A Comparative Study on Diffusion Models for Tabular Data Synthesis in Healthcare

Neetu Kumari and Enayat Rajabi
Shannon School of Business, CBU, Nova Scotia, Canada

INTRODUCTION

BACKGROUND:

Synthetic data generation helps overcome critical challenges in healthcare:

- **Data scarcity** for rare events or arbitrary events
- **Cost** and Time Efficiency
- **Privacy** and Confidentiality

Despite its growing use, there is a notable **lack of comparative** studies on the latest diffusion models for tabular datasets in healthcare.

RESEARCH OBJECTIVE:

This study compares two advanced diffusion models, **TabDDPM** and **TabSyn**, on healthcare datasets. It assesses their performance based on:

- **Data Similarity**: Checks how closely synthetic data matches the original.
- **Utility for Machine Learning**: Tests effectiveness in ML applications.
- **Privacy Preservation**: Ensures the synthetic data maintains confidentiality.

DATASET

The Obesity and Diabetes datasets were selected from the UCI Machine Learning Repository for several reasons including **Variability, Sensitivity, Entry Volume Diversity, Feature Diversity.**

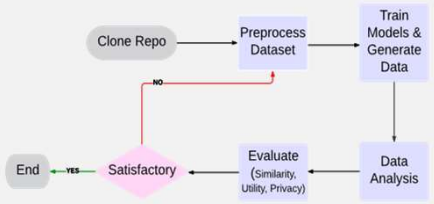
Statistics of datasets used in the study # - Number

Dataset	#Entries	#Num	#Cat	#Train	#Test	Task
Obesity	2111	8	9	1899	212	MultiClass
Diabetes	253680	7	15	228312	25368	BiClass

METHODS

GENERAL APPROACH:

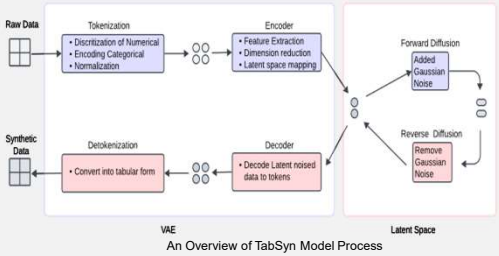
- Training, generating artificial data, and evaluation



Flowchart of the process of the study

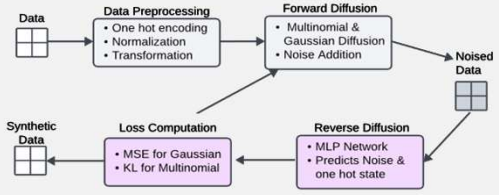
The following are the Models used in the study:

1. TabSYN:



An Overview of TabSyn Model Process

2. TabDDPM:



An Overview of TabDDPM Model Process

MLP - multi-layer perceptron
MSE – Mean Squared Error

EVALUATION METRICES:

Data Similarity	Utility	Privacy Preservation
• Variable Correlation • Distribution Similarity • Pair-wise Correlation	• TSTR Method in Ml Models	• Distance to Closest Record (DCR) • Alpha - Precision • Beta Recall

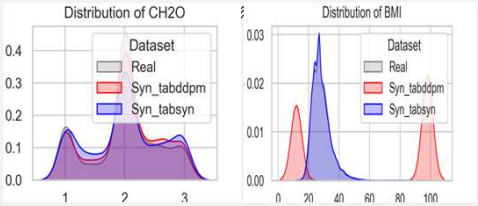
TSTR (Train Synthetic Test Real)

RESULTS

This research assesses following key metrics to maintain privacy while optimizing data utility:

SIMILARITY EVALUATION:

- Variable Correlation: Statistical similarities for continuous (Mean & Median) and categorical (Ratio of categories)
- Distribution and Pair-wise Correlation: Kolmogorov-Smirnov(KS) Test, Chi-square test, and analyzes



[Left] CH2O Distribution: "Obesity" dataset via TabDDPM & TabSyn Models
[Right] BMI Distribution: "Diabetes" dataset via TabDDPM & TabSyn Models

Comparative table of OQS, Column Shapes and Column pair Trends

	Obesity		Diabetes	
Metrices/Models	*TabDDPM*	*TabSyn*	*TabDDPM*	*TabSyn*
Overall Quality Score	**95.36**	94.68	62.25	**97.94**
Column Shapes	96.37	**96.84**	72.04	**98.51**
Column Pair Trends	**94.35**	92.52	52.46	**97.36**

Note:
- **Obesity: No notable difference.**
- **Diabetes: TabSyn outperforms TabDDPM.**

UTILITY EVALUATION:

- Machine learning usability:

TSTR (Training-Set Test-Set) approach.

AUC score by using XGB Classifier (Higher the Score, Better performance)

Dataset\Model	TabDDPM	TabSyn
Obesity	**0.9981**	0.9962
Diabetes	0.6773	**0.8275**

PRIVACY PRESERVATION:

- Alpha-Precision & Beta-Recall Metrics
- Distance to Closest Record (DCR)

Comparing Alpha-Precision and Beta-Recall values for both models
Alpha value : Fidelity & Beta value : Diversity

	Obesity		Diabetes	
Metrices/Models	*TabDDPM*	*TabSyn*	*TabDDPM*	*TabSyn*
Alpha-Precision	0.897	**0.975**	0.655	**0.978**
Beta-Recall	**0.380**	0.304	0.0003	**0.566**

DCR Scores: Datasets generated by using TabDDPM and TabSyn

Metrices/Models	*Obesity*	*Diabetes*
TabDDPM	0.93	0.89
TabSyn	**0.92**	**0.87**

DISCUSSION

- **Smaller** Dataset (Obesity): **No significant** differences between model's performance on all metrices.
- **Larger** Dataset (Diabetes): *TabSyn* **outperforms** *TabDDPM* significantly on all evaluation metrices.
- TabSyn is well-suited for synthetic healthcare data,

FUTURE DIRECTION

- **Enhance** generative models to more effectively produce high-dimensional, small healthcare datasets.
- **Differential Privacy Preservation:**

1. **Re-identification Risk:** Evaluating the traceability of synthetic data back to original data.
2. **Membership Inference Attacks:** Determine if an individual's data was used to create the synthetic dataset.