Applications of Fairness: Label Bias and Recovery of Ground Truth

Lina Battikha
lbattikha@ucsd.edu

Sai Poornasree Balamurugan
s2balamurugan@ucsd.edu

Mentor: Babak Salimi
bsalimi@ucsd.edu

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Introduction

Bias in artificial intelligence is an important problem to be addressed as such models disproportionately affect those who are of minority groups. One common reason that bias arises is due to sensitive features that are included in the dataset. Inclusion of sensitive features can cause the observed outcome label to pick up on unfair dependencies that will lead to label bias. ***Our project focuses on discovering fair labels that are not influenced by bias.***

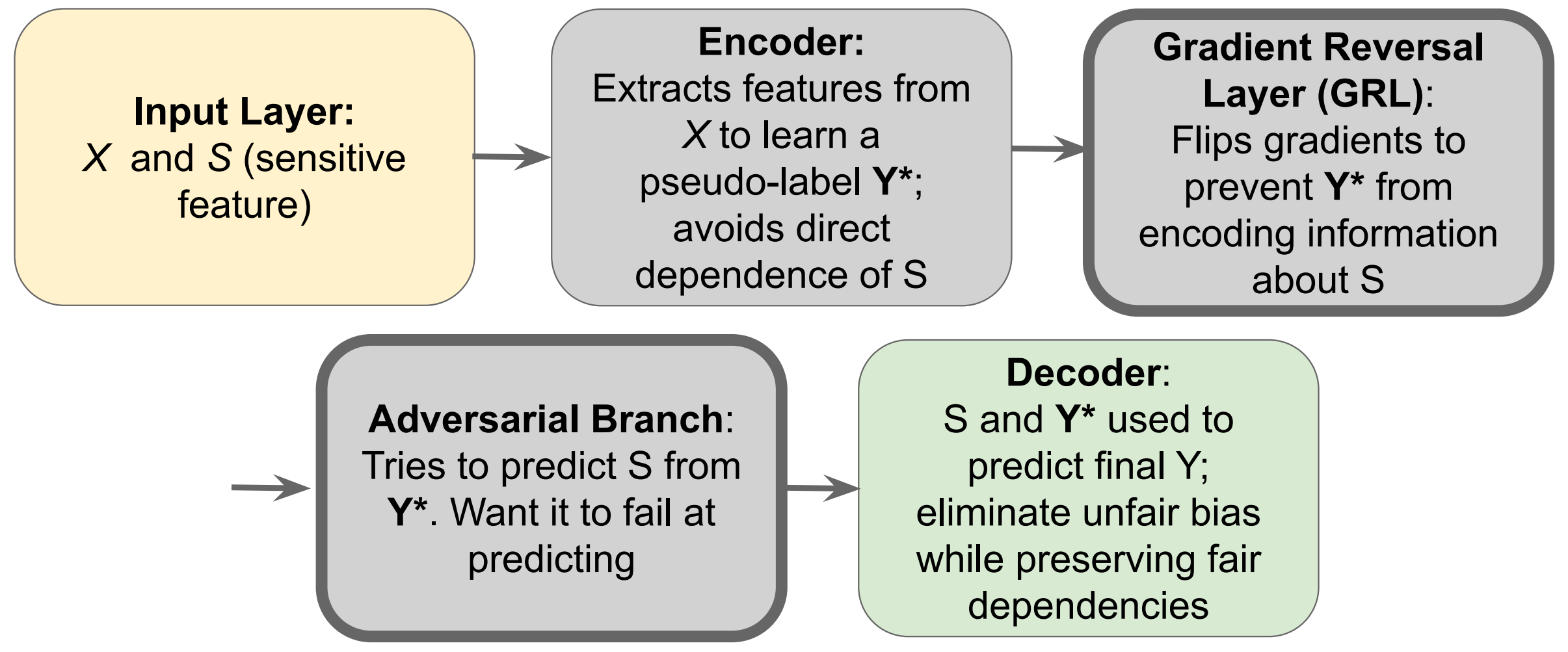
- Objective:
  - Uncover ground truth fair latent labels provided the biased labels
  - **Enhance fairness, ensuring equitable outcomes for all subpopulations**

Datasets

- **Synthetic Datasets:** A fair dataset with injected error to evaluate effectiveness of model
  - The sensitive feature and observed biased label have 2 classes
- **COMPAS Dataset:** Ethical concerns regarding racial bias in judicial decision-making
  - Sensitive Feature = African American/Not African American
  - Observed Label = Recidivism prediction within 2 years
- **UCI Adult Dataset**: Predict income levels and analyze fairness in wage distribution based on demographic and employment
  - Sensitive Feature = Male/Female
  - Observed Label = Income > $50K
- **UCI Drug Consumption Dataset:** Predict how likely an individual is to use cannabis
  - Sensitive Feature = Received Higher Education/Didn't Receive Higher Education
  - Observed Label = Likeliness of Cannabis Use (4 classes)

Methodology

- Use an encoder-decoder framework inspired by prior research for latent variable model [1].
- Establish independence between the sensitive feature and the observed label [2].
- **Fair Label Inference:** Treats the observed label Y as a proxy to infer the fair latent variable, ensuring it is independent of the sensitive feature S.



Input Layer: X and S (sensitive feature)

Encoder: Extracts features from X to learn a pseudo-label **Y***; avoids direct dependence of S

Gradient Reversal Layer (GRL): Flips gradients to prevent **Y*** from encoding information about S

Adversarial Branch: Tries to predict S from **Y***. Want it to fail at predicting

Decoder: S and **Y*** used to predict final Y; eliminate unfair bias while preserving fair dependencies

- **Outcome**: Effectively removes label bias, promotes fairness in representations, and maintains predictive accuracy.
- ***Focus on Y* for evaluation to see how well model establishes independence.***

Results

[Binary] Synthetic Dataset

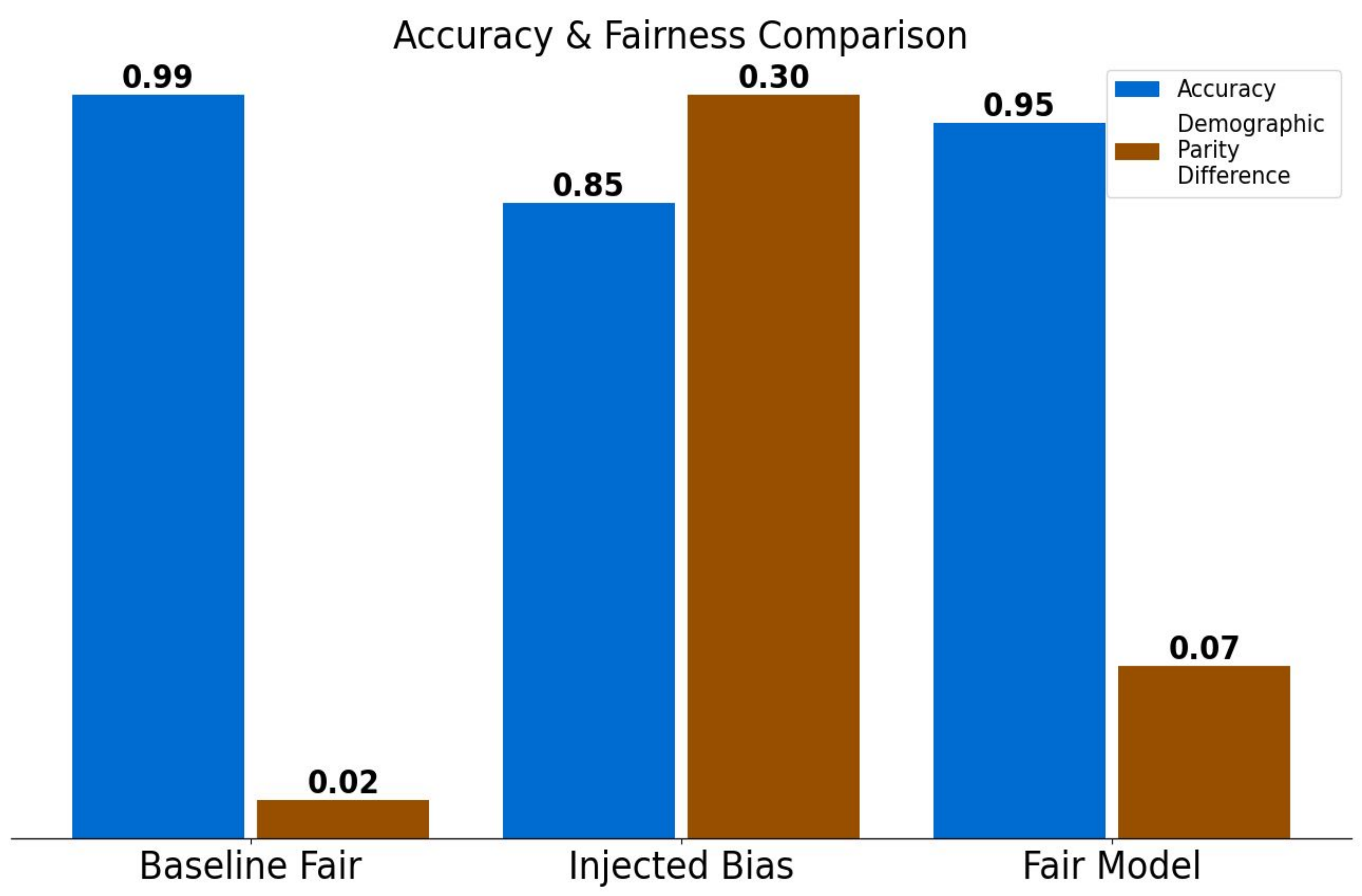


Figure 1: Proof of Model With Synthetic Data
Fair model was able to retrieve high accuracy while significantly reducing demographic parity difference, compared to the injected bias scenario. The baseline scenarios serves as reference to show that the model effectively removes label bias and recovers fair labels. Observed label is binary.

[Binary] UCI Adults Dataset

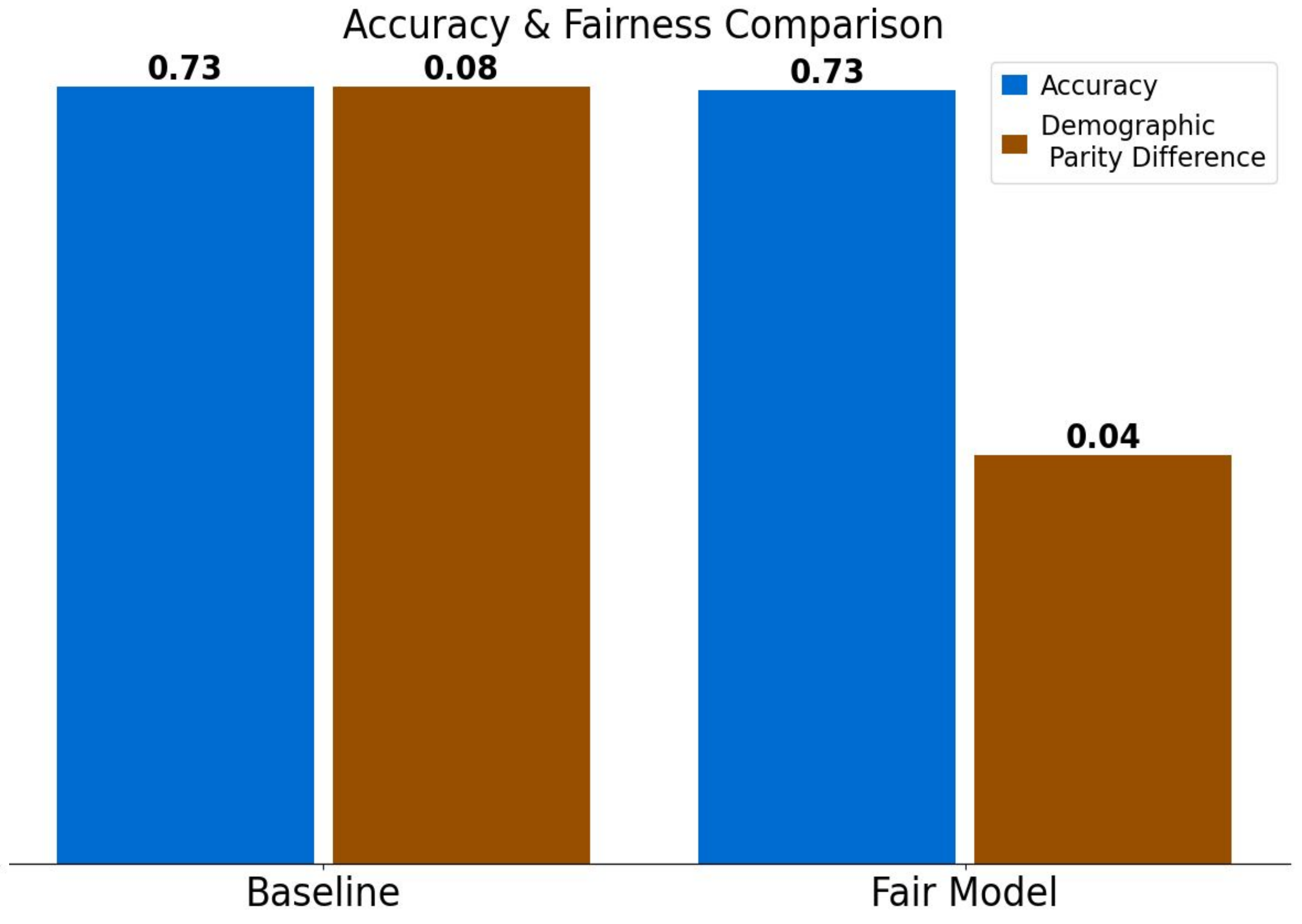


Figure 2: Maintaining High Accuracy
On real-world data, the fair latent variable model reduces demographic parity difference while maintaining nearly the same accuracy. This demonstrates model's ability to mitigate bias while preserving performance. Observed label is binary.

[Binary] COMPAS Dataset

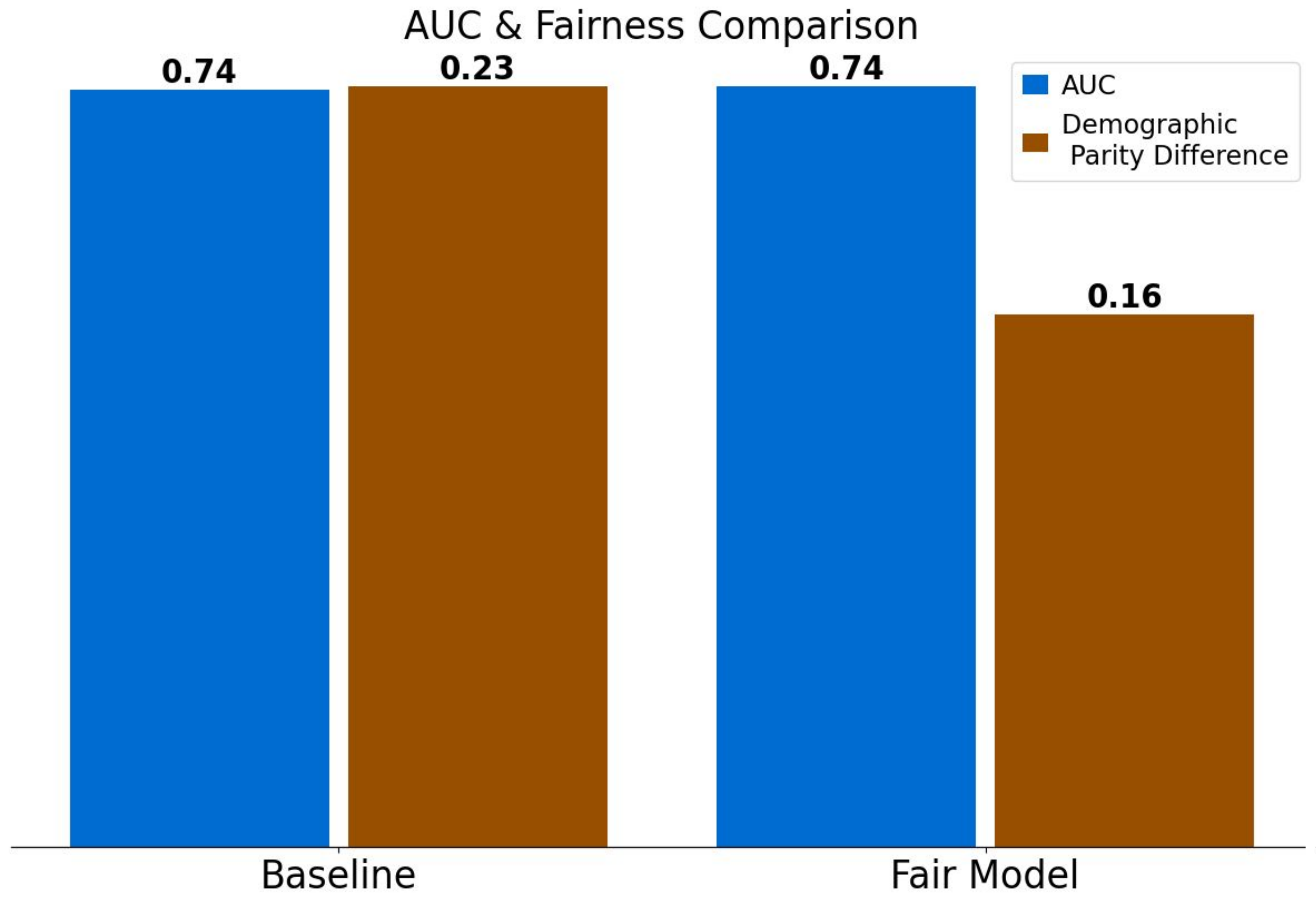


Figure 3: Maintaining High Area Under the Curve Score (AUC)
Fair latent variable model effectively removes bias, reducing demographic parity difference, while maintaining the same AUC score. This indicates that model is able to balance fairness with predictive capabilities. Observed label is binary.

[Multi-Class] Cannabis Consumption

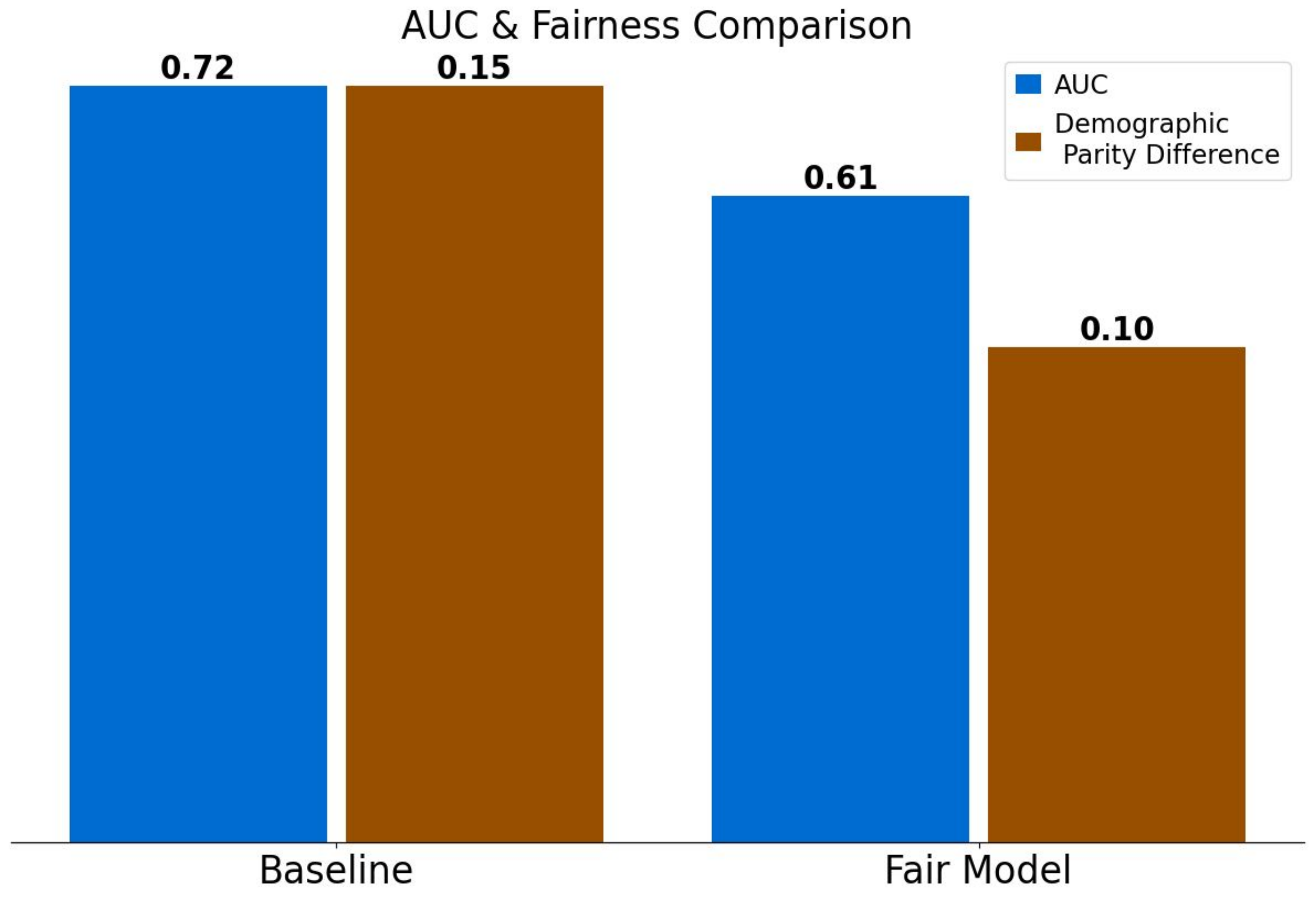
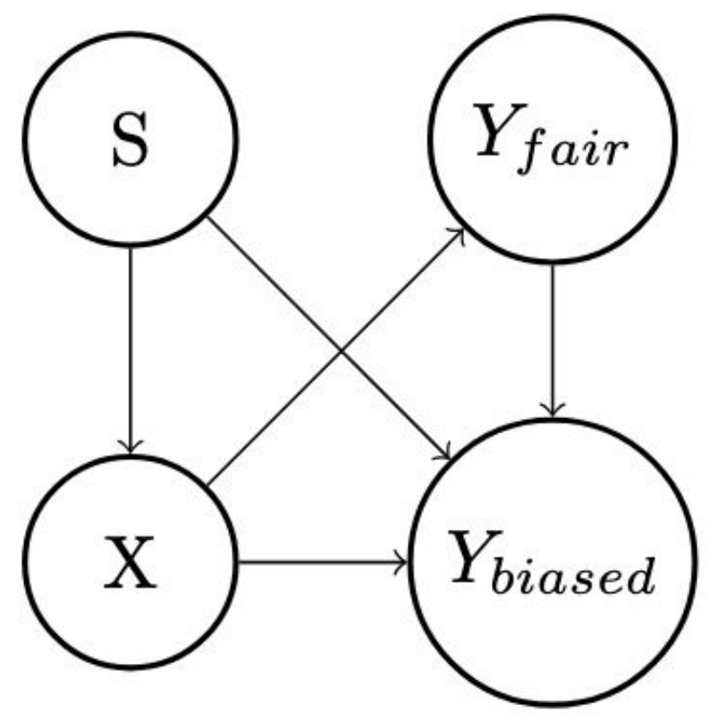


Figure 4: Trade-off Between Performance and Fairness
In a setting where the observed label is multi-class, the model reduces demographic parity difference, indicating the mitigation of label bias. However, there is a drop in the area under the curve (AUC) score, highlighting a common fairness-performance trade-off.

Underlying Causal Graph



Y_{biased} is generated from Y_{fair}, with its probabilities affected by the sensitive feature

With this, we can recover latent Y_{fair} by establishing independence

Future Works

- Explore the model's effectiveness when the sensitive feature is non-binary, in scenarios where the biased labels are either binary or multi-class
- Effectively adjust the model so that it performs well for cases where the biased label has more than 2 classes
- Provide clear guidance for situation in which accuracy should be prioritized over independence and vice versa when working with the model

Discussion

This quarter, our work focused on developing a fairness-aware algorithm to mitigate label bias by recovering the fair ground truth label. By leveraging an encoder-decoder framework with adversarial biasing, we ensure that the identified sensitive feature is independent of the observed biased labels. With that, we found the fair latent label. Our approach is evaluated on well-established fairness datasets, demonstrating its ability to produce fair and unbiased labels. As data-driven decision-making continues to shape critical domains like healthcare and criminal justice, our method provides a robust solution for addressing structural biases and enhancing the trustworthiness of AI systems.

Acknowledgements

Thank you to Dr. Babak Salimi, Parjanya Prashant, and Dr. Umesh Bellur for their support their support and guidance!

References

[1] **Choi, Y., Dang, M., & Van den Broeck, G. (2021). Group Fairness by Probabilistic Modeling with Latent Fair Decisions.** *Proceedings of the AAAI Conference on Artificial Intelligence,* 35(13), 12051-12059. https://doi.org/10.1609/aaai.v35i13.17431

[2] **Prashant, Parjanya Prajakta, Seyedeh Baharan Khatami, Bruno Ribeiro, and Babak Salimi.** 2025. "Scalable Out-of-Distribution Robustness in the Presence of Unobserved Confounders." *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS).* https://openreview.net/forum?id=eIyOtZ9tgl