Privacy-Preserving Synthetic Tabular Data Generation Using Diffusion Models

SEDS500 Graduation Project

Umut Akın

Izmir Institute of Technology
Graduate School of Engineering and Sciences
Department of Computer Engineering

January 2026

Abstract

Research Summary

This project investigates **diffusion models** as a privacy-preserving approach for generating synthetic tabular data.

Method: TabDDPM-style diffusion with hybrid Gaussian-Multinomial noise

Comparison: Against CTGAN (GAN-based) and SMOGN (interpolation-based)

Key Results:

- **87–98%** of baseline model performance with synthetic data alone
- Significantly outperforms CTGAN (35%) and SMOGN (fails completely)
- **Zero privacy leakage** (membership inference AUC = 0.51)

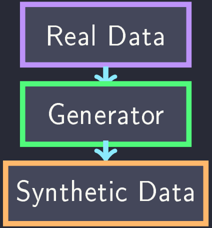
Conclusion: Diffusion models are superior for generating high-utility, privacy-preserving synthetic tabular data.

Motivation & Problem Definition

The Challenge:

- Organizations need to share data for ML collaboration
- Privacy regulations (GDPR, KVKK) restrict data sharing
- Traditional methods (GANs, interpolation) struggle with mixed data types

Solution: Synthetic Data



Research Question

Do diffusion models produce more realistic synthetic tabular data than traditional methods?

Evaluation Criteria:

1. **Utility**: ML performance on real data
2. **Privacy**: No information leakage

Proposed Solution

Implement TabDDPM-style diffusion for tabular data generation

Key innovations:

- Hybrid noise handling
- Gaussian for numerical features
- Multinomial for categorical features
- Log-space operations for stability
- KL divergence loss

Methods compared:

- **TabDDPM-style** (ours)
- CTGAN (GAN-based)
- SMOGN (interpolation)

Datasets:

- Production (5,370 samples)
- Ozel Rich (2,670 samples)

Related Research

Paper	Venue	Approach	Key Innovation
TabDDPM	ICML 2023	Diffusion	Hybrid Gaussian-Multinomial noise
CTGAN	NeurIPS 2019	GAN	Mode-specific normalization
STaSy	ICLR 2023	Score-based	Self-paced learning
TabSyn	ICLR 2024	Latent diffusion	Transformer VAE encoder

Why diffusion over GANs?

- GANs suffer from mode collapse and training instability
- Diffusion models have stable training dynamics
- Iterative refinement captures full data distribution

Our contribution: Implement and evaluate TabDDPM-style diffusion on real organizational datasets with privacy validation.

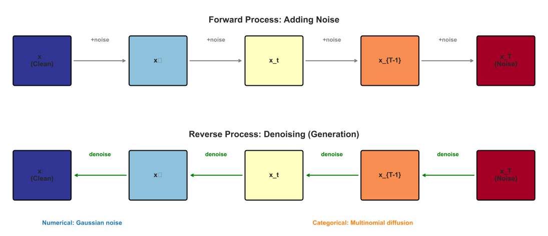
Solution: TabDDPM Diffusion Model

How Diffusion Works:

- **Forward**: Gradually add noise (T=1000 steps)
- **Reverse**: Learn to denoise step-by-step
- Neural network: noise $\rightarrow$ realistic data

TabDDPM Hybrid Approach:

- **Numerical**: Gaussian diffusion
- **Categorical**: Multinomial diffusion
- Process both simultaneously



Iterative denoising from random noise

Key Advantage

Stable training + captures full distribution (no mode collapse like GANs)

Datasets & Experimental Setup

Two real-world organizational datasets (Turkish fastener company):

Dataset	Domain	Samples	Features	Target
Production	Sales quotation	5,370	7 num + 35 cat	Quote amount
Ozel Rich	Custom mfg	2,670	2 num + 4 cat	Machine time

Evaluation scenarios:

1. **Replacement**: Train on synthetic only, test on real data
2. **Augmentation**: Train on real + synthetic, test on real data

Metrics: R^2 score (utility), MIA AUC (privacy)

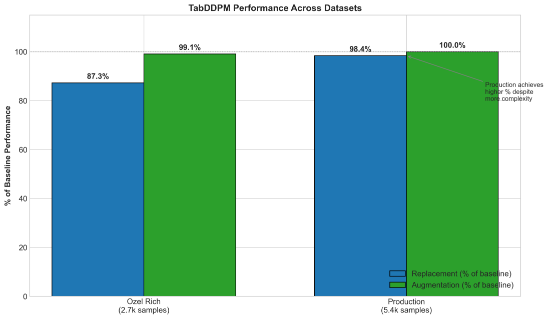
Results: Utility Comparison

Ozel Rich Dataset (Replacement scenario)

Method	R^2	%
Baseline	0.645	100%
SMOGN	-0.14	FAILED
CTGAN	0.229	35.5%
TabDDPM	**0.563**	**87.3%**

Production Dataset

Scenario	R^2	%
Replacement	0.979	**98.4%**
Augmentation	0.994	**100%**



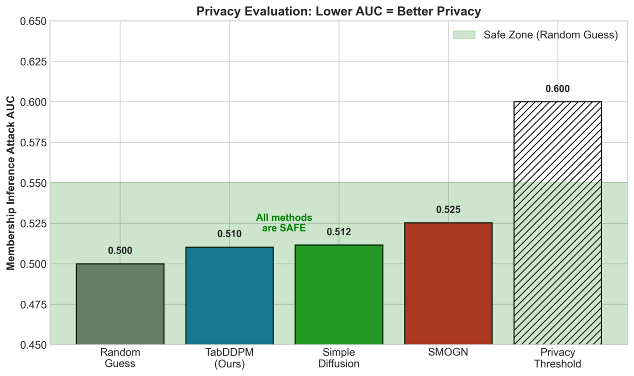
Key findings:

- TabDDPM: **87–98%** of baseline
- CTGAN: 35% | SMOGN: Failed
- Generalizes across datasets

Results: Privacy Evaluation

Membership Inference Attack: Can attacker identify training records?

Method	AUC	Status
Random	0.50	–
TabDDPM	**0.51**	SAFE
Simple Diff	0.51	SAFE
SMOGN	0.53	SAFE



AUC $\approx$ 0.5 = Random guessing = No privacy leakage

Key Result

TabDDPM: **Highest utility** (87–98%) + **Excellent privacy** (AUC = 0.51)

Conclusion & Future Work

Key Findings:

- TabDDPM: **87–98%** utility vs 35% (CTGAN)
- Privacy-safe: MIA AUC = 0.51 (random guessing)
- SMOGN fails on complex tabular data
- Generalizes across different datasets

Main Conclusion

Diffusion models are superior for privacy-preserving synthetic data

Limitations:

- 2 organizational datasets
- Basic privacy evaluation

Future Work:

- TabSyn (latent diffusion)
- Differential privacy
- Public benchmarks

Applications:

- Safe data sharing
- GDPR/KVKK compliance

Thank You

Questions?

Umut Akın
Izmir Institute of Technology
SEDS500 Graduation Project
January 2026