

# Causally Augmented Synthetic Data for Robust Cross-Domain Tabular Foundation Models

Assignee Research

June 7, 2026

## Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: Can cross-domain transfer learning with causally augmented synthetic data enhance the robustness of tabular foundation models across domains (e.g., healthcare vs. finance) while maintaining high. 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: Can cross-domain transfer learning with causally augmented synthetic data enhance the robustness of tabular foundation models across domains (e.g., healthcare vs. finance) while maintaining high accuracy on downstream tasks like regression or classification?.

## 2 Methodology

Systematic literature search across multiple databases yielded 11 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

11 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 pre-trained model in fine-tuning performance	×	0.09
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, followed by the default fine-tuning baseline.	×	0.06
Purely synthetic generators, including CTGAN, SCM, TabEBM, TableAugment, and Mixed-Model, show negative median improvement	×	0.12
The experiments were conducted on the Mitra model across 33 classification datasets with 10 folds each from the TabArena	×	0.11
Model performance is reported as normalized ROC-AUC relative to the pre-trained model.	×	0.08
CausalMixFT extends the fine-tuning framework by mixing real and causally grounded synthetic samples into the fine-tuning	×	0.12
SCM-Based Synthetic Augmentation (CausalMixFT) uses SCMs fitted to the target dataset to generate synthetic data.	×	0.13
SCMs explicitly encode causal dependencies among features through a directed acyclic graph (DAG) and a set of structural	×	0.14
The structural relations between the features are estimated using the PC and FCI algorithms.	×	0.04
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 in the SCM framework.	×	0.03
Synthetic samples are generated by sampling exogenous noise and propagating it through the fitted SCM.	×	0.04

## References

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

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