Hybrid Causal-Non-Causal Pre-Training for Fair TabPFN Models

Assignee Research

June 12, 2026

Abstract

With the growing adoption of machine learning (ML) systems in areas like law enforcement, criminal justice, finance, hiring, and admissions, it is increasingly critical to guarantee the fairness of decisions assisted by ML. In this paper, we study the tradeoff between fairness and accuracy under the statistical notion of equalized odds. We present a new upper bound on the accuracy (that holds for any classifier), as a function of the fairness budget. In addition, our bounds also exhibit dependence on the underlying statistics of the data, labels and the sensitive group attributes. We validate

1 Introduction

This paper examines: Intrinsic Fairness-Accuracy Tradeoffs under Equalized Odds. Research question: Can a hybrid causal-non-causal pre-training framework for TabPFN achieve better fairness-accuracy trade-offs than purely causal models, as evaluated on biased tabular datasets using accuracy scores and fairness metrics?.

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 7.4/10.

3 Results

16 papers retrieved. 13 claims extracted; 10 independently verified. Quality review score: 7.4/10.

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 paper focuses on the group fairness notion of equalized odds (EO).	✓	0.16
A binary classifier f satisfies ϵEO-Equalized Odds if the maximum difference in prediction probabilities between sensiti	×	0.12
When ϵEO = 0, the prediction f(X) is conditionally independent of the sensitive attribute Z given the label Y.	✓	0.16
The Total Variation (TV) distance between two probability distributions P and Q is defined as half the integral of the a	×	0.13
An unconstrained upper bound on accuracy for any binary classifier is given by the formula: Acc = max(1 - α, α) + min(1	✓	0.25
The unconstrained accuracy upper bound is attainable by the Bayes optimal classifier when the class distribution α is 0,	✓	0.23
For a fixed class distribution α, accuracy is directly proportional to the total variation distance dTV(P1, P0).	✓	0.16
The paper presents a new classifier-independent upper bound on accuracy as a function of the EO budget (ϵEO).	✓	0.24
The derived bounds are determined by the underlying statistics of the data, labels, and sensitive groups.	✓	0.17
The primary technique used to derive the bounds involves adapting Le Cam's bound to encompass Equalized Odds fairness co	✓	0.24
The original Le Cam's bound is based on the total variation distance between two class distributions (dTV(P0, P1)).	✓	0.29
Prior works [17], [25] study fair Bayes-optimal classifiers subject to equal opportunity, which is a weaker notion than	✓	0.27
Prior bounds established in [26] are intrinsically linked to the characteristics of the classifier itself.	×	0.11

References

- http://arxiv.org/abs/2601.17912v2
- http://arxiv.org/abs/2405.07393v1
- http://arxiv.org/abs/2501.14710v1