Causal Structure Preservation in CausalMixFT Enhances Robustness of Transformers Under Adversarial Attacks

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Can the causal structure preservation in CausalMixFT's synthetic samples enhance the robustness of TFMs under adversarial attacks compared to standard fine-tuning, as measured by accuracy on. 19 claims were extracted from source literature; 2 were 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: Can the causal structure preservation in CausalMixFT's synthetic samples enhance the robustness of TFMs under adversarial attacks compared to standard fine-tuning, as measured by accuracy on perturbed tabular datasets such as those in TabBench?.

2 Methodology

Systematic literature search across multiple databases yielded 14 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

14 papers retrieved. 19 claims extracted; 2 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 with 10 folds each from the TabArena ben	×	0.07
The study totaled 2,310 fine-tuning runs.	×	0.12
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.	×	0.09
Purely synthetic augmentation methods (CTGAN, SCM, TabEBM, TableAugment, and MixedModel) show negative median improvemen	×	0.08
CausalMixFT has a performance variability of $\pm$0.63, while default fine-tuning has a variability of $\pm$0.98.	×	0.10
In average rank analysis across datasets, CausalMixFT ranks first overall.	×	0.03
In average rank analysis, the default fine-tuning baseline ranks second, followed by purely synthetic generators.	×	0.09
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 = metricsign $\times$ (score_method / score_baseline - 1) $\times$ 100%.	×	0.00
In the normalization formula, metricsign is 1 for metrics where higher is better (e.g., ROC-AUC) and -1 for metrics wher	×	0.02
The method generates synthetic data using Structural Causal Models (SCMs) fitted to the target dataset.	✓	0.22
SCMs encode causal dependencies among features through a directed acyclic graph (DAG) and structural equations.	✓	0.16
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.	×	0.01
Synthetic samples are generated by sampling ex-	×	0.04

References

- http://arxiv.org/abs/2601.04110v2
- http://arxiv.org/abs/2512.03307v1
- http://arxiv.org/abs/2403.10075v2