

What is the trade-off between pre-training computational cost (measured in FLOPs) and downstream fairness (mea

Assignee Research

June 10, 2026

Abstract

Foundation models for tabular data, such as the Tabular Prior-data Fitted Network (TabPFN), are pre-trained on a massive number of synthetic datasets generated by structural causal models (SCM). They leverage in-context learning to offer high predictive accuracy in real-world tasks. However, the fairness properties of these foundational models, which incorporate ideas from causal reasoning during pre-training, remain underexplored. In this work, we conduct a comprehensive empirical evaluation of TabPFN and its fine-tuned variants, assessing predictive performance, fairness, and robustness acro

1 Introduction

This paper examines: Causal Pre-training Under the Fairness Lens: An Empirical Study of TabPFN. Research question: What is the trade-off between pre-training computational cost (measured in FLOPs) and downstream fairness (measured via demographic parity) when scaling the number of SCM features in TabPFN?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

15 papers retrieved. 17 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
The Heart dataset contains 303 samples and 13 features, with sensitive attributes being age and sex.	×	0.02
The Bank dataset contains 5,000 samples and 14 features, with sensitive attributes being education and family.	×	0.03
The Law dataset contains 21,000 samples and 8 features, with sensitive attributes being race and sex.	×	0.03
The Adult dataset contains 48,000 samples and 14 features, with sensitive attributes being race and sex.	×	0.02
Logistic Regression (LR) is evaluated with L2 regularization ($C=0.1$).	×	0.00
Random Forest (RF) is evaluated with 50 trees of maximum depth 5 and minimum samples per split of 10.	×	0.01
Multi-Layer Perceptron (MLP) is evaluated with one hidden layer of 50 units, L2 regularization ($\alpha=0.01$), and up to 300 t	×	0.02
TabPFN is evaluated in zero-shot mode with pretraining limits disabled, <code>n_estimators=2</code> , and inference using 5000 subsamp	×	0.05
FT-TabPFN is fine-tuned for 10 epochs via Adam ($lr=1 \times 10^{-5}$), meta-batch size 1, inner batch size 5000, and cross-entropy	×	0.05
TabPFN is designed for small- to medium-sized tabular datasets (up to roughly 10k samples).	×	0.04
For the Adult and Law datasets, TabPFN is evaluated on subsets of 500, 1k, and 10k samples to analyze how fairness varie	×	0.04
Fairness metrics used are Equal Opportunity (EO) and Demographic Parity (DP), implemented via FairLearn.	×	0.05
Spurious correlation perturbation experiment introduces an artificial feature Z_{spur} that is strongly correlated with the	×	0.04
Z_{spur} is a synthetic continuous variable $Z_{spur} \sim N(\mu_y, 0.5)$, where $\mu_y=1$ if $Y=1$ and 0 otherwise in training data, and fl	×	0.05
Flip Consistency Rate measures the proportion of unchanged predictions when spurious attributes are flipped.	×	0.03
FT-TabPFN and TabPFN exhibit superior flip consistency, reaching 0.82–0.97 across datasets and leading the pack on Heart	×	0.03
FT-TabPFN and TabPFN have the highest accuracy on most datasets (Bank 0.97, Heart/Adult 0.75-0.88).	×	0.05

References

- <http://arxiv.org/abs/2603.10254v1>
- <http://arxiv.org/abs/1909.03564v2>
- <http://arxiv.org/abs/2601.17912v2>