How do diffusion-based tabular generators compare to GANs in preserving high-order feature correlations under

Assignee Research

June 10, 2026

Abstract

Tabular data is one of the most prevalent and important data formats in real-world applications such as healthcare, finance, and education. However, its effective use in machine learning is often constrained by data scarcity, privacy concerns, and class imbalance. Synthetic tabular data generation has emerged as a powerful solution, leveraging generative models to learn underlying data distributions and produce realistic, privacy-preserving samples. Although this area has seen growing attention, most existing surveys focus narrowly on specific methods (e.g., GANs or privacy-enhancing technique

1 Introduction

This paper examines: A Comprehensive Survey of Synthetic Tabular Data Generation. Research question: How do diffusion-based tabular generators compare to GANs in preserving high-order feature correlations under differential privacy constraints?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

12 papers retrieved. 18 claims extracted; 2 independently verified. Quality review score: 4.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
TabDDPM and TABDIFF solve heterogeneity problems in tabular data generation through multimodal diffusion.	×	0.11
CLLM leverages the internal knowledge of Large Language Models (LLMs) to improve generation quality in limited data scen	×	0.12
DP-LLMTGen, DPGAN, and PATE-GAN enhance privacy protection by incorporating differentially private strategies.	×	0.03
Existing surveys [20, 21, 22] focus specifically on GANs for generating medical tabular data.	×	0.10
Existing surveys [23, 24] focus on class imbalance, privacy, and fairness in synthetic tabular data generation.	✓	0.19
Synthetic data may match the original distribution but still violate commonsense or logical rules, such as negative age	×	0.06
Diffusion models offer advantages over traditional generative methods regarding stability and representation capabilitie	×	0.08
LLMs offer advantages over traditional generative methods regarding their ability to model semantic structure.	×	0.06
The survey categorizes existing approaches into three categories: traditional generation methods, diffusion model method	✓	0.28
The survey classifies post-processing techniques into two primary categories: sample enhancement and label enhancement.	×	0.06
CLLM was published in 2024 at ICML.	×	0.02
LITO was published in 2024 at NeurIPS.	×	0.02
EPIC was published in 2024 at ICLR.	×	0.02
CLLM utilizes GPT-2 and GPT-3.5 models.	×	0.02
LITO utilizes GPT-3.5 and Mistral-7b-v0.1 models.	×	0.01
EPIC utilizes DistilGPT-2 and GPT-2 models.	×	0.01
Tabular data plays a crucial role in domains such as healthcare, finance, education, transportation, and psychology.	×	0.10
GDPR and CCPA are data privacy regulations affecting tabular data usage.	×	0.07

References

- http://arxiv.org/abs/2507.19211v1
- http://arxiv.org/abs/2504.16506v3
- http://arxiv.org/abs/2205.11090v1