

Promoting Intersectional Fairness through Knowledge Distillation

Md Fahim Sikder^a, Resmi Ramachandranpillai^b, Daniel de Leng^a, and Fredrik Heintz^a

^aDepartment of Computer and Information Science (IDA), Linköping University, Sweden

^bInstitute of Experiential AI, Northeastern University, USA



The Problem: Intersectional Bias in AI Systems

Discrimination can manifest at intersections of multiple sensitive attributes. Existing methods address single-attributes, thereby missing compounded discrimination.

Our Approach

Knowledge distillation + Modular Intersectional Loss (*False positive rate (FPR)* + *Demographic parity (DP)* + *Conditional Independence (CI)*) targeting Intersectional Groups

Key Contribution

- Novel two-stage framework
- Teacher focuses on accuracy; student inherits knowledge while enforcing fairness
- 52% accuracy improvement, 61% FPR reduction compared to the Baseline model

