Calibration Error of Tabular Foundation Models Across Synthetic Data Generators on OpenML-CC18

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does the calibration error of tabular foundation models compare across different synthetic data generators (e.g., CTGAN, TVAE, GANs) when evaluated on the OpenML-CC18 benchmark with varying noise. 13 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: Causal Data Augmentation for Robust Fine-Tuning of Tabular Foundation Models. Research question: How does the calibration error of tabular foundation models compare across different synthetic data generators (e.g., CTGAN, TVAE, GANs) when evaluated on the OpenML-CC18 benchmark with varying noise magnitudes?.

2 Methodology

Systematic literature search across multiple databases yielded 8 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

8 papers retrieved. 13 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
CausalMixFT achieves the highest median improvement of (+0.12$\pm$0.63) over the pretrained model on 33 classification data	×	0.09
Default fine-tuning has a variability of $\pm$0.98, while CausalMixFT has a variability of $\pm$0.63.	×	0.09
CausalMixFT ranks first overall in average ranks across datasets, followed by the default fine-tuning baseline.	×	0.08
Purely synthetic generators, including CTGAN, SCM, TabEBM, TableAugment, and MixedModel, show negative median improvemen	×	0.07
The normalization strategy used to compare performance across different data generators is based on Gorishniy et al. [12	×	0.05
The base model's (Mitra's) zero-shot performance is used as the performance baseline.	×	0.05
The normalized performance is computed as scorenormalized = metricsign $\times$ (scoremethod / scorebaseline -1) $\times$ 100%.	×	0.02
CausalMixFT extends the fine-tuning framework of Bhler et al. [5] by mixing real and causally grounded synthetic sample	×	0.11
SCMs explicitly encode causal dependencies among features through a directed acyclic graph (DAG) and a set of structural	×	0.05
The PC and FCI algorithms are used to estimate the structural relations between the features.	×	0.02
DoWhy's SCM framework with additive noise models is used to sample and fit DAGs.	×	0.03
Numerical features are modeled with regressors, and categorical features with classifiers in the SCM framework.	×	0.04
Synthetic samples are generated by sampling exogenous noise and propagating it through the fitted SCM.	×	0.05

References

- http://arxiv.org/abs/2601.04110v2

- http://arxiv.org/abs/1002.1148v1
- http://arxiv.org/abs/2512.03307v1