## Promoting Intersectional Fairness through Knowledge Distillation

Md Fahim Sikder<sup>a</sup>, Resmi Ramachandranpillai<sup>b</sup>, Daniel de Leng<sup>a</sup>, and Fredrik Heintz<sup>a</sup>

<sup>a</sup>Department of Computer and Information Science (IDA), Linköping University, Sweden

<sup>b</sup>Institute of Experential AI, Northeastern University, USA



# The Problem: Intersectional Bias in Al Systems

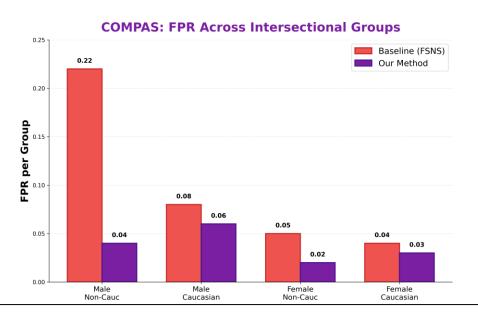
Discrimination can manifest at intersections of multiple sensitive attributes. Existing method address single-attributes, thereby missing compounded discrimination.

### 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



#### **Overall Predictive Performance**

