Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29436

TableDiffusion: A Diffusion-Based Framework for Differentially Private Mixed-Type Tabular Data Synthesis

Zichen Xu

Department of Computer Science, University of Minnesota – Twin Cities, Minneapolis, USA
zichenxu1017@hotmail.com

Abstract. With growing concerns about data privacy, differential privacy (DP) has become a central technique in privacy-preserving data sharing and machine learning. While many DP generative models have been developed for continuous data such as images, generating synthetic tabular data—especially with mixed data types—remains a challenging task. In this work, we present TableDiffusion, the first diffusion-based generative model specifically designed for differentially private tabular data synthesis. TableDiffusion introduces a hybrid data generation framework capable of handling both continuous and categorical features. It uses a cosine noise scheduler to control the diffusion process and adopts a composite loss function that combines mean squared error (MSE) and Kullback-Leibler (KL) divergence to better adapt to different data types. The model is implemented using the Opacus library to provide formal DP guarantees, and privacy accounting is performed through the Rényi Differential Privacy (RDP) mechanism. To assess the effectiveness of TableDiffusion, we implemented a set of benchmark DP generative models, including attention-based methods such as DPattentionVAE and DPattentionGAN, as well as classical approaches such as PATE-GAN and DP-WGAN. Experiments are conducted on a real-world tabular dataset in the cybersecurity domain, containing more than 50,000 samples and 45 features. We evaluated model performance using metrics such as marginal distribution similarity, PMSE ratio, and Alpha-Beta membership inference scores, focusing on both data utility and privacy preservation. The experimental results are currently being analyzed. Preliminary findings suggest that TableDiffusion offers promising performance in a range of privacy budgets and shows advantages in training stability and diversity over GAN-based methods. Our main contributions include: (1) proposing the first diffusion-based method for differentially private tabular data generation; (2) designing a unified architecture for mixed-type feature synthesis; and (3) providing a comprehensive implementation and evaluation framework to facilitate further research.

Keywords: Differential Privacy, Diffusion Models, Tabular Data Synthesis, Mixed Data Types, Privacy-Preserving Machine Learning

1. Introduction

In recent years, the demand for privacy-preserving data synthesis has grown significantly, driven by the increasing need to share and analyze sensitive data without compromising individual privacy.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29436

This demand is particularly pressing in fields such as healthcare, finance and cybersecurity, where regulatory restrictions prohibit direct access to real-world datasets. While generative models such as GANs and VAEs have made notable progress in synthetic data generation (e.g., VAEs [1]), generating high-fidelity tabular data under rigorous differential privacy (DP) constraints remains a considerable challenge due to the structured, heterogeneous nature of tabular data [2]. Moreover, differential privacy has also been pursued via teacher–student knowledge transfer—most notably PATE, which aggregates votes from private teacher models to train a public student model [3].

This paper introduces dpgai, a modular and extensible framework designed to generate synthetic tabular datasets with formal differential privacy guarantees. The framework leverages several state-of-tive architectures, including attention-based GANs, variational autoencoders (VAEs), and diffusion models [4] such as TableDiffusion, the first diffusion-based sthe-art generaynthesizer tailored for tabular data with DP integration.

dpgai incorporates privacy-preserving mechanisms such as DP-SGD [2] and the PATE framework into the training process, ensuring that generated datasets retain the statistical utility of the original while protecting sensitive information. Furthermore, the framework provides a rich suite of utilities for data set transformation, privacy accounting, fidelity evaluation, and visual comparison, making it a practical tool for both researchers and data practitioners.

Through extensive experiments on real-world tabular datasets, we demonstrate that dpgai not only preserves data utility but also robustly meets differential privacy requirements. Our work highlights the potential of integrating diffusion-based models into privacy-aware synthetic data generation and provides a foundation for secure data sharing and analysis across a wide range of domains.

2. Method

2.1. Problem definition

We consider a tabular dataset D comprising n samples and d characteristics, denoted as $D=\{x_1,x_2,\ldots,x_n\}$, where each data point x_i is a vector of dimensions d . The tabular data often contains mixed data types, including continuous features $x^{(c)}\in\mathbb{R}^{d_c}$ and categorical features $x^{(cat)}\in\mathbb{Z}^{d_{cat}}$, where $d=d_c+d_{cat}$. Our objective is to learn a generative model G_θ that can produce synthetic data $\tilde{x}$, which closely approximates the distribution of the original dataset D while satisfying (ϵ,δ) -differential privacy.

2.2. TableDiffusion model architecture

2.2.1. Diffusion process design

TableDiffusion is based on diffusion probabilistic models [4] , which involve two key phases: a forward (diffusion) process and a reverse (denoising) process.

Forward Process: Starting from an original clean input x_0 , Gaussian noise is progressively added over T time steps:

$$q(x_t|x_{t-1})=N(\sqrt{1-\beta_t}\cdot x_{t-1},\beta_t I) \tag{1}$$

The variance schedule β_t is defined by a cosine-based function following [5] :

DOI: 10.54254/2755-2721/2026.TJ29436

$$\beta_t = \frac{1-\cos\left(\pi \cdot \frac{t}{T}\right)}{2} + 0.1 \tag{2}$$

Reverse Process: A neural network ϵ_θ predicts the added noise, allowing the denoising distribution to be approximated as:

$$p_\theta(x_{t-1}|x_t) = N(\mu_\theta(x_t, t), \sigma_t^2 I) \tag{3}$$

2.2.2. Mixed-Type Generator

The MixedTypeGenerator neural network uses an MLP architecture. Continuous features are handled by direct regression, while categorical features are passed through softmax layers based on category indices.

2.2.3. Loss function design

We define a composite loss function for mixed data types:

Numerical feature loss (MSE):

$$\mathcal{L}_{num} = \frac{1}{d_c} \sum \|x_{0_i} - \hat{x}_{0_i}\|^2 \tag{4}$$

Categorical feature loss (KL divergence):

$$\mathcal{L}_{cat} = \frac{1}{d_{cat}} \sum KL(\log(\hat{x}_{0_j}), x_{0_j}) \tag{5}$$

Total loss:

$$\mathcal{L}_{total} = \mathcal{L}_{num} + \mathcal{L}_{cat} \tag{6}$$

2.3. Data preprocessing

2.3.1. Continuous features

Continuous columns are normalized using quantile transformation to reduce outlier sensitivity and maintain rank order.

2.3.2. Categorical features

Categorical variables are transformed using one-hot encoding with safeguards for unknown values during inference.

2.4. Baseline model implementations

To benchmark performance, we compare TableDiffusion with several baselines:

- DPattentionVAE: A VAE with attention modules.
- DPattentionGAN: A GAN with attention mechanisms.
- PATE-GAN: A GAN using teacher aggregation [5] .
- DP-WGAN: A differentially private Wasserstein GAN.
- DP-autoGAN: A hybrid GAN-autoencoder model.

DOI: 10.54254/2755-2721/2026.TJ29436

2.5. Sampling procedure

Synthetic samples are generated by reversing the diffusion process from pure noise. If noise prediction is enabled, the model estimates noise directly and subtracts it at each time step.

2.6. Post-processing

Generated data are converted back to human-readable form through inverse transformations:

- Inverse quantile transformation for continuous features
- Inverse one-hot decoding for categorical features

This ensures schema compatibility and usability for downstream tasks.

3. Experiments

3.1. Experimental setup

To evaluate the effectiveness of TableDiffusion, we conducted extensive experiments on real-world tabular datasets under differential privacy constraints. The experiments were designed to test data utility, privacy preservation, and model robustness across a wide range of privacy budgets and baseline models.

Dataset. We use a preprocessed variant of the UCI Adult dataset [6] , a widely adopted benchmark in tabular data synthesis. The dataset includes 26 features encompassing both continuous (e.g., id, dur, rate, sload, dload) and categorical (e.g., proto, service, state, attack_cat, label) types. Continuous features are quantile-transformed while categorical features are one-hot encoded.

Privacy Settings. We explore a range of privacy budgets $\epsilon \in \{0.5,1.0,5.0,10.0,50.0,100.0,500.0,1000.0\}$, with δ fixed at 10^{-5} . Differential privacy is enforced using DP-SGD via the Opacus library [2] , and cumulative privacy loss is tracked using Rényi Differential Privacy (RDP) accounting.

3.2. Baseline models

We compare TableDiffusion against the following differentially private generative models:

- DPattentionVAE: Variational autoencoder with attention layers
- DPattentionGAN: Attention-based GAN
- DPautoGAN: Hybrid autoencoder-GAN architecture
- DP-WGAN: Wasserstein GAN with differential privacy
- PATE-GAN: Teacher-student GAN with private aggregation [5]
- SAINT-AE: Tabular transformer with autoencoding [7]
- TableDiffusion (ours): Diffusion-based model with mixed-type support [4,8]

Each model is trained for 5 epochs and evaluated across 10 independent runs to ensure statistical robustness.

3.3. Metrics

Data Utility. We assess data utility from multiple perspectives:

- PMSE Ratio: Measures classifier generalization from synthetic to real data
- Marginal Distribution Similarity: Includes statistical tests such as Chi-squared, Kolmogorov-Smirnov, and Wasserstein distance

DOI: 10.54254/2755-2721/2026.TJ29436

• Alpha-Beta Metrics: Evaluate synthetic fidelity via α -precision and β -recall [9]

Privacy Risk. Privacy protection is evaluated using:

• Privacy Budget Tracking: Real-time accounting of ϵ throughout training

• Singling-out Risk Assessment (SRA): Measures re-identification risks at record level

3.4. Implementation details

Each model is trained with a batch size of $512 - 1024$, depending on architecture. TableDiffusion-specific configuration is as follows:

```
{
"batch_size": 1024,
"lr": 0.005, "dims": (128, 128),
"diffusion_steps": 5,
"epsilon_target": 1.0,
"max_grad_norm": 1.0,
"delta": 1e-5
}
```

A cosine-based noise scheduler is used in the forward process:

$$\beta_t = \frac{1-\cos\left(\frac{\pi t}{T}\right)}{2} + 0.1 \tag{6}$$

Generated synthetic data ($\sim$ 500 samples per run) are post-processed via inverse quantile and one-hot decoding for usability in downstream ML tasks.

3.5. Results and analysis

• Total Experiments: 263 unique training/evaluation runs

• Performance: TableDiffusion consistently outperforms GAN-based models in training stability, fidelity under small ϵ , and diversity of generated samples

• Ablation Studies: Experiments varying diffusion steps (3–5) and loss weights show tradeoffs between sample fidelity and privacy leakage

Preliminary analysis shows that TableDiffusion achieves a superior balance between data utility and privacy preservation, especially under strict budgets ($\epsilon \leq 1.0$). In contrast, traditional GANs are more prone to overfitting and training instability.

4. Conclusion

This work presents TableDiffusion, the first diffusion-based generative model tailored for differentially private synthesis of mixed-type tabular data. By extending the diffusion paradigm beyond continuous domains like images [4] , TableDiffusion effectively handles both numerical and categorical features through a hybrid generation mechanism and a cosine noise scheduler [8] , addressing longstanding challenges in heterogeneous table modeling.

We developed a comprehensive evaluation framework encompassing seven baseline models and multiple metrics, including PMSE ratio, Wasserstein distance, Alpha-Beta certification, and individual privacy risk. Experimental results across a wide range of privacy budgets ($\epsilon \in [0.5, 1000.0]$) show that TableDiffusion achieves superior synthetic data quality and strong privacy protection, particularly for complex mixed-type data. Technically, TableDiffusion leverages

DOI: 10.54254/2755-2721/2026.TJ29436

a progressive denoising process, which not only stabilizes training compared to adversarial methods but also facilitates the integration of differential privacy mechanisms [2] .

$$\beta(t,T)=\frac{1-\cos(\pi\, t/T)}{2}+0.1 \tag{7}$$

Our cosine noise scheduler provides an effective trade-off between training efficiency and generation quality. A reproducible implementation with MLflow tracking is made publicly available to foster future research.

Nonetheless, certain limitations remain. The current implementation is optimized for medium-scale datasets, and its scalability to very large datasets requires further investigation. Additionally, while results on standard datasets such as UCI Adult are promising [6] , broader evaluations across diverse real-world datasets are needed to assess generalization.

Future work includes (1) designing more efficient diffusion schedules to reduce inference time and computational cost, (2) developing adaptive privacy budgeting mechanisms for finer control, (3) integrating with federated learning for distributed private synthesis, and (4) extending to time-series and hierarchical tabular structures.

In summary, TableDiffusion represents a significant step forward in differentially private tabular data generation. By combining the generative strength of diffusion models with rigorous privacy guarantees, it opens new possibilities for safe and responsible data sharing in fields such as healthcare, finance, and social science.

References

[1] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv preprint arXiv: 1312.6114. https: //arxiv.org/abs/1312.6114

[2] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318. https: //doi.org/10.1145/2976749.2978318

[3] Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I., & Talwar, K. (2017). Semi-supervised knowledge transfer for deep learning from private training data. arXiv preprint arXiv: 1610.05755. https: //arxiv.org/abs/1610.05755

[4] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851. https: //arxiv.org/abs/2006.11239

[5] Jordon, J., Yoon, J., & van der Schaar, M. (2018). PATE-GAN: Generating synthetic data with differential privacy guarantees. International Conference on Learning Representations. https: //openreview.net/forum?id=S1zk9iRqF7

[6] Dua, D., & Graff, C. (2017). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. http: //archive.ics.uci.edu/ml

[7] Somepalli, G., Cherepanova, V., Goldblum, M., Schwarzschild, A., & Goldstein, T. (2021). SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. arXiv preprint arXiv: 2106.01342. https: //arxiv.org/abs/2106.01342

[8] Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. Proceedings of the 38th International Conference on Machine Learning, 8162–8171. https: //arxiv.org/abs/2102.09672

[9] Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., & Aila, T. (2019). Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems, 32. https: //arxiv.org/abs/1904.06991