Adversarial Training Enhances Robustness in Tabular Foundation Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of incorporating adversarial training during fine-tuning on the robustness of tabular foundation models, as measured by accuracy on the TabMNAR and TabFair benchmarks compared to. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. 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: What is the impact of incorporating adversarial training during fine-tuning on the robustness of tabular foundation models, as measured by accuracy on the TabMNAR and TabFair benchmarks compared to standard fine-tuning approaches?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

13 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 4.2/10.

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
CausalMixFT achieves the highest median improvement of (+0.12$\pm$0.63) over the pretrained model.	×	0.05
Default fine-tuning has a variability of $\pm$0.98, while CausalMixFT has a variability of $\pm$0.63.	×	0.06
CausalMixFT ranks first overall in average ranks across datasets.	×	0.03
The experiments were conducted on the Mitra model across 33 classification datasets with 10 folds each from the TabArena	×	0.11
SCM-based augmentation stabilizes fine-tuning under small-data conditions by introducing causally structured synthetic d	×	0.11
The normalization strategy suggested by Gorishniy et al. [12] is used to compare the performance across different data g	×	0.05
The base model's (Mitra's) zero-shot performance is used as the performance baseline.	×	0.08
The normalized performance is computed as scorenormalized = metricsign $\times$ (scoremethod / scorebaseline -1) $\times$ 100%.	×	0.02
SCMs explicitly encode causal dependencies among features through a directed acyclic graph (DAG) and a set of structural	×	0.05
The structural relations between the features are estimated using the PC and FCI algorithms.	×	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.	×	0.04
Synthetic samples are generated by sampling exogenous noise and propagating it through the fitted SCM.	×	0.05

References

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

- http://arxiv.org/abs/2412.14097v1