Noise Magnitude Effects on Tabular Foundation Model Calibration in OpenML-CC18

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does varying the noise magnitude in synthetic tabular data generators impact the calibration error of tabular foundation models when evaluated on the OpenML-CC18 benchmark. 17 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.6/10. 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 does varying the noise magnitude in synthetic tabular data generators impact the calibration error of tabular foundation models when evaluated on the OpenML-CC18 benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.6/10.

3 Results

14 papers retrieved. 17 claims extracted; 5 independently verified. Quality review score: 5.6/10.

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 experimental analysis was conducted on three standard network intrusion detection datasets.	✓	0.24
The study compares the proposed metrics with established evaluation methods including Fidelity, Utility, TSTR, and TRTS.	✓	0.21
FAED effectively captures generative modeling issues that are overlooked by existing metrics.	✓	0.30
FPCAD exhibits promising performance but requires further refinements to enhance reliability.	✓	0.20
Recent advancements in diffusion models and transformers have led to success in domains such as images, videos, text, an	×	0.05
Existing metrics such as SDV Fidelity, Utility, Privacy through DCR, TSTR, and TRTS have insufficiently explored robustn	×	0.13
The study introduces three novel evaluation metrics named FAED, FPCAD, and RFIS tailored for tabular data.	✓	0.22
The study simulates three specific challenges in real datasets: Quality Decrease, Mode Drop, and Mode Collapse.	×	0.04
Experimental results show that FAED successfully detects all synthesized problems (Quality Decrease, Mode Drop, and Mode	×	0.04
Existing metrics fail to identify key generative modeling issues in the conducted experiments.	×	0.13
Quantitative metrics like Inception Score (IS) and Frchet Inception Distance (FID) are standard for evaluating generati	×	0.11
TSTR involves training a classifier on synthetic data and testing it on real data.	×	0.06
A high TSTR accuracy suggests that synthetic data effectively approximates real-world distributions.	×	0.07
TRTS involves training a classifier on real data and testing it on synthetic data.	×	0.06
A high TRTS score indicates that the synthetic data retains key characteristics of the real data.	×	0.05
TSTR is particularly useful for detecting cases where synthetic data only partially represents real data.	×	0.04
TRTS assesses whether synthetic samples introduce patterns absent in real data.	×	0.05

References

- http://arxiv.org/abs/2601.04110v2
- http://arxiv.org/abs/2506.16791v4
- http://arxiv.org/abs/2504.20900v1