Revisiting Preprocessing for Fair ML: Statistically Robust Evaluation and a Novel EMD-Driven Optimisation Method

THE UNIVERSITY OF SYDNEY

Hyeonggeun Yun (Supervisor: Dr Shahadat Uddin)

Summer Vacation Research Internship

School of Electrical and Computer Engineering, The University of Sydney

1. Introduction

➤ **Motivation:** ML models used in high-stakes domains can exhibit systematic bias against protected groups like gender or race.

➤ **Gap:** Prior studies on preprocessing-based bias mitigation often lack statistical validation and broad comparative evaluation.

➤ **Approach:** We compare instance removal, resampling, and feature transformation methods across multiple datasets, models, and fairness metrics using cross-validation and hypothesis testing.

➤ **Key contribution:** An automated, statistically grounded optimisation of Earth Mover's Distance (EMD)-based instance removal.

➤ **Findings:** Distribution-aware preprocessing achieves substantial fairness improvements, while preserving predictive performance.
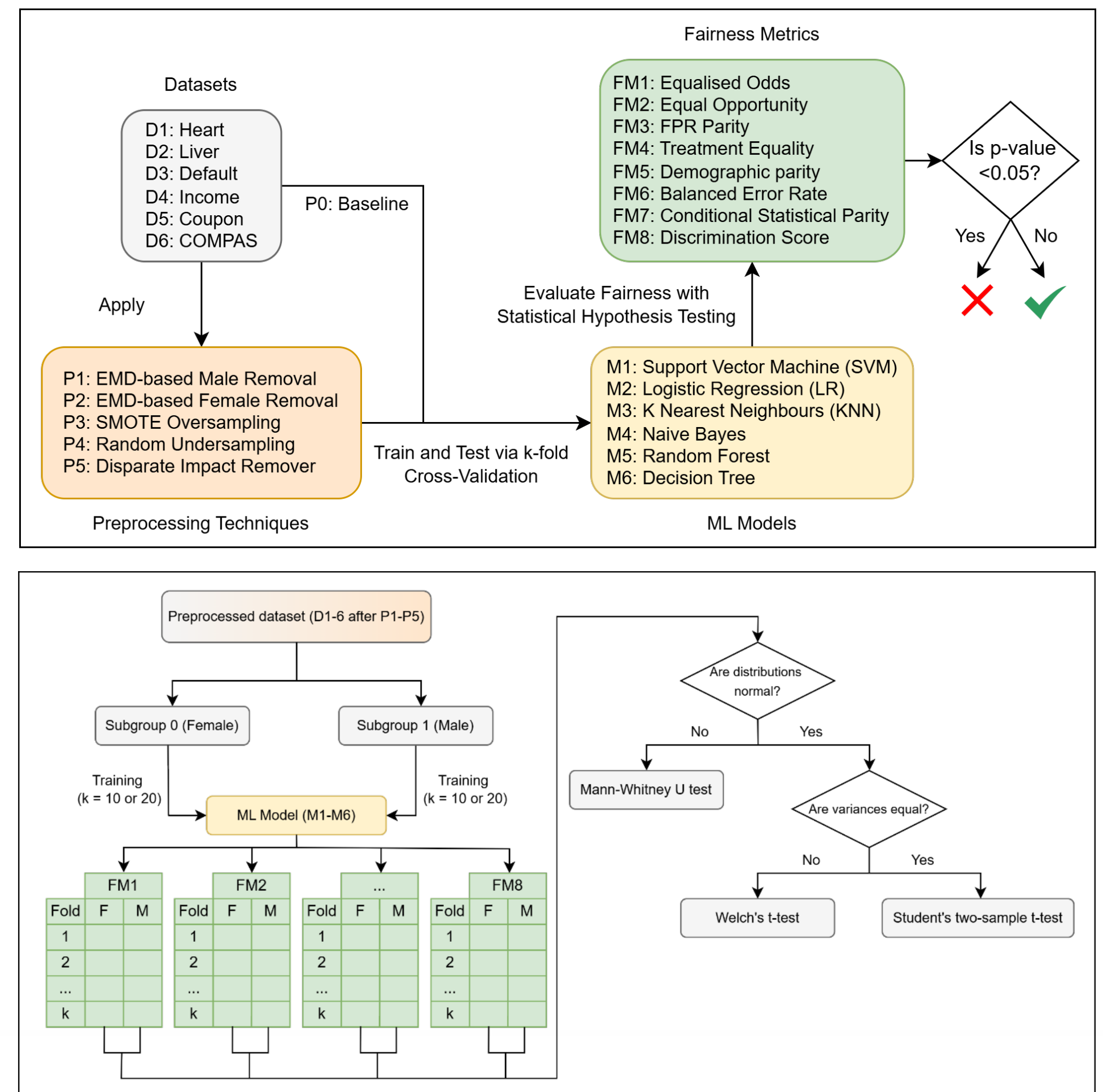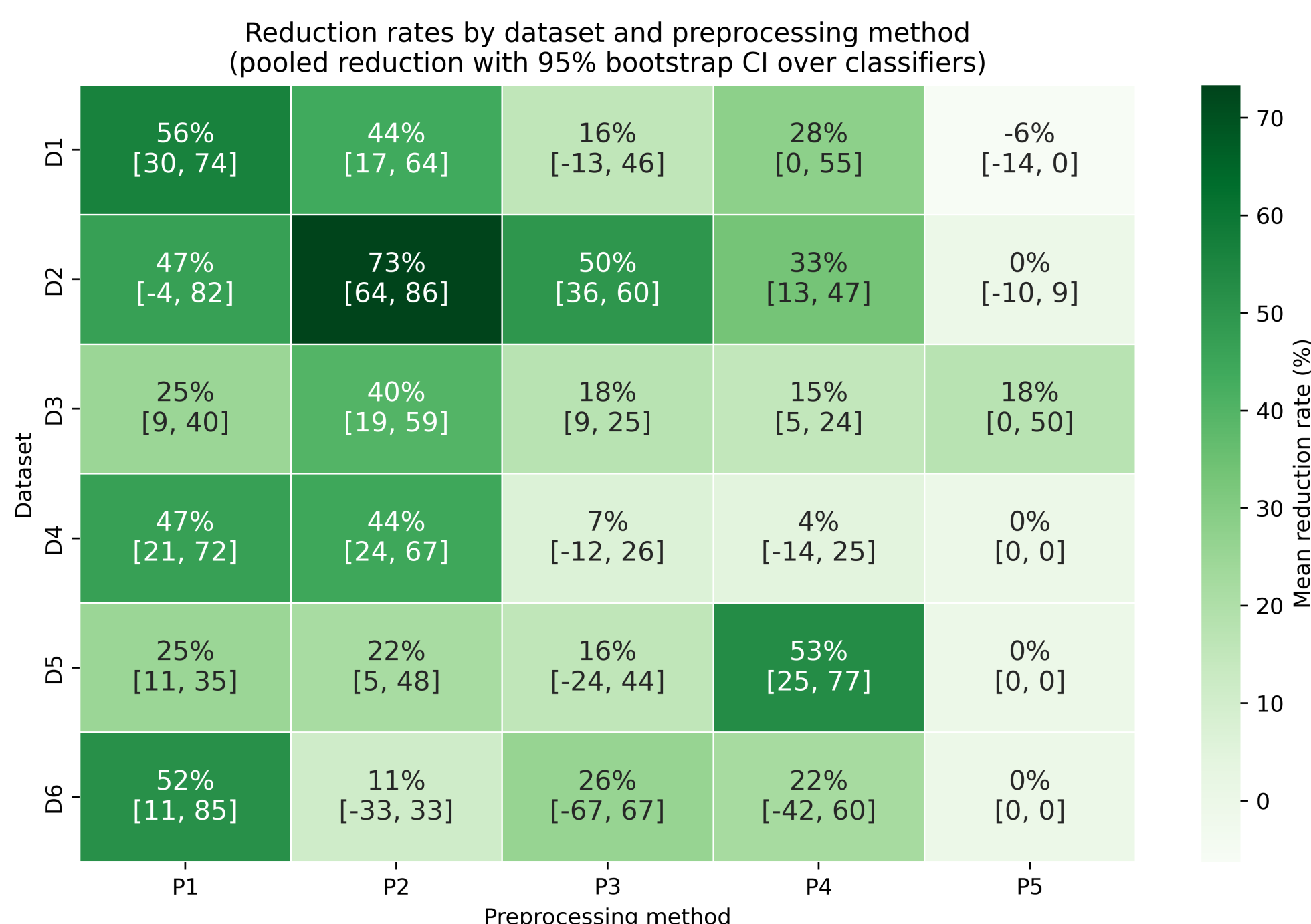
2. Methodology

For code and relevant artefacts! → SCAN ME
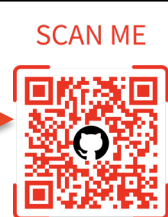
Fairness is evaluated within a cross-validated statistical framework that treats group disparities as inferential quantities rather than single-run estimates. For each dataset–model configuration, fairness violations are identified using parametric or non-parametric tests across multiple fairness definitions, enabling robust comparison of both the magnitude and stability of bias mitigation effects. Furthermore, the proposed EMD-based instance removal applies a coarse-to-fine search to identify the minimal number and type of instances required to be removed, achieving statistically insignificant distributional disparities.


3. Fairness Results

EMD-based instance removal (P1, P2) reduced statistically significant fairness violations by approximately 40% on average, compared to around 25% for resampling methods (P3, P4) and 2% for feature transformation approach (P5).

Reduction rates by dataset and preprocessing method
(pooled reduction with 95% bootstrap CI over classifiers)

4. Performance Results

Despite the fairness gains, EMD-based removal largely preserved predictive performance, whereas resampling methods introduced more noticeable trade-offs, including reduced accuracy but increased recall in several datasets.

Performance comparison between datasets and methods
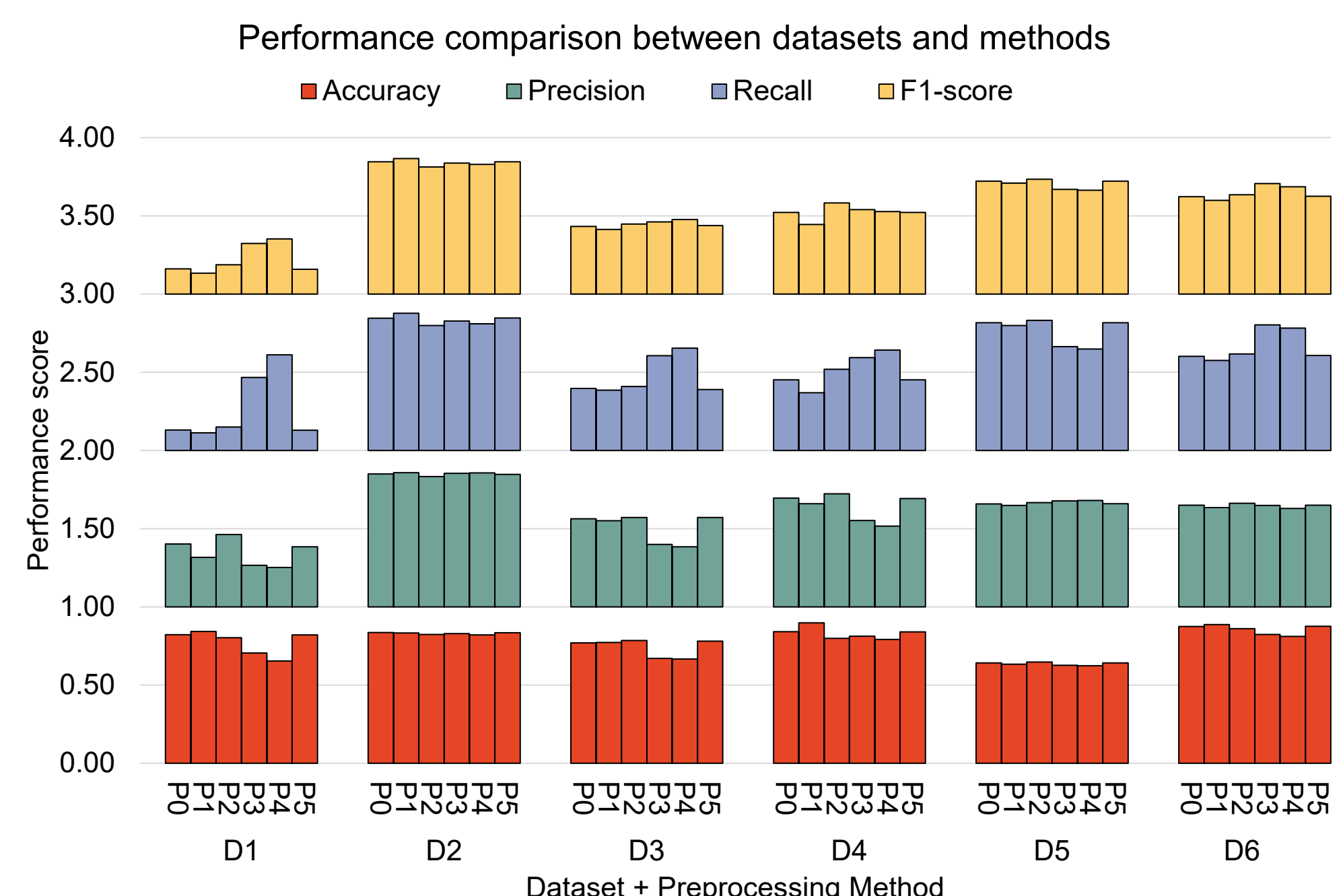
5. Conclusion and Future Work

➤ Statistically grounded preprocessing, particularly optimised EMD-based instance removal, provides consistent and substantial reductions in fairness violations without sacrificing predictive performance.

➤ Future work will explore hybrid EMD strategies that jointly consider male (P1) and female (P2) instance removal to balance interventions across groups, as well as extensions to intersectional attributes and multi-class datasets.

References

SCAN ME