Adversarial Latent Feature Augmentation For Fairness

Hoin Jung, Junyi Chai, Xiaoqian Wang

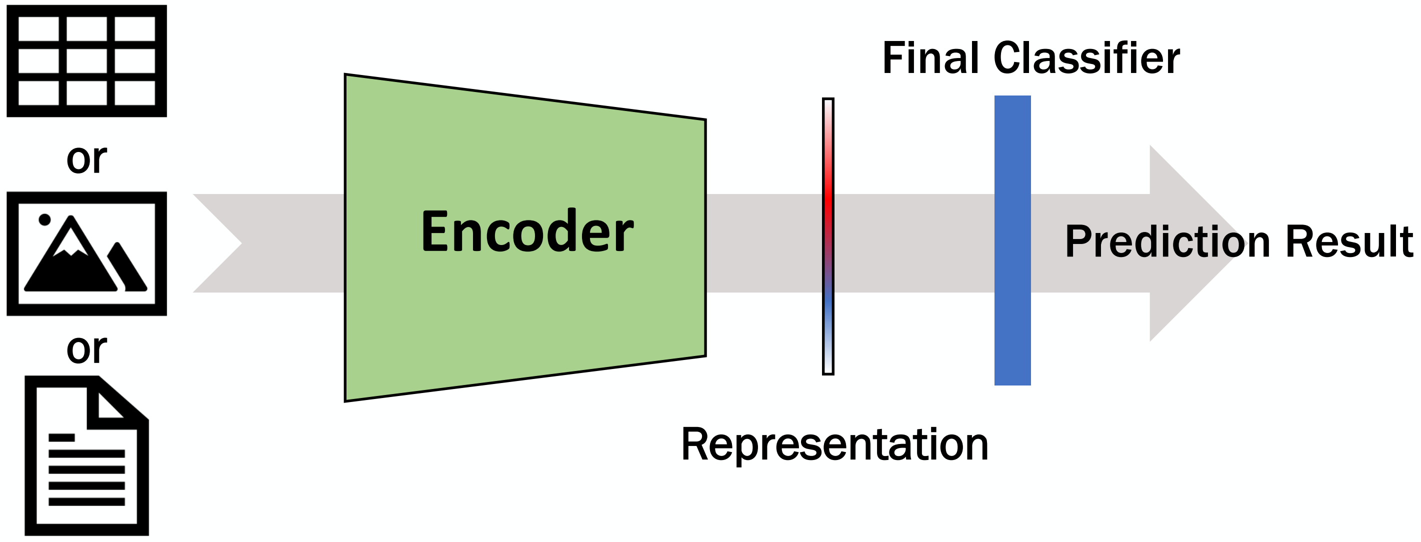
PURDUE UNIVERSITY

Elmore Family School of Electrical and Computer Engineering

Background

Typical Classification Tasks

- Most classification tasks across data modalities (e.g., tabular, image, text) share a common structure:
  - An **encoder**
  - A **latent representation**
  - A **final classifier**

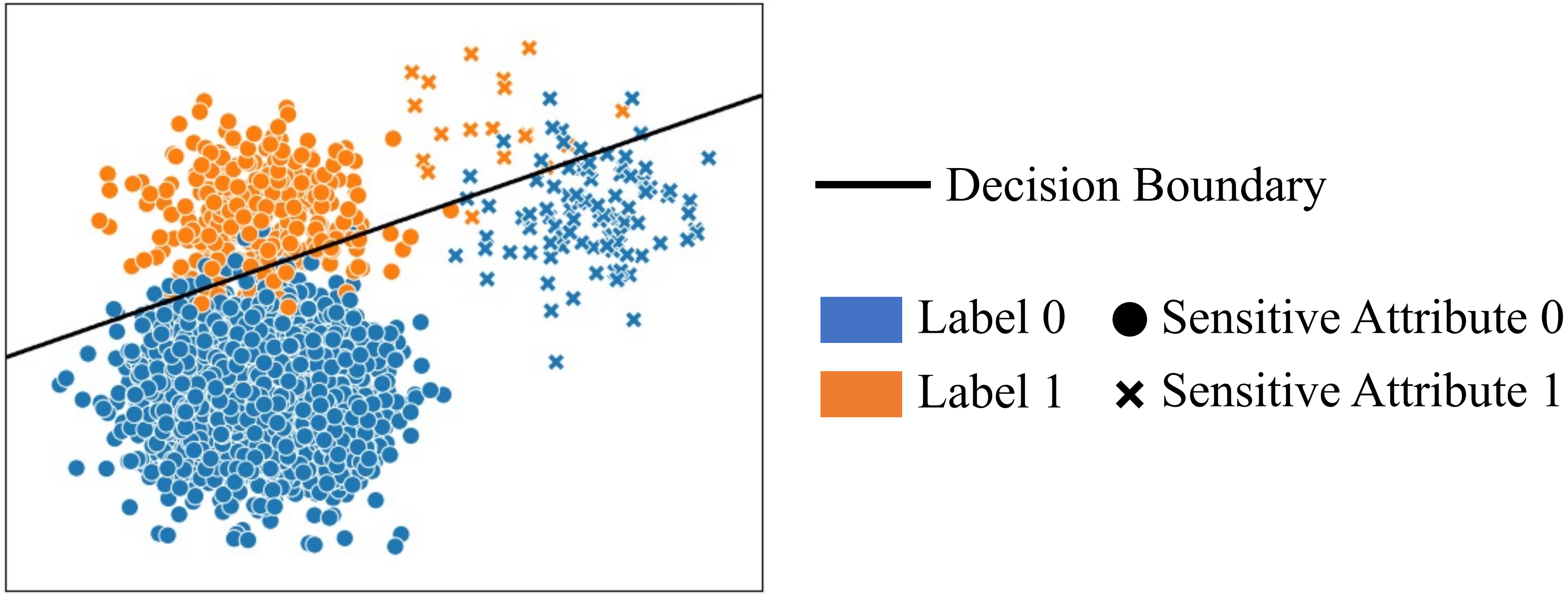


- The output prediction is based on this pipeline.

Motivation

Property of Unfair Classification

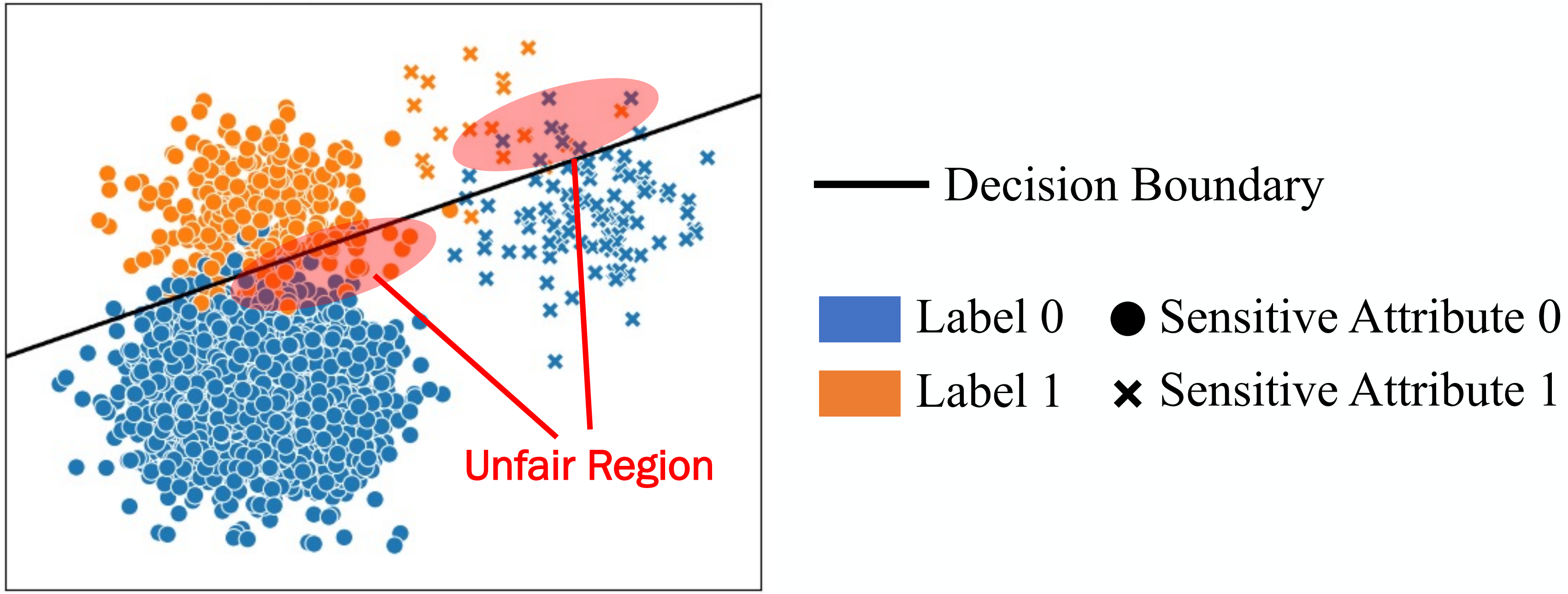
- Unfair representations can lead to biased decision boundaries and, consequently, unfair predictions.
- However, achieving fair representations typically requires re-training the entire model, which is computationally intensive.



Motivation

Property of Unfair Classification

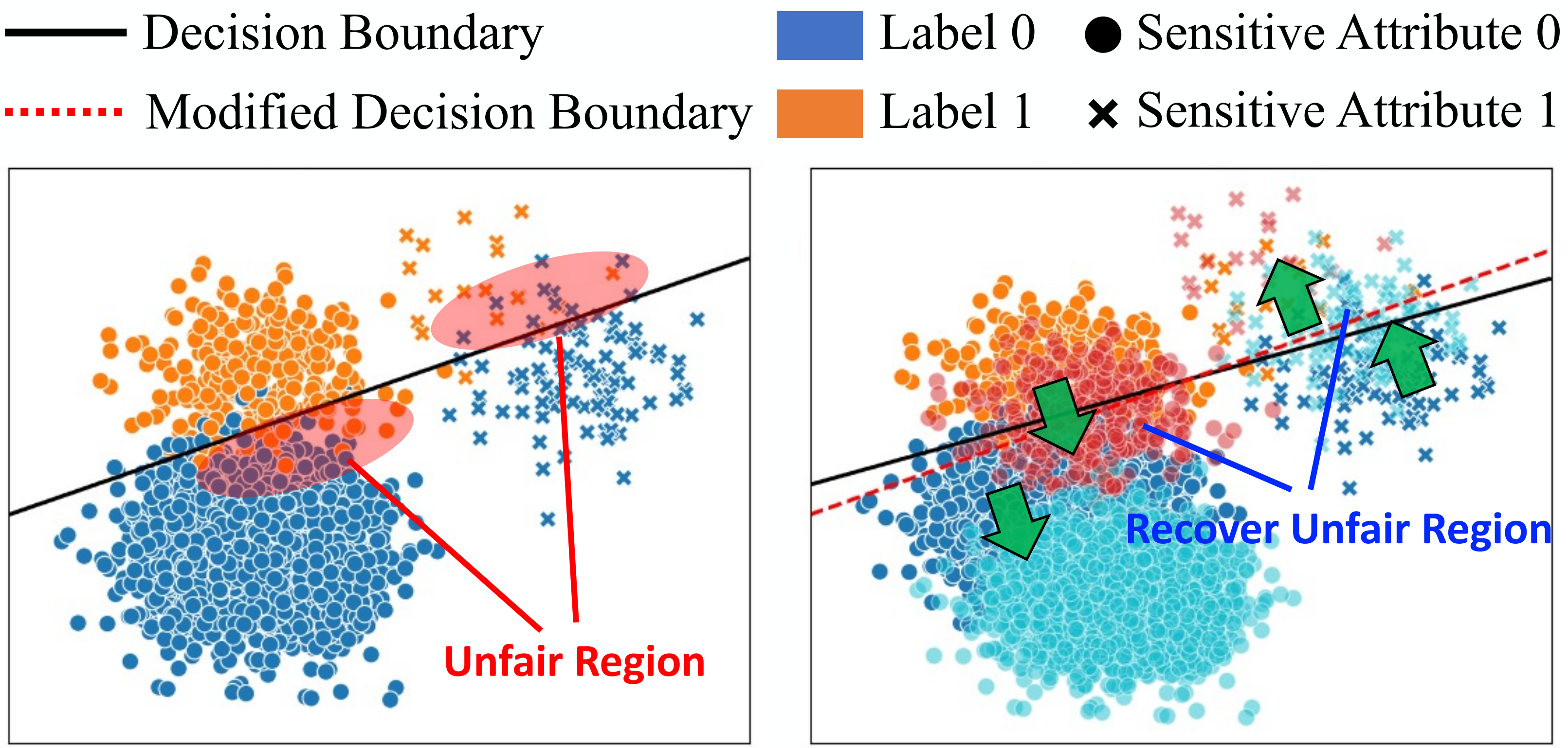
- We propose Adversarial Latent Feature Augmentation (ALFA) to mitigate unfairness without re-training the model.
- ALFA manipulates data directly in the latent representation space.
- We first identify unfair regions that contribute to biased decision boundaries.



Motivation

What is the Unfair Region?

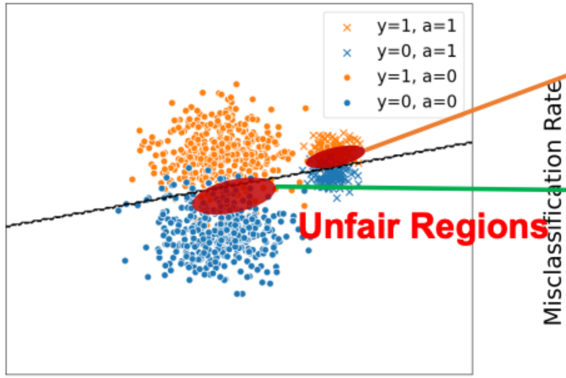
- **Unfair Region**: A subspace that highlights areas where misclassification rates for certain demographic groups are disproportionately high.
- Correcting this region helps to improve fairness in predictions.



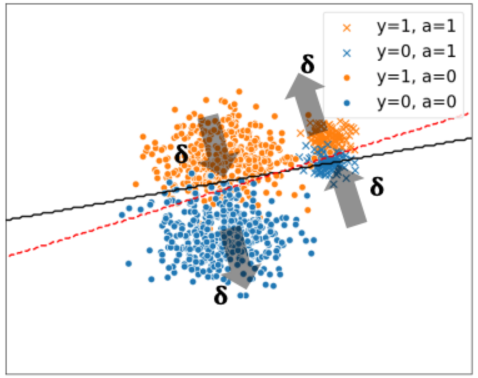
Motivation

What is the Unfair Region?

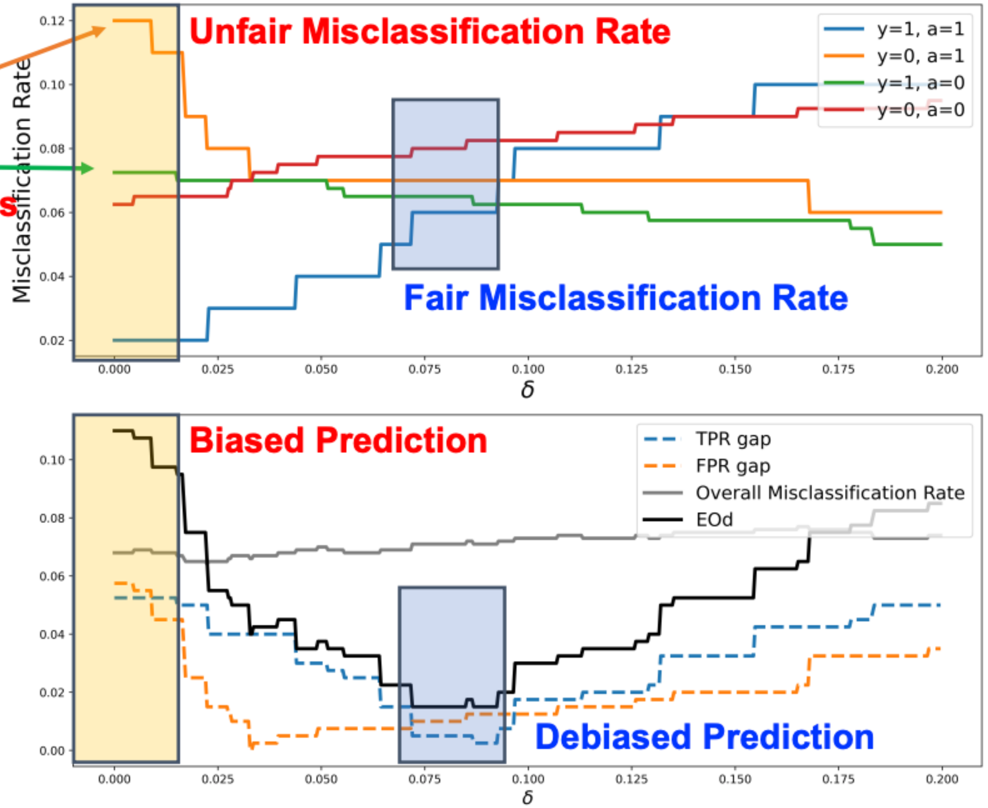
- How do we correct the unfair region?
  - By **perturbing** latent features toward these regions, we can gradually shift subgroup misclassification rates.



(a) Synthetic Data



(b) Perturbation on Synthetic Data

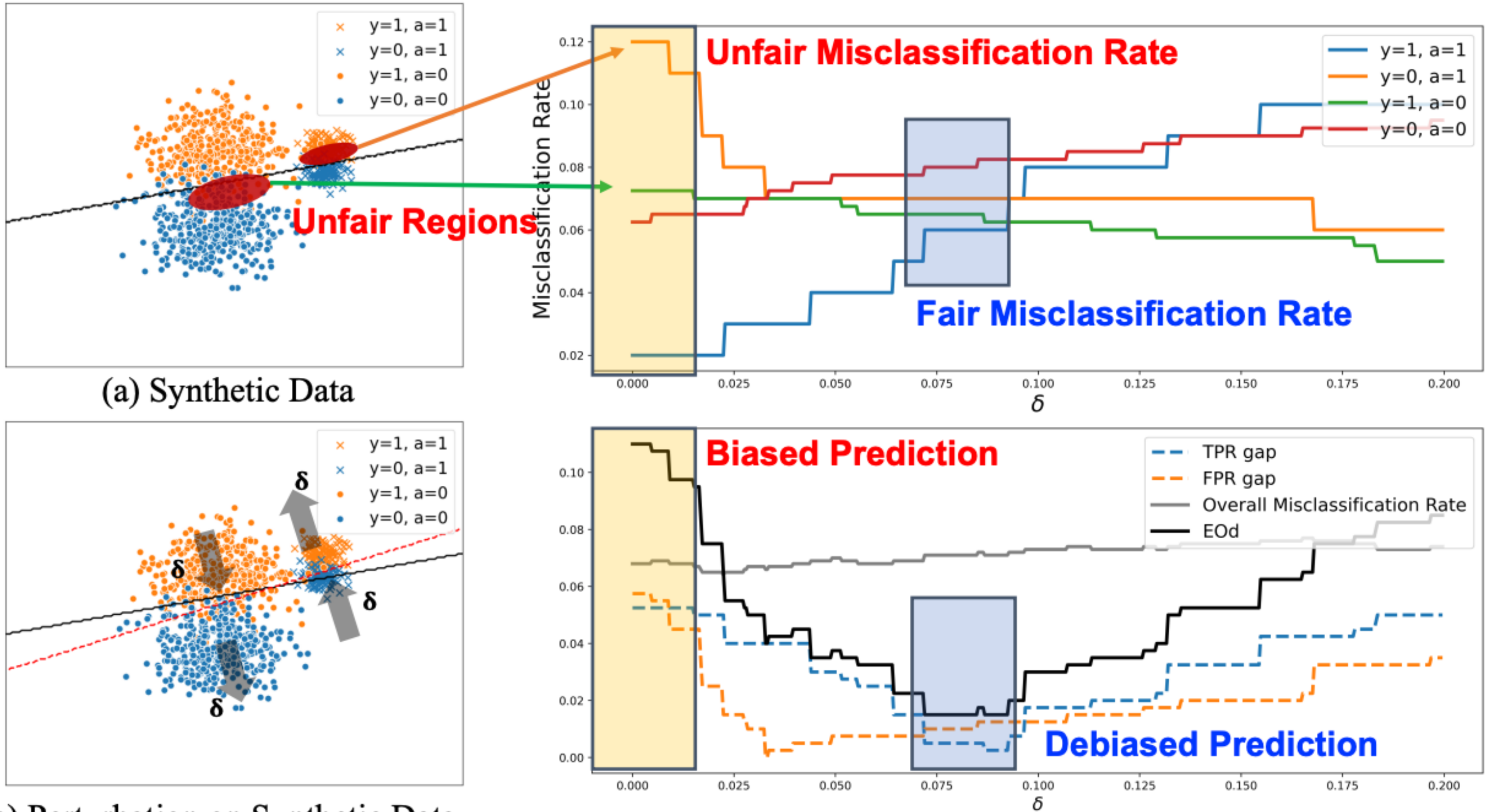


(c) Misclassification Rate VS. **δ**

Motivation

What is the Unfair Region?

- How do we correct the unfair region?
  - There exists an optimal perturbation δ that equalizes misclassification rates across subgroups.



(a) Synthetic Data

(b) Perturbation on Synthetic Data

(c) Misclassification Rate VS. $\boldsymbol{\delta}$

Proposed Method

Adversarial Latent Augmentation for Fairness

- How do we determine the correct direction and magnitude of perturbation?
- We employ an adversarial attack guided by a fairness constraint.
- Specifically, the **covariance** between the sensitive attribute and classifier output.

$$\mathcal{L}_{\text{fair}} = |Cov(a, g(\tilde{\boldsymbol{z}}))| = \left|\mathbb{E}\left[(a-\bar{a})\big(g(\tilde{\boldsymbol{z}}) - \mathbb{E}[g(\tilde{\boldsymbol{z}})]\big)\right]\right| \approx \frac{1}{N_p}\left|\sum_{i=1}^{N_p}(a_i - \bar{a})\big(d_i - \bar{d}\big)\right|$$

  where the sensitive attribute a, linear classifier g, latent feature $\boldsymbol{z}$, its perturbation $\tilde{\boldsymbol{z}} = \boldsymbol{z} + \boldsymbol{\delta}$, and the signed distance d from $\boldsymbol{z}$ to decision boundary,
- Counterintuitively, the perturbed latent features will be located in Unfair Regions.

Proposed Method

Adversarial Latent Augmentation for Fairness

- To preserve feature integrity, we minimize the Sinkhorn distance between original and perturbed features.

$$\max_{\|\boldsymbol{\delta}\|_2 \leq \epsilon} \Big(\mathcal{L}_{\text{fair}} - \alpha D(\boldsymbol{z}, \boldsymbol{z} + \boldsymbol{\delta}) \Big)$$

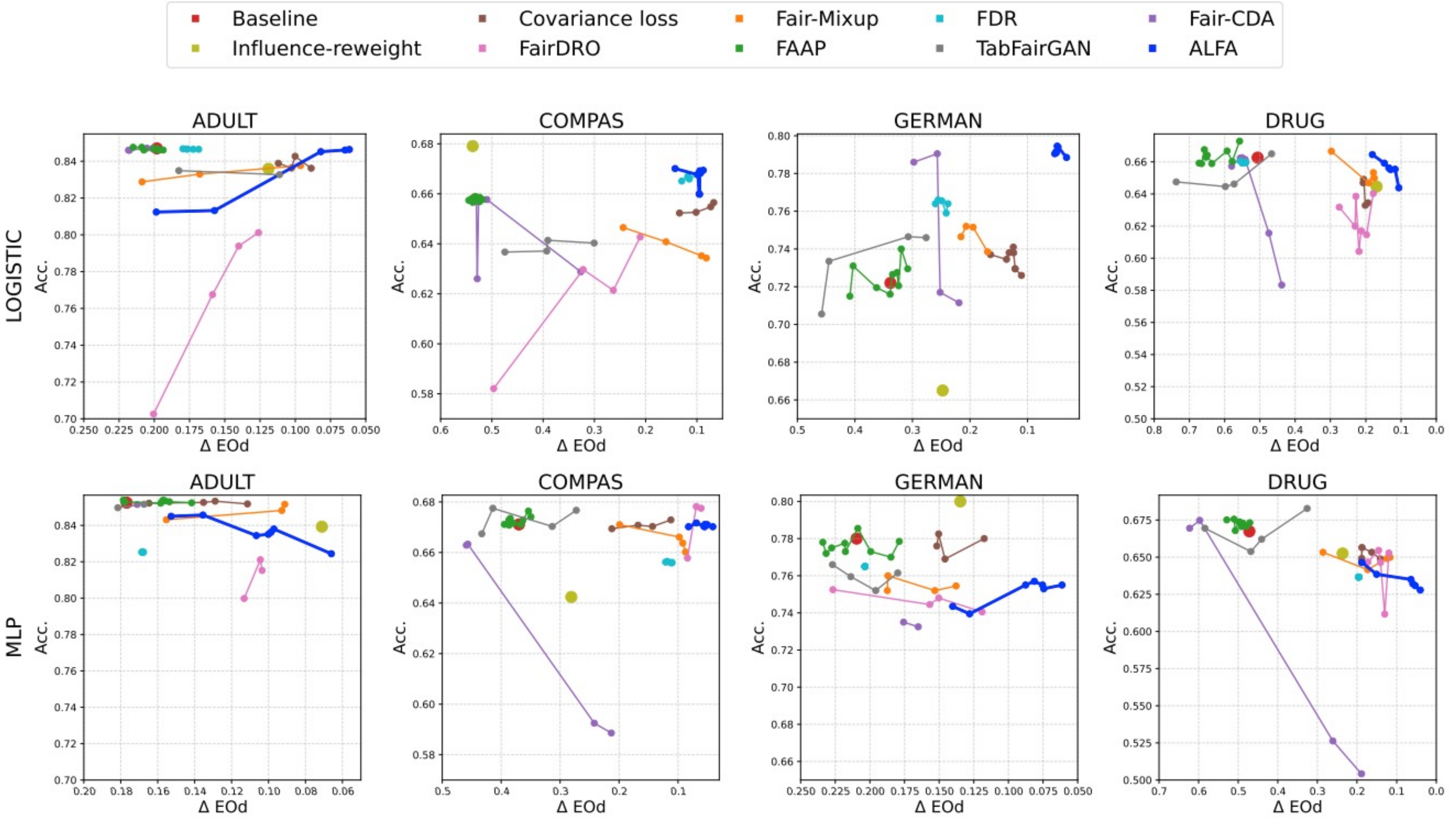
- The classifier g is then **fine-tuned** on both original and perturbed features, while The encoder remains frozen throughout this process.

$$\min_{\theta} \frac{1}{|\boldsymbol{X}_c| + |\boldsymbol{Z}_p|} \Big(\sum_{\boldsymbol{x}_i \in \boldsymbol{X}_c} \mathcal{L}_{\text{ce}}\big(g(f(\boldsymbol{x}_i)), y_i, \theta\big) + \sum_{\boldsymbol{z}_j \in \boldsymbol{Z}_p} \mathcal{L}_{\text{ce}}\big(g(\boldsymbol{z}_j + \boldsymbol{\delta}_j^*), y_j, \theta\big) \Big)$$

Experimental Results

Result on Tabular Datasets

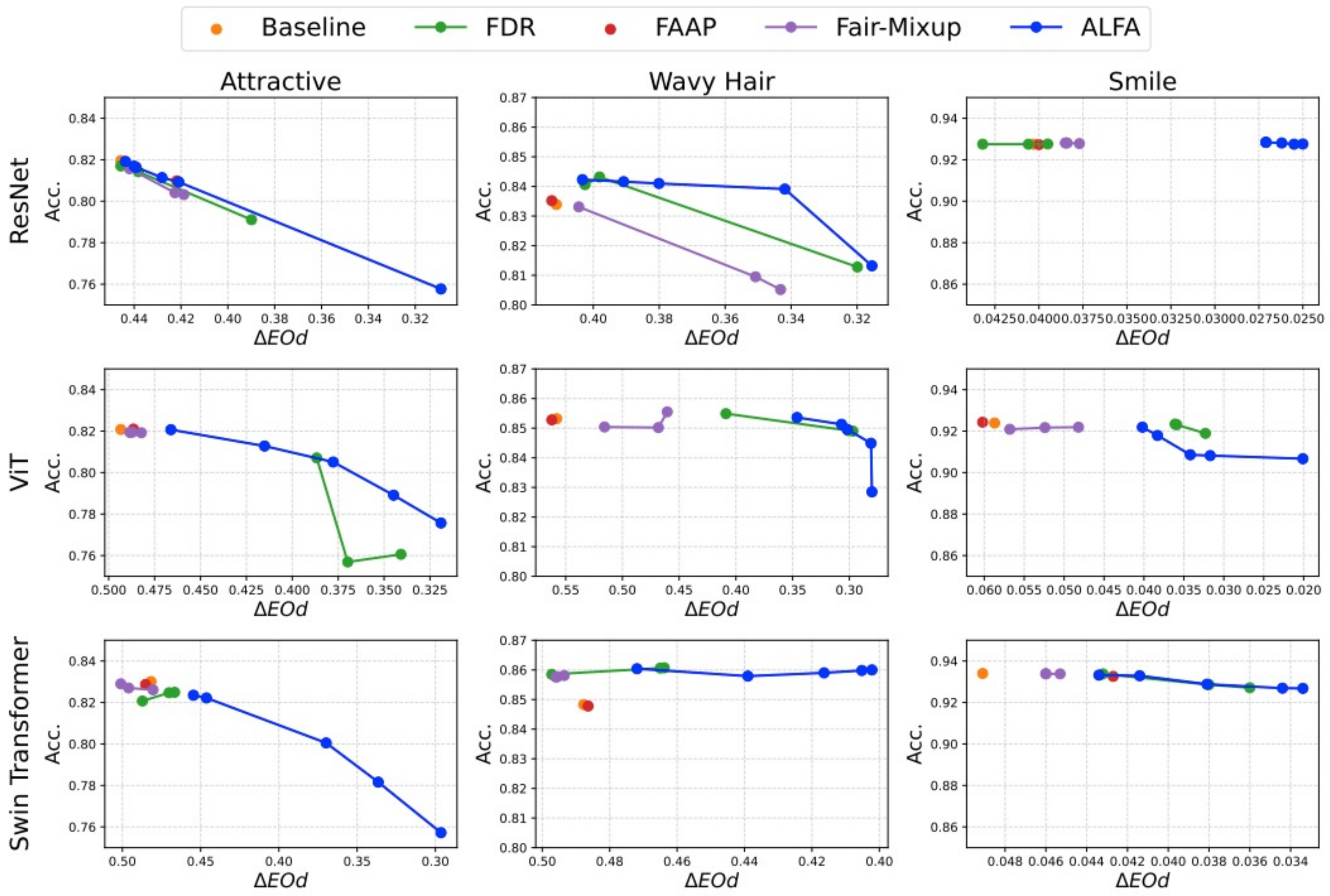
Sensitive Attribute: Gender



Experimental Results

Result on Image Datasets (CelebA)

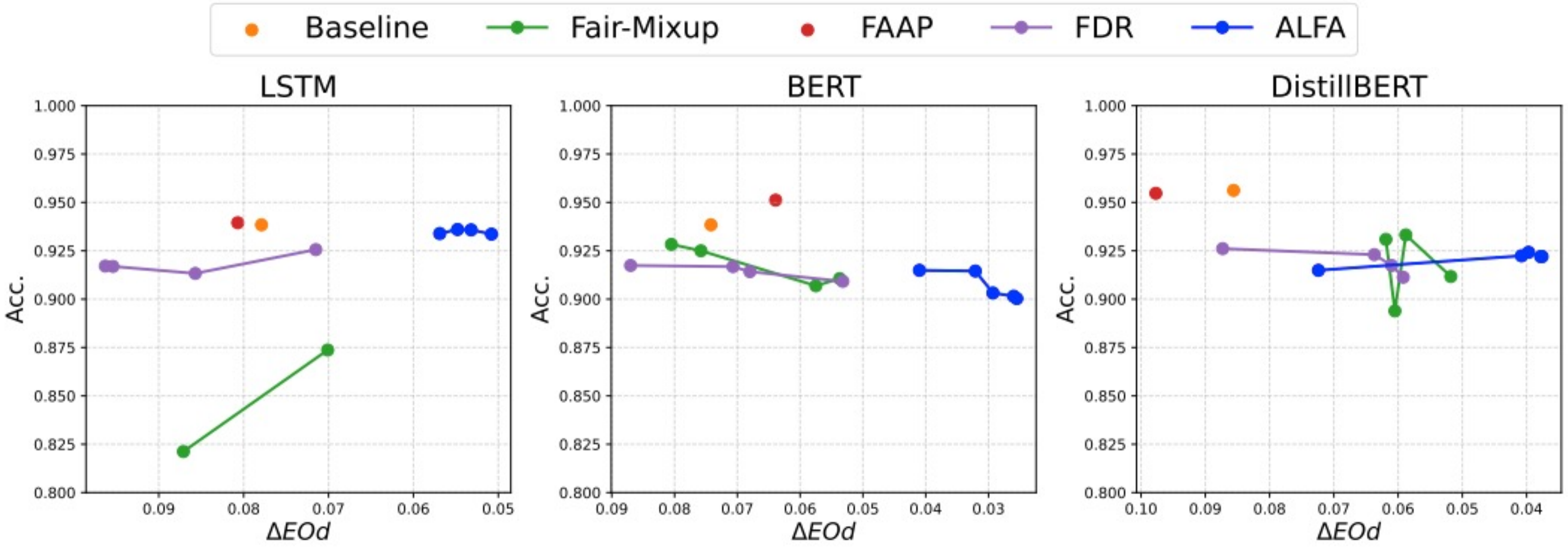
Sensitive Attribute: Gender



Experimental Results

Result on NLP Datasets (Wikipedia Toxicity)

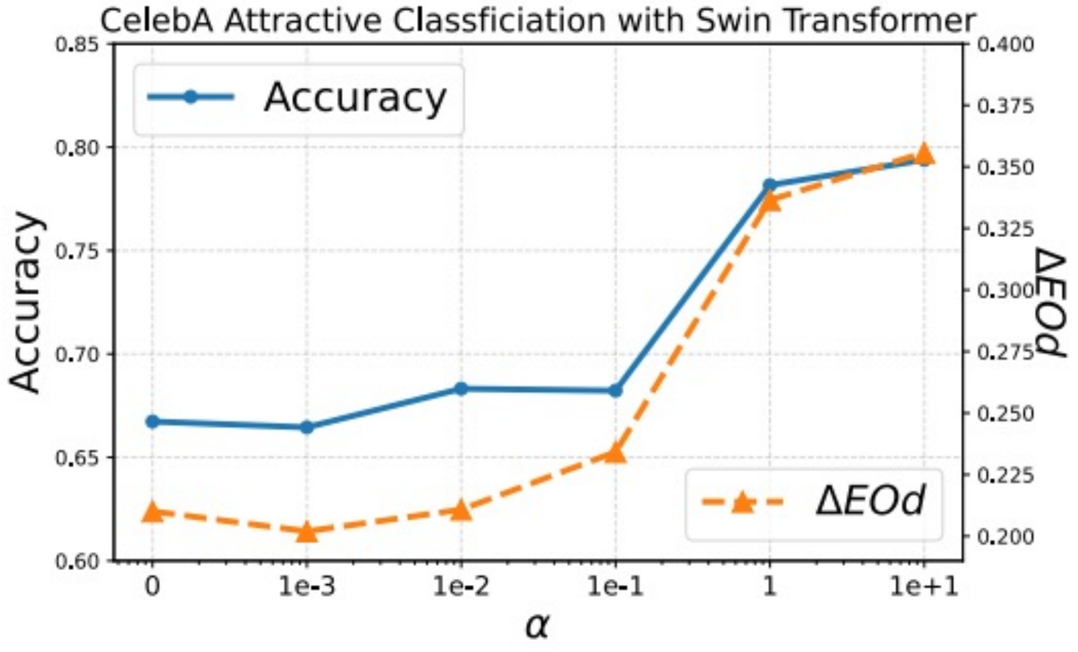
Sensitive Attribute: Sexuality Term

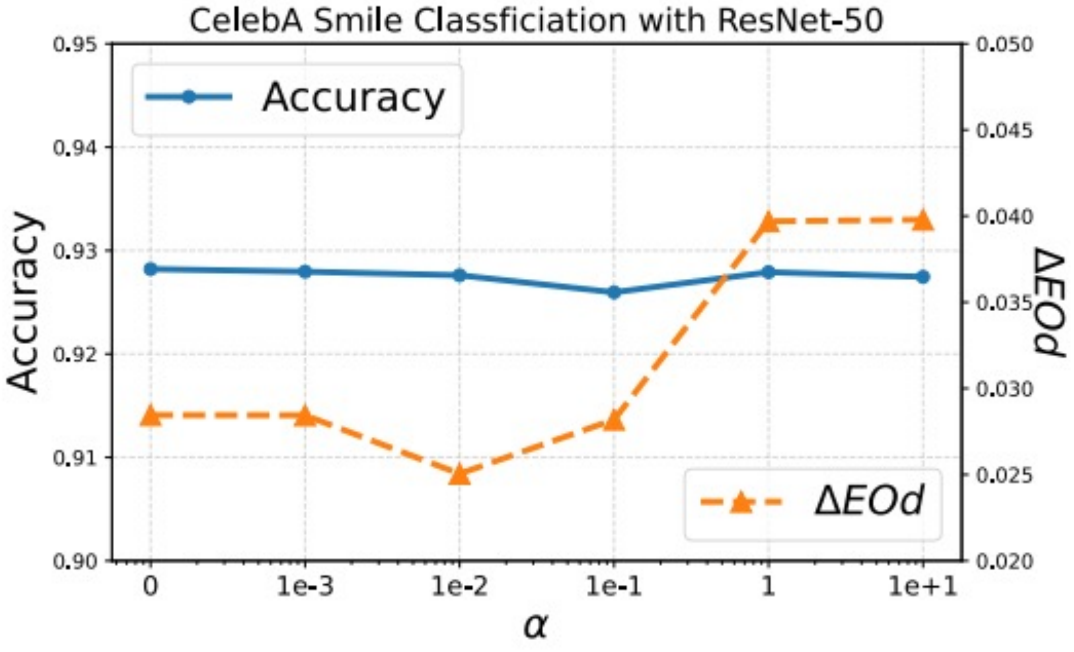


Result Analysis

Adversarial Latent Augmentation for Fairness

- Ablation Study
  - Investigated the effect of Sinkhorn Distance with parameter α.





Conclusion

Adversarial Latent Augmentation for Fairness

- **Performance-Fairness Trade-off**
  ALFA balances fairness and performance, achieving notable fairness improvements with minimal accuracy loss.
- **Key Advantages:**
  - Eliminates the need to re-train the encoder.
  - Perturbation step is performed **once**, prior to classifier fine-tuning.
  - Applicable across data types and encoders (operates in latent space).
  - Guarantees fairness improvement (**see Appendix A**).
  - Supports various fairness constraints.
- **Future Work:**
  - Extend ALFA to more domains and tasks.

Thank You

Hoin Jung

jung414@purdue.edu