Synthetic Data Dimensionality and Structural Complexity in Tabular Foundation Model Fine-Tuning

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the dimensionality and structural complexity of synthetic data generated via SCMs influence the fine-tuning efficiency (measured in training steps and validation accuracy) of tabular. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. 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 dimensionality and structural complexity of synthetic data generated via SCMs influence the fine-tuning efficiency (measured in training steps and validation accuracy) of tabular foundation models on sparse 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 3.8/10.

3 Results

12 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 3.8/10.

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
Experiments were conducted on the Mitra model across 33 classification datasets from the TabArena benchmark suite.	×	0.08
The experimental setup involved 10 folds per dataset, totaling 2,310 fine-tuning runs.	×	0.10
Model performance was reported as normalized ROC-AUC relative to the pre-trained model.	×	0.08
CausalMixFT achieved a median improvement of +0.12$\pm$0.63 over the pre-trained model.	×	0.05
The default fine-tuning baseline achieved a median improvement of +0.10$\pm$0.98 over the pre-trained model.	×	0.09
Purely synthetic augmentation methods (CTGAN, SCM, TabEBM, TableAugment, and MixedModel) showed negative median improvem	×	0.08
CausalMixFT demonstrated lower performance variability ($\pm$0.63) compared to default fine-tuning ($\pm$0.98).	×	0.09
In average rank analysis, CausalMixFT ranked first overall, followed by the default fine-tuning baseline.	×	0.07
The normalization strategy uses the base model's (Mitra's) zero-shot performance as the baseline.	×	0.03
The normalization formula is defined as: score_normalized = metric_sign $\times$ (score_method / score_baseline - 1) $\times$ 100%.	×	0.00
The method generates synthetic data using Structural Causal Models (SCMs) fitted to the target dataset.	✓	0.23
Structural relations between features are estimated using the PC and FCI algorithms.	×	0.03
The estimation process produces a probabilistic adjacency matrix encoding edge strengths between variables.	×	0.03
DAGs are sampled and fitted using DoWhy's SCM framework with additive noise models.	×	0.03
Numerical features are modeled with regressors and categorical features with classifiers within the SCM.	×	0.04

References

- http://arxiv.org/abs/2601.04110v2
- http://arxiv.org/abs/2603.10254v1
- http://arxiv.org/abs/2512.03307v1