Evaluating Tabular Data Generation Metrics on Large-Scale Mixed-Type Datasets

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How do the proposed evaluation metrics for tabular data generation perform on mixed-type datasets compared to existing metrics like KSD or KID in terms of computational efficiency and accuracy when. 6 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating Generative Models for Tabular Data: Novel Metrics and Benchmarking. Research question: How do the proposed evaluation metrics for tabular data generation perform on mixed-type datasets compared to existing metrics like KSD or KID in terms of computational efficiency and accuracy when scaled to datasets with over 1M samples?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.5/10.

3 Results

13 papers retrieved. 6 claims extracted; 3 independently verified. Quality review score: 6.5/10.

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
FAED effectively captures generative modeling issues overlooked by existing metrics.	✓	0.28
FPCAD exhibits promising performance but requires further refinements to enhance its reliability.	✓	0.19
Existing metrics such as SDV Fidelity, Utility, Privacy through DCR, TSTR, and TRTS have been employed in specific scena	×	0.12
FAED successfully detects all synthesized problems, whereas FPCAD shows promising but improvable performance.	×	0.06
Existing metrics fail to identify key issues, underscoring the need for more robust evaluation measures for tabular data	✓	0.16
TSTR is particularly useful for detecting cases where synthetic data only partially represents real data, while TRTS ass	×	0.08

References

- http://arxiv.org/abs/2412.00381v1
- http://arxiv.org/abs/2504.20900v1
- http://arxiv.org/abs/2507.05904v1