Robustness of Fine-Tuned Tabular Foundation Models with CausalMixFT-Scale Synthetic Data

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the robustness of fine-tuned tabular foundation models differ when trained with CausalMixFT-scale synthetic data versus standard mixing strategies under distribution shift conditions (e.g.,. 18 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. 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 robustness of fine-tuned tabular foundation models differ when trained with CausalMixFT-scale synthetic data versus standard mixing strategies under distribution shift conditions (e.g., measured by OOD detection accuracy on benchmarks like WILDS)?.

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 4.5/10.

3 Results

12 papers retrieved. 18 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
Experiments were conducted on the Mitra model across 33 classification datasets from the TabArena benchmark suite.	×	0.08
The experimental setup involved 10 folds for each of the 33 datasets, totaling 2,310 fine-tuning runs.	×	0.09
Model performance was reported as normalized ROC-AUC relative to the pre-trained model.	×	0.07
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$0.98) over the pre-trained model.	×	0.08
Purely synthetic augmentation methods (CTGAN, SCM, TabEBM, TableAugment, and MixedModel) showed negative median improvem	×	0.08
CausalMixFT demonstrated lower performance variability ($\pm$0.63) compared to the default fine-tuning baseline ($\pm$0.98).	×	0.07
In average rank analysis across datasets, CausalMixFT ranked first overall.	×	0.03
The default fine-tuning baseline ranked second in average rank analysis, followed by purely synthetic generators.	×	0.07
The normalization strategy used sets the base model's (Mitra's) zero-shot performance as the baseline.	×	0.05
The normalization formula is defined as: score_normalized = metric_sign $\times$ ((score_method / 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 proposed 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 structural equations.	×	0.05
Structural relations between features were estimated using the PC and FCI algorithms.	×	0.03
DAGs were 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.	×	0.03
Synthetic samples are generated by sampling ex	×	0.04

References

- http://arxiv.org/abs/2601.04110v2
- http://arxiv.org/abs/2410.02152v1
- http://arxiv.org/abs/2512.03307v1