Novel Evaluation Metrics for Tabular Generative Models Beyond Inception Score and FID

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How do novel evaluation metrics for tabular generative models like CausalMixFT compare to traditional metrics (e.g., Inception Score, FID) in quantifying structural fidelity and diversity of. 7 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: The Vendi Score: A Diversity Evaluation Metric for Machine Learning. Research question: How do novel evaluation metrics for tabular generative models like CausalMixFT compare to traditional metrics (e.g., Inception Score, FID) in quantifying structural fidelity and diversity of generated samples across heterogeneous datasets?.

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 3.5/10.

3 Results

12 papers retrieved. 7 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
The Vendi Score (vs) identifies some model weaknesses that are not detected by IntDiv.	×	0.12
VS increases proportionally with diversity in three sets of synthetic datasets.	×	0.04
VS behaves consistently and intuitively in all three settings: in each case, VS can be interpreted as the effective numb	×	0.02
VS captures a more fine-grained notion of diversity than number of modes (nom).	×	0.02
Presgan and Self-cond.gan both capture all the 1000 modes.	×	0.01
VS reveals that Presgan is more diverse than Self-cond.gan and that they both are less diverse than the original dataset	×	0.02
VS is able to capture the HMM's lack of diversity while IntDiv cannot.	×	0.03

References

- http://arxiv.org/abs/2504.20900v1
- http://arxiv.org/abs/2207.05295v2
- http://arxiv.org/abs/2210.02410v2