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

Outline

1 Abstract

2 Introduction

3 Related Research

4 Solution Approach

5 Validation Approach

6 Conclusion & Future Work

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: Data Sharing vs Privacy

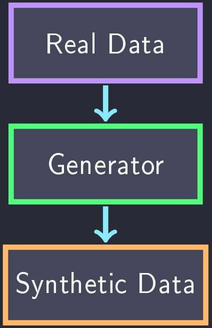
Organizations want to:

- Share data with partners
- Enable ML research
- Collaborate across teams

But they face:

- Privacy regulations (GDPR, KVKK)
- Sensitive customer data
- Competitive concerns

Solution: Synthetic Data



Same statistical properties,
no original records exposed

Problem Definition & Goal

Problem

Traditional synthetic data methods (interpolation, GANs) struggle with complex tabular data containing mixed numerical and categorical features.

Research Question

When generating synthetic tabular data for privacy purposes,
do diffusion models produce more realistic data
than traditional methods?

Evaluation Criteria:

1. **Utility**: Can ML models trained on synthetic data perform well on real data?
2. **Privacy**: Does the synthetic data leak information about training records?

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.

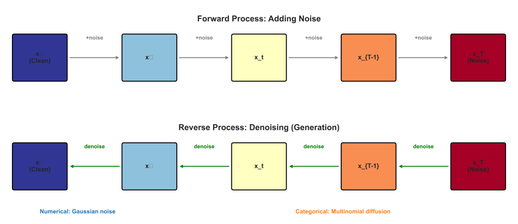
Diffusion Models: Core Idea

Forward Process (Training):

- Gradually add noise to data
- Over $T = 1000$ timesteps
- Data becomes pure noise

Reverse Process (Generation):

- Learn to denoise step-by-step
- Neural network predicts noise
- Random noise $\rightarrow$ realistic data



Key advantage: Stable training, captures full distribution (no mode collapse)

TabDDPM: Handling Mixed Data Types

Challenge: Tabular data has both numerical and categorical features

Numerical Features:

- Standard Gaussian diffusion
- Add/remove continuous noise
- Example: price, quantity

Categorical Features:

- Multinomial diffusion
- Transition between categories
- Example: product type, material

Hybrid Approach

Process numerical and categorical features simultaneously
with type-appropriate noise schedules

Implementation Details

Key improvements over simple diffusion (26% → 87%):

Improvement	Problem Solved	Impact
Log-space operations	Probability underflow	Prevents NaN/Inf
KL divergence loss	Wrong loss for categories	Learns distributions
Gumbel-softmax	Non-differentiable argmax	Enables gradients
Proper posterior	Incorrect reverse process	Faithful reconstruction

Technical setup:

- Framework: PyTorch | Hardware: NVIDIA RTX 4070 Ti Super (16GB)
- Training: 1000 epochs, batch size 128, LR 10^{-4}, cosine schedule

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:

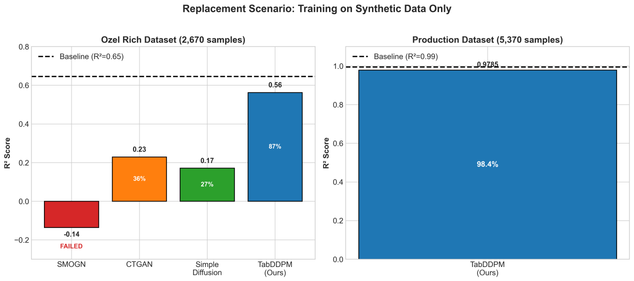
1. **Replacement**: Train on synthetic only, test on real data
2. **Augmentation**: Train on real + synthetic, test on real data

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

Results: Replacement Scenario

Train on synthetic data only, evaluate on real test data

Method	R^2	%
Baseline	0.645	100%
SMOGN	-0.14	FAILED
CTGAN	0.229	35.5%
Simple Diff	0.171	26.5%
TabDDPM	**0.563**	**87.3%**



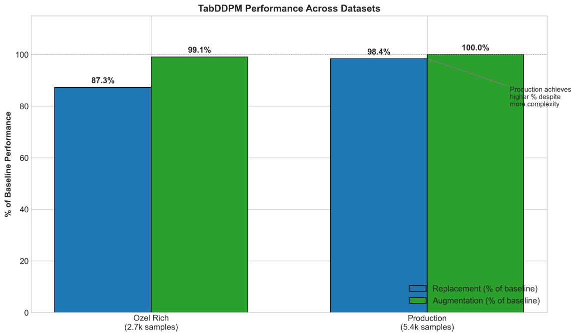
Key findings:

- TabDDPM: 2.5× better than CTGAN
- SMOGN: Catastrophic failure

Results: Production Dataset

Larger, more complex dataset (5,370 samples, 117 features after encoding)

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



Cross-dataset comparison:

- Ozel Rich: 87.3% of baseline
- Production: **98.4%** of baseline

Insight

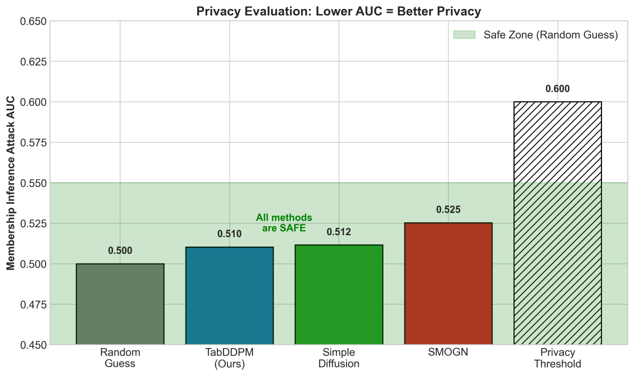
TabDDPM generalizes well to larger, more complex datasets

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)

Discussion: Why Methods Differ

Why TabDDPM succeeds:

- Stable training dynamics
- Learns full distribution
- Iterative refinement
- Preserves rare patterns

Why CTGAN is moderate:

- Mode collapse risk
- Adversarial instability
- May miss rare samples

Why SMOGN fails:

- Interpolation in high dimensions
- Cannot handle categoricals
- Creates unrealistic combinations

Critical Finding

SMOGN is not just "less effective"
but **actively harmful** on complex data
(corrupts training, R^2 goes negative)

Key Findings

Finding	Evidence
TabDDPM achieves highest utility	87–98% vs 35% (CTGAN)
Generalizes across datasets	Ozel: 87%, Production: 98%
Diffusion is privacy-safe	MIA AUC $= 0.51$ (random guessing)
SMOGN fails on complex data	Negative R^2
TabDDPM improvements essential	$3.3\times$ better than simple diffusion

Main Conclusion

Diffusion models are superior for privacy-preserving synthetic tabular data generation

Limitations & Future Work

Limitations:

- Evaluated on 2 organizational datasets (not public benchmarks)
- CTGAN used default hyperparameters
- Basic MIA (no shadow models)
- Slower generation than GANs

Future Work:

- TabSyn (latent diffusion)
- Differential privacy integration
- Standard benchmarks (Adult, Covertype)
- Web interface for practitioners
- Conditional generation

Practical applications:

1. Share data safely with partners (87–98% utility)
2. Enable ML collaboration without exposing real records
3. Comply with privacy regulations (GDPR, KVKK)

Thank You

Questions?

Umut Akın
Izmir Institute of Technology
SEDS500 Graduation Project
January 2026