

CausalMixFT Synthetic Data Generation and Sample Complexity in Tabular Benchmark Performance

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the sample complexity of CausalMixFT-scale synthetic data generation compare to standard mixing strategies when evaluating downstream task performance (e.g., accuracy, F1-score) on tabular. 19 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. 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 sample complexity of CausalMixFT-scale synthetic data generation compare to standard mixing strategies when evaluating downstream task performance (e.g., accuracy, F1-score) on tabular benchmarks like TabularGLUE or TabMNR?.

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 4.4/10.

3 Results

12 papers retrieved. 19 claims extracted; 1 independently verified. Quality review score: 4.4/10.

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 with 10 folds each from the TabArena ben	×	0.07
The study totaled 2,310 fine-tuning runs.	×	0.11
Model performance is reported as normalized ROC-AUC relative to the pre-trained model.	×	0.07
CausalMixFT achieves a median improvement of $+0.12 \pm 0.63$ over the pre-trained model.	×	0.05
The default fine-tuning baseline achieves a median improvement of $+0.10 \pm 0.98$ over the pre-trained model.	×	0.08
Purely synthetic augmentation methods (CTGAN, SCM, TabEBM, TableAugment, and MixedModel) show negative median improvement	×	0.08
CausalMixFT has a performance variability of ± 0.63 , while default fine-tuning has a variability of ± 0.98 .	×	0.09
In average rank analysis, CausalMixFT ranks first overall, followed by the default fine-tuning baseline.	×	0.07
Purely synthetic generators occupy lower ranks than CausalMixFT and the default fine-tuning baseline in average rank ana	×	0.09
The normalization strategy uses the base model's (Mitra's) zero-shot performance as the baseline.	×	0.02
The normalization formula is defined as: $\text{score_normalized} = \text{metric_sign} \times (\text{score_method} / \text{score_baseline} - 1) \times 100\%$.	×	0.00
In the normalization formula, metric_sign is 1 for metrics where higher is better (e.g., ROC-AUC) and -1 for metrics whe	×	0.04
The method generates synthetic data using Structural Causal Models (SCMs) fitted to the target dataset.	✓	0.21
SCMs encode causal dependencies among features through a directed acyclic graph (DAG) and a set of structural equations.	×	0.04
Structural relations between features are estimated using the PC and FCI algorithms.	×	0.03
The estimation of structural relations produces a probabilistic adjacency matrix encoding edge strengths between variabl	×	0.03
DAGs are sampled and fitted using DoWhy's SCM framework with additive noise models.	×	0.03
In the SCM framework, numerical features are modeled with regressors and categorical features with classifiers	×	0.05

References

- <http://arxiv.org/abs/2601.04110v2>
- <http://arxiv.org/abs/2407.21523v1>
- <http://arxiv.org/abs/2504.20900v1>