Comparative Accuracy of Learned Structural Causal Models in CausalMixFT Versus State-of-the-Art Methods on Tabular Benchmarks

Assignee Research

June 11, 2026

Abstract

Fine-tuning tabular foundation models (TFMs) under data scarcity is challenging, as early stopping on even scarcer validation data often fails to capture true generalization performance. We propose CausalMixFT, a method that enhances fine-tuning robustness and downstream performance by generating structurally consistent synthetic samples using Structural Causal Models (SCMs) fitted on the target dataset. This approach augments limited real data with causally informed synthetic examples, preserving feature dependencies while expanding training diversity. Evaluated across 33 classification datas

1 Introduction

This paper examines: Causal Data Augmentation for Robust Fine-Tuning of Tabular Foundation Models. Research question: How does the accuracy of learned Structural Causal Models (SCMs) in CausalMixFT compare to state-of-the-art causal discovery methods on tabular benchmarks like Camelyon17 or CatBoost datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

16 papers retrieved. 19 claims extracted; 14 independently verified. Quality review score: 7.2/10.

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	✓	0.21
The study totaled 2,310 fine-tuning runs.	×	0.10
Model performance is reported as normalized ROC-AUC relative to the pre-trained model.	✓	0.27
CausalMixFT achieves a median improvement of +0.12$\pm$0.63 over the pre-trained model.	✓	0.20
The default fine-tuning baseline achieves a median improvement of +0.10$\pm$0.98 over the pre-trained model.	✓	0.23
Purely synthetic augmentation methods (CTGAN, SCM, TabEBM, TableAugment, and MixedModel) show negative median improvemen	✓	0.22
CausalMixFT has a variability of $\pm$0.63, while default fine-tuning has a variability of $\pm$0.98.	×	0.15
In average rank analysis, CausalMixFT ranks first overall, followed by the default fine-tuning baseline.	✓	0.21
Purely synthetic generators occupy lower ranks than CausalMixFT and the default fine-tuning baseline in average rank ana	✓	0.22
The normalization strategy uses the base model's (Mitra's) zero-shot performance as the baseline.	×	0.06
The normalization formula is 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	×	0.09
The method generates synthetic data using SCMs fitted to the target dataset.	✓	0.15
SCMs encode causal dependencies among features through a directed acyclic graph (DAG) and a set of structural equations.	✓	0.24
Structural relations between features are estimated using the PC and FCI algorithms.	✓	0.16
The estimation process produces a probabilistic adjacency matrix that encodes edge strengths between variables.	✓	0.16
DAGs are sampled and fitted using DoWhy's SCM framework with additive noise models.	✓	0.24
Numerical features are modeled with regressors and categorical features with classifiers within the SCM.	✓	0.18
Synthetic samples are generated by sampling exogenous noise and propagating it through the	✓	0.23

References

- http://arxiv.org/abs/2601.04110v2
- http://arxiv.org/abs/2312.07577v3
- http://arxiv.org/abs/2506.16791v4