Promoting Intersectional Fairness through Knowledge Distillation

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

[a]Reasoning and Learning lab (ReaL), Department of Computer and Information Science (IDA), Linköping University, Sweden

[b]Institute of Experential AI, Northeastern University, USA

The problem: Intersectional Bias in AI Systems

Discrimination can manifest at intersections of multiple sensitive attributes (gender x race x ethnicity). For example: evidence showed commercial facial recognition software gives errors of 0.8% for lighter-skinned males versus 34.7% for darker-skinned females, while giving good performance when considering single-attributes alone [1]. Existing fairness methods address single-attributes, thereby missing compounded discrimination.

Our Approach: Novel Two-stage Framework for Mitigating Intersectional Bias

Knowledge distillation + Modular Intersectional Loss (FPR + DP + CI + Adversarial) targeting Intersectional Groups

Key Contribution

Novel Two-Stage Framework

- Presented a knowledge distillation framework that transfers predictive performance from a teacher model to a fair student model
- Ensures high utility while enabling fairness interventions
- Teacher focuses on accuracy; student inherits knowledge while enforcing fairness

Comprehensive Architecture

- Modular loss functions targeting False-positive rate parity (FPR), Demographic parity (DP), Conditional independence (CI)
- Support both binary and multi-class classification settings

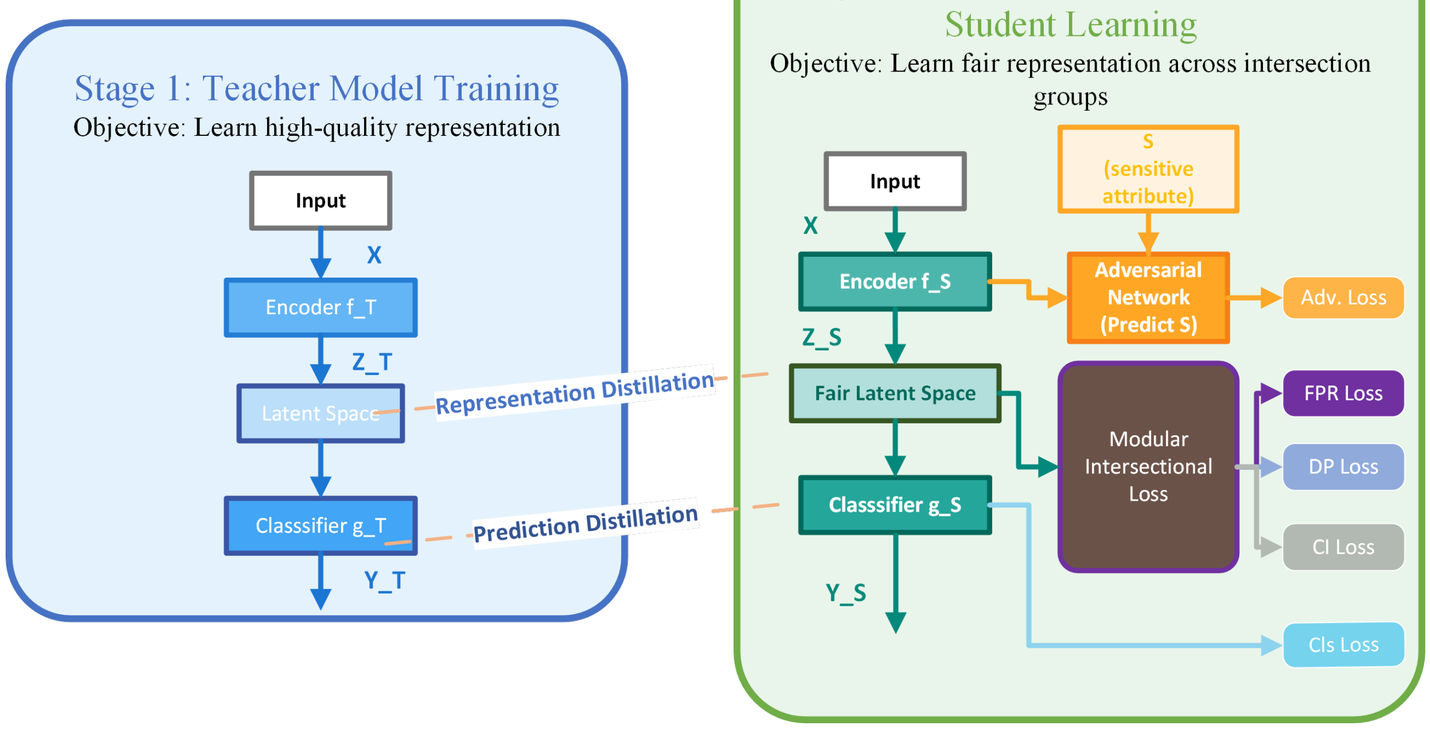
Improved Performance

- +52% accuracy, 61% FPR reduction, robust model

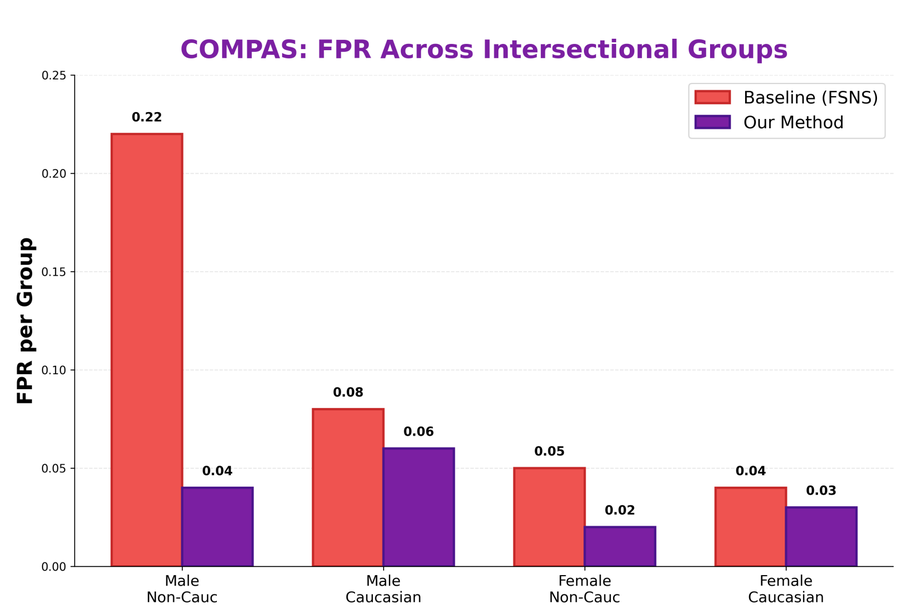
Practical Deployment

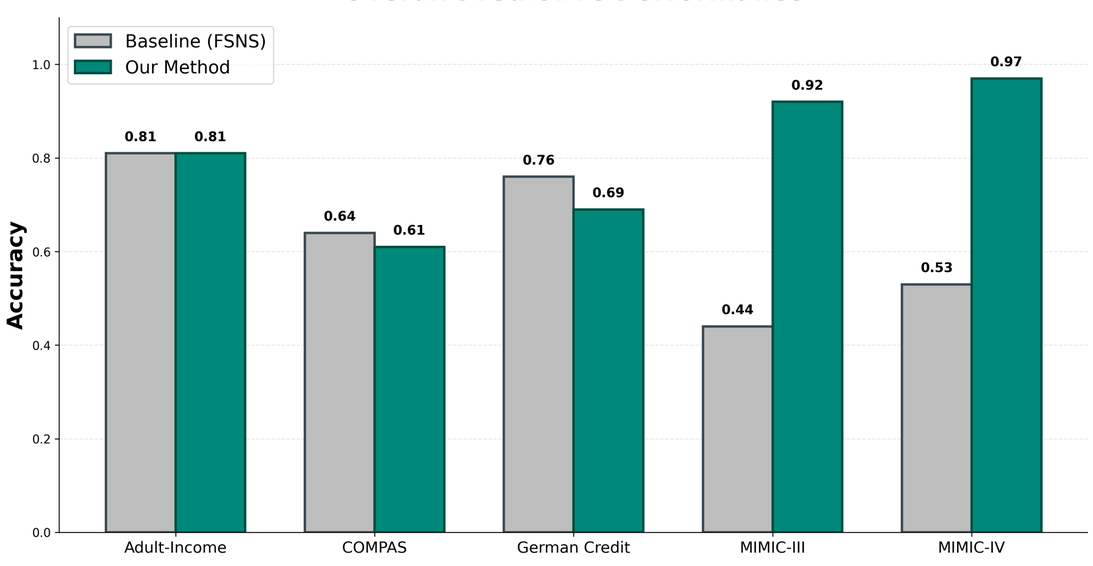
- 2.5 – 5 x parameter compression

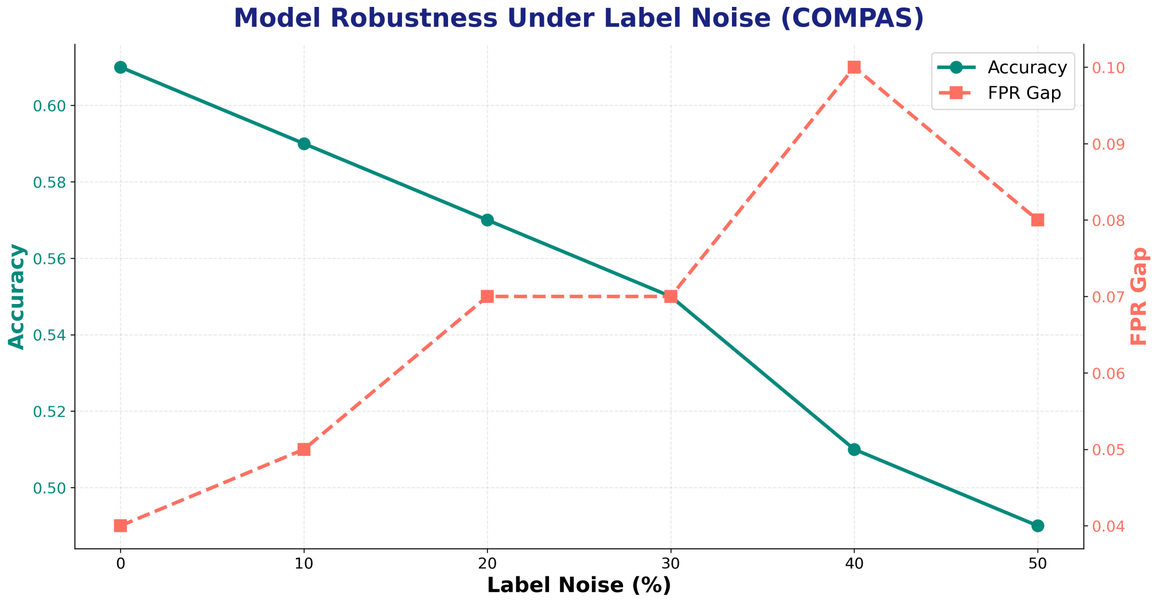
Framework Architecture



Experimental Results







Acknowledgement

This work was funded by the Knut and Alice Wallenberg Foundation, Sweden, the ELLIIT Excellence Center at Linköping-Lund for Information Technology, Sweden (portions of this work were carried out using the AIOps/Stellar). The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

References

[1] Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on fairness, accountability and transparency*. PMLR, 2018.

[2] Jang, Taeuk, et al. "Achieving fairness through separability: A unified framework for fair representation learning." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024.

Paper

LINKÖPING UNIVERSITY

Department of Computer and Information Science (IDA)
Artificial Intelligence and Integrated Computer Systems (AIICS)

ecai 2025 BOLOGNA