Demographic Parity in TabPFN Predictions Across Causal Structures and Quantification via Counterfactual Fairness Metrics

Assignee Research

June 12, 2026

Abstract

Machine learning systems are becoming increasingly ubiquitous. These systems's adoption has been expanding, accelerating the shift towards a more algorithmic society, meaning that algorithmically informed decisions have greater potential for significant social impact. However, most of these accurate decision support systems remain complex black boxes, meaning their internal logic and inner workings are hidden to the user and even experts cannot fully understand the rationale behind their predictions. Moreover, new regulations and highly regulated domains have made the audit and verifiability o

1 Introduction

This paper examines: Machine Learning Interpretability: A Survey on Methods and Metrics. Research question: How does the demographic parity of TabPFN's downstream predictions vary when trained on SCMs with different causal structures, and can a causal fairness metric like counterfactual fairness be used to quantify this effect?.

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 8.3/10.

3 Results

14 papers retrieved. 14 claims extracted; 14 independently verified. Quality review score: 8.3/10.

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
Machine learning systems are becoming increasingly ubiquitous.	✓	0.19
The adoption of machine learning systems has been expanding, accelerating the shift towards a more algorithmic society.	✓	0.22
Algorithmic decisions have greater potential for significant social impact.	✓	0.21
Most accurate decision support systems remain complex black boxes.	✓	0.25
The internal logic and inner workings of machine learning systems are hidden to the user.	✓	0.23
Experts cannot fully understand the rationale behind the predictions of machine learning systems.	✓	0.24
New regulations and highly regulated domains have made the audit and verifiability of decisions mandatory.	✓	0.27
There is an increasing demand for the ability to question, understand, and trust machine learning systems.	✓	0.27
Interpretability is indispensable for the audit and verifiability of machine learning systems.	✓	0.18
The research community has recognized the interpretability problem and focused on developing both interpretable models a	✓	0.34
There is no consensus on how to assess the explanation quality of machine learning systems.	✓	0.24
The aim of the article is to provide a review of the current state of the research field on machine learning interpretab	✓	0.31
The article focuses on the societal impact and on the developed methods and metrics of machine learning interpretability	✓	0.26
A complete literature review is presented in the article to identify future directions of work on machine learning inter	✓	0.27

References

- https://doi.org/10.1016/j.future.2024.02.023

- https://doi.org/10.1007/s10618-023-00933-9
- https://doi.org/10.3390/electronics8080832