50e-02 |
| Avg. TVD ↓ | 3.90e-02 | 2.90e-02 | 3.62e-02 | 3.60e-02 | 3.40e-02 | 3.50e-02 |
| Corr. Diff. ↓ | 6.99e-01 | 2.32e-01 | 2.15e-01 | 1.60e-01 | 1.40e-01 | 1.06e-01 |
| MI Diff. ↓ | 3.45e+00 | 5.84e+00 | 2.66e+00 | 5.84e+00 | 5.84e+00 | 5.84e+00 |
| $\Delta R^2$ ↑ | -1.54e-02 | -6.00e-03 | -1.56e-02 | -2.50e-02 | -2.00e-02 | -6.00e-03 |
| $\Delta F_1$ ↑ | -3.90e-02 | -1.30e-02 | -1.11e-02 | -1.30e-02 | -8.00e-03 | -1.30e-02 |

Table 13: Synthetic data quality metrics across experimental settings for target TabDiff models trained on Berka.

| Category | Setting | $\beta$ Cov.↑ | $\alpha$ Prec.↑ | Avg. KS↓ | Avg. TVD↓ | Corr. Diff.↓ | MI Diff.↓ |
|---|---|---|---|---|---|---|---|
| Train Steps | 5e03 | 0.592 | 0.914 | 0.020 | 0.014 | 0.147 | 0.107 |
| | 1e04 | 0.590 | 0.925 | 0.012 | 0.010 | 0.091 | 0.101 |
| | 2e04 | 0.519 | 0.953 | 0.009 | 0.007 | 0.053 | 0.090 |
| | 4e04 | 0.534 | 0.957 | 0.010 | 0.008 | 0.057 | 0.099 |
| | 6e04 | 0.583 | 0.926 | 0.010 | 0.007 | 0.057 | 0.097 |
| | 8e04 | 0.580 | 0.934 | 0.009 | 0.007 | 0.064 | 0.094 |
| | 1e05 | 0.587 | 0.944 | 0.009 | 0.006 | 0.059 | 0.096 |
| Architectures | 512x1 | 0.544 | 0.950 | 0.009 | 0.007 | 0.071 | 0.101 |
| | 1024x1 | 0.562 | 0.941 | 0.009 | 0.006 | 0.061 | 0.103 |
| | 2048x1 | 0.558 | 0.936 | 0.009 | 0.013 | 0.061 | 0.098 |
| | 512x2 | 0.584 | 0.933 | 0.008 | 0.006 | 0.063 | 0.095 |
| | 1024x2 | 0.587 | 0.944 | 0.009 | 0.006 | 0.059 | 0.096 |
| | 2048x2 | 0.525 | 0.938 | 0.006 | 0.008 | 0.048 | 0.089 |
| | 512x3 | 0.567 | 0.935 | 0.009 | 0.006 | 0.062 | 0.095 |
| | 1024x3 | 0.589 | 0.927 | 0.009 | 0.008 | 0.064 | 0.099 |
| | 2048x3 | 0.525 | 0.951 | 0.008 | 0.005 | 0.067 | 0.097 |
| Diffusion Steps | 5 | 0.360 | 0.728 | 0.205 | 0.054 | 1.155 | 1.003 |
| | 50 | 0.587 | 0.944 | 0.009 | 0.006 | 0.059 | 0.096 |
| | 300 | 0.515 | 0.965 | 0.009 | 0.005 | 0.045 | 0.024 |
| Train Size (Equal Synth.) | 2e04 | 0.587 | 0.944 | 0.009 | 0.006 | 0.059 | 0.096 |
| | 1e05 | 0.579 | 0.933 | 0.038 | 0.005 | 0.064 | 0.091 |
| | 2e05 | 0.526 | 0.944 | 0.037 | 0.005 | 0.069 | 0.090 |
| Batch Size | 2048 | 0.566 | 0.925 | 0.009 | 0.008 | 0.060 | 0.097 |
| | 4096 | 0.587 | 0.944 | 0.009 | 0.006 | 0.059 | 0.096 |
| | 8192 | 0.588 | 0.928 | 0.009 | 0.007 | 0.080 | 0.098 |
| Synth. Size (2e04 Train) | 2e04 | 0.547 | 0.944 | 0.009 | 0.006 | 0.059 | 0.096 |
| | 1e05 | 0.585 | 0.927 | 0.007 | 0.005 | 0.056 | 0.100 |
| | 2e05 | 0.564 | 0.938 | 0.006 | 0.004 | 0.052 | 0.097 |

### F.3 TabDiff: Berka Quality Results

Quality metrics for TabDiff, trained on the Berka dataset, for the various levers investigated in Sections 4.1 and 4.2 are displayed in Table 13. Overall, the quality metrics are generally stable across the range of hyperparameters and alternative settings investigated. At the extremes, there are some signs of model quality degradation. For example, when training for only $5000$ steps, $\alpha$ precision, Avg. KS, and Corr. Diff begin to show signs of stress. Similarly, using a very small number of diffusion steps reduces synthetic data quality along a number of metrics. Even still, the attack shows non-trivial success for this dataset. Several takeaways are shared with the results for ClavaDDPM. Training models for longer improves quality to a point. However, beyond that, privacy leakage significantly scales without in-kind quality benefits. More training data does not uniformly

improve all quality metrics, but it does markedly reduce privacy risk. Finally, larger models can produce some quality improvements, but also increased leakage. Measuring quality, along with privacy, allows for informed tradeoff decisions by the data modeler.

An important observation from the quality metrics is that synthetic data quality for Berka does not vary significantly with batch size. There are some variations with batch size changes for Diabetes, but they are not consistent in the sense that most of the "best" metric values occur at the default batch size of $4096$ rather than on either end of the batch size spectrum. This is interesting due to the observed differences in MIA success rate between ClavaDDPM and TabDiff. Data utility does not appear to suggest clear explanations for the attenuating MIA success with increasing batch size seen in the TabDiff experiments. Future work is required to more deeply understand the mechanisms at work in these results.

### F.4 TabDiff: Diabetes Quality Results

Table 14 exhibits the quality metrics for TabDiff models trained using the Diabetes dataset. The table includes measures across the different configurations reported in Sections 4.1 and 4.2. As noted above, the Diabetes dataset is smaller than Berka, but has more columns and representation diversity. It also tends to exhibit privacy leakage at a higher rate for both ClavaDDPM and TabDiff models. While large synthetic data quality degradation is only present at certain setting extremes, model quality is more sensitive for Diabetes, compared to Berka. For example, When training for only $3000$ steps, $\alpha$ precision is much lower, Avg. KS and MI Diff. are much higher than such values at $24000$ steps. Similar quality drop off is seen when using the smallest model configurations and very few diffusion steps to generate the data. Nevertheless, the tradeoffs around utility and privacy seen for TabDiff models trained on Berka are also seen here.

Table 14: Synthetic data quality metrics across experimental settings for target TabDiff models trained on Diabetes.

| Category | Setting | $\beta$ Cov.↑ | $\alpha$ Prec.↑ | Avg. KS↓ | Avg. TVD↓ | Corr. Diff.↓ | MI Diff.↓ |
|---|---|---|---|---|---|---|---|
| Train Steps | 3e03 | 0.420 | 0.781 | 0.047 | 0.015 | 1.168 | 4.008 |
| | 6e03 | 0.450 | 0.865 | 0.032 | 0.011 | 1.256 | 3.543 |
| | 9e03 | 0.449 | 0.862 | 0.028 | 0.009 | 1.037 | 3.807 |
| | 1.2e04 | 0.444 | 0.822 | 0.025 | 0.008 | 0.967 | 3.375 |
| | 1.5e04 | 0.449 | 0.852 | 0.026 | 0.008 | 0.979 | 2.894 |
| | 1.8e04 | 0.456 | 0.894 | 0.026 | 0.009 | 0.981 | 3.089 |
| | 2.1e04 | 0.452 | 0.891 | 0.026 | 0.008 | 1.102 | 2.337 |
| | 2.4e04 | 0.462 | 0.936 | 0.026 | 0.008 | 0.992 | 2.010 |
| Architecture | 512x1 | 0.443 | 0.845 | 0.026 | 0.008 | 1.016 | 3.624 |
| | 1024x1 | 0.454 | 0.894 | 0.025 | 0.008 | 0.991 | 2.997 |
| | 512x2 | 0.452 | 0.908 | 0.026 | 0.008 | 0.994 | 4.126 |
| | 1024x2 | 0.462 | 0.936 | 0.026 | 0.008 | 0.992 | 2.010 |
| | 512x3 | 0.448 | 0.881 | 0.025 | 0.008 | 1.055 | 2.956 |
| | 1024x3 | 0.456 | 0.892 | 0.024 | 0.008 | 1.005 | 2.162 |
| Diffusion Steps | 5 | 0.102 | 0.242 | 0.308 | 0.087 | 2.437 | 5.748 |
| | 50 | 0.462 | 0.936 | 0.026 | 0.008 | 0.992 | 2.010 |
| | 300 | 0.454 | 0.891 | 0.026 | 0.008 | 1.062 | 2.888 |
| Train size (Equal Synth.) | 5e03 | 0.391 | 0.732 | 0.048 | 0.010 | 1.243 | 3.951 |
| | 1e04 | 0.462 | 0.936 | 0.026 | 0.008 | 0.992 | 2.010 |
| | 2e04 | 0.476 | 0.921 | 0.016 | 0.008 | 0.687 | 2.940 |
| Batch size | 2048 | 0.467 | 0.895 | 0.026 | 0.011 | 1.460 | 4.547 |
| | 4096 | 0.462 | 0.936 | 0.026 | 0.008 | 0.992 | 2.010 |
| | 8192 | 0.432 | 0.834 | 0.028 | 0.008 | 1.255 | 3.946 |
| Synth. size (1e04 Train) | 5e03 | 0.419 | 0.764 | 0.026 | 0.011 | 1.021 | 4.827 |
| | 1e04 | 0.462 | 0.936 | 0.026 | 0.008 | 0.992 | 2.010 |
| | 2e04 | 0.452 | 0.885 | 0.025 | 0.006 | 0.962 | 2.009 |

## G Dataset Preprocessing

Prior to training the ClavaDDPM target models in the experiments, both datasets undergo light preprocessing steps. Any columns corresponding to unique identifiers are dropped. Missing values for categorical columns are treated as a distinct category. For Diabetes, some categorical columns also have entries of "NaN" and "?". These values are similarly treated as distinct categories for the columns in which they appear. For numerical columns in Berka, missing values are imputed as $0$, while no missing numerical values are present for the Diabetes data. Finally, categorical values are ordinally encoded for uniform representation with encoding performed over the entire dataset to avoid issues with less frequent values. For the TabDiff experiments, the Diabetes dataset is preprocessed in an identical way to the ClavaDDPM experiments. For Berka, we use the preprocessed dataset from (Mahyar et al., 2026).

When computing the metrics discussed in Appendix F and DCR, preprocessing differed depending on the metric computed. For $\alpha$-precision and $\beta$-coverage, categorical values are one-hot encoded and numerical columns are left unchanged. When computing KS and TVD statistics or Correlation and MI matrix differences, the SynthEval (Lautrup et al., 2024) preprocessing approach is applied before downstream computations, which leaves categorical columns ordinally encoded and min-max transforms numerical values. For the MLE quality measures, min-max transforms are also applied to numerical values, but categorical columns are one-hot encoded. Finally, for DCR, categorical variables are one-hot encoded and numerical columns are normalized to the interval $[-1, 1]$.

## H Compute Resources

Generally, the experiments conducted in this work are quite computationally intensive. In addition to training the target models, both the TF and Ensemble attacks require training numerous shadow models, each of which is a large tabular diffusion model itself. In the default setting for Berka, the TF MIA trains $20$ such models, while the Ensemble approach constructs $17$, one primary shadow model and $16$ RMIA models. Both attacks also need to train membership classifiers and require non-trivial feature extraction computations. For the TF attack, this involves extracting loss values via forward and backward diffusion processes for each data point for $2100$ $(\epsilon, t)$ pairs for ClavaDDPM and $5000$ such pairs for TabDiff. In the Ensemble MIA, computationally intensive distance-, DOMIAS-, and RMIA-based features need to be derived.

The experiments in this work leverage both A40 and A100 GPUs with 48GB and 80GB of GPU memory, respectively. Depending on the experimental setup, end-to-end attack training and inference times vary. For example, attacks applied to larger models require more time compared to the default settings. However, a single default ClavaDDPM configuration for the TF approach requires approximately $2.5$ hours for Berka and $1$ hour for the Diabetes dataset on A40 GPUs. For an individual setup, the Ensemble attack takes around $16$ hours to complete for Berka and $11$ hours to complete for Diabetes using A100s. This increases to $22$ and $16$ hours, respectively, on A40 GPUs.