"Can Simple Sampling Fix Bias in Face Gender Classification?"

Fairness Without Labels: Pseudo-Balancing for Bias Mitigation in Face Gender Classification

Haohua Dong[1,2], Ana Manzano Rodríguez[1,3], Camille Guinaudeau[4], Shin'ichi Satoh[1]

[1]National Institute of Informatics, Japan [2]INESC-ID, Instituto Superior Técnico, University of Lisbon, Portugal [3]University of Amsterdam, The Netherlands [4]LIMSI, CNRS / Université Paris-Saclay, France

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Motivation

Automated face gender classifiers often exhibit **significant performance disparities across demographic groups** (e.g., race, gender, age)

- **Biased training data** that under-represents non-Western populations like East Asians or Black faces
- **Collecting and retraining** with new, balanced labeled datasets is **expensive and time-consuming**

Key Ideas

- Achieve fairness without ground-truth demographic labels by using **demographically balanced data**
- Fine-tune a CNN using **pseudo-labelling** methods
- Enforce **demographic balance** during pseudo-labeling selection

Model & Datasets

- ResNet18-based CNN pre-trained on the Kaggle Gender dataset
- FairFace (FF) ➔ balanced across 7 racial groups (unlabeled training set)
- All-Ages-Face (AAF) ➔ predominantly Asian (test set)

Model	Acc [%] (Male / Female)	SR [%]
Baseline	73.28 97.94 / 48.61	49.63
Fine-tuned with FF	87.54 95.36 / 79.71	83.59

Method

Pseudo-Balancing (PB): Manually enforce gender class balance when sampling pseudo-labels during self-training

- Preserves model accuracy through confidence-based sample selection
- Prevents majority-class domination

➔**PB** works with FixMatch, FlexMatch and other pseudo-labeling techniques

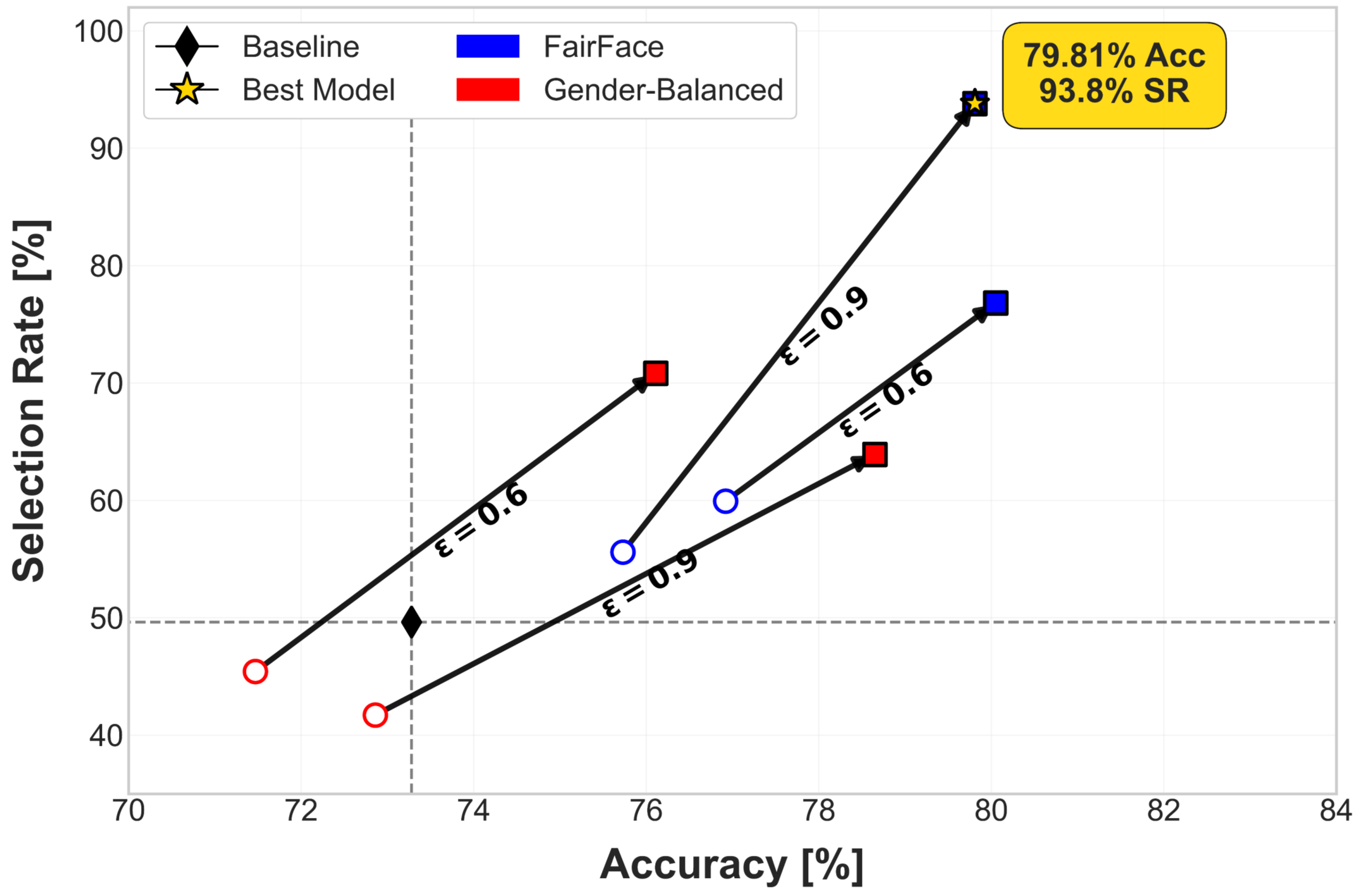
Experiments

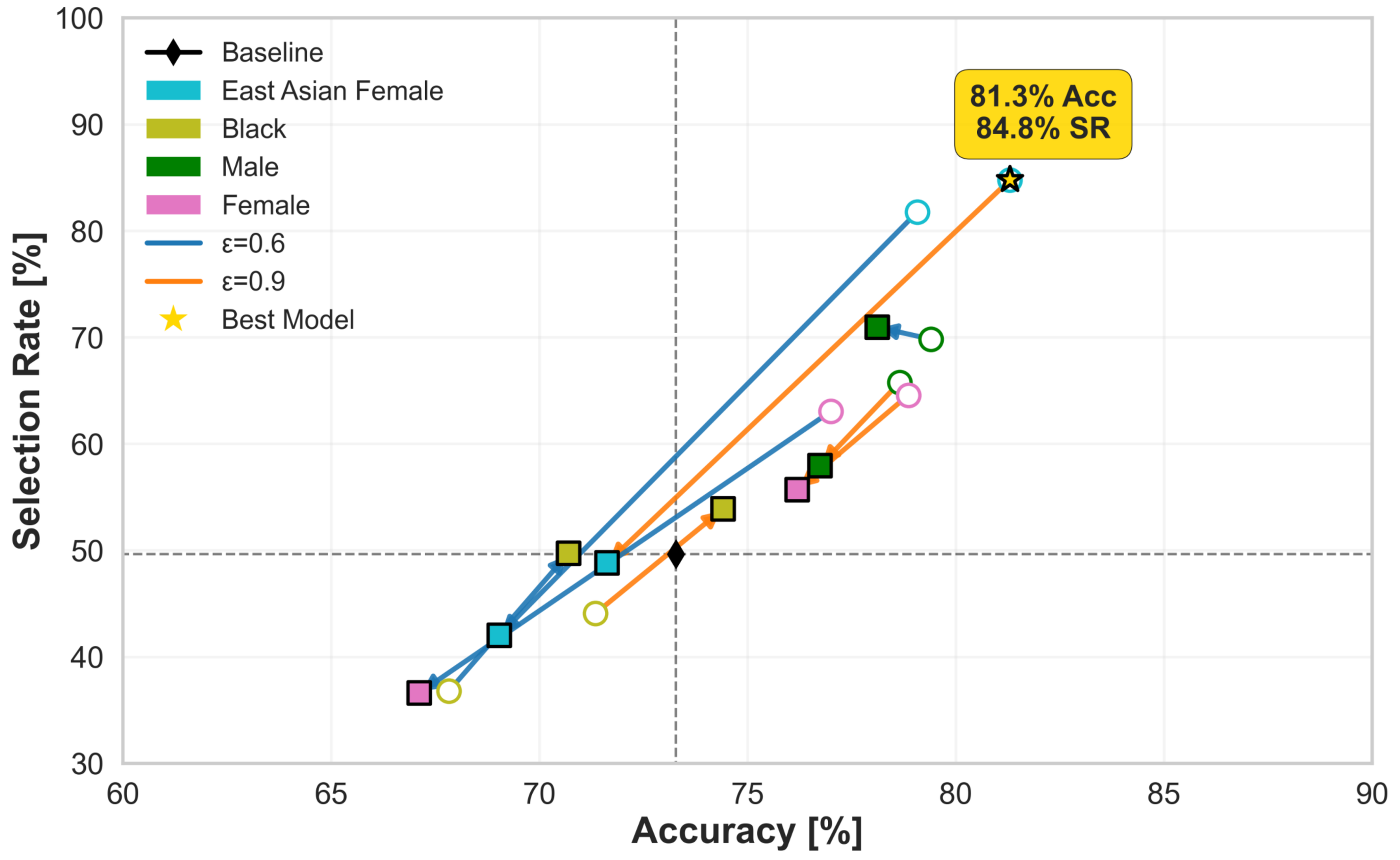
- Evaluated PB across **balanced, moderately biased, and severely biased FairFace** subsets (race and/or gender biases)
- Tested with **FixMatch** and **FlexMatch**, with/without **PB**, on **AAF** benchmark

FixMatch Analysis

- **PB improves accuracy and fairness in most cases!**

Pseudo-Balancing Effect on Accuracy and Selection Rate





Results

FixMatch	PB	Acc [%] (M / F)	SR[%]
FF ($\epsilon = 0.6$)	✓	80.05 91.12 / 69.98	76.80
FF ($\epsilon = 0.6$)	✗	76.92 96.20 / 57.65	59.93
FF ($\epsilon = 0.9$)	✓	79.81 82.37 / 77.26	93.80
FF ($\epsilon = 0.9$)	✗	75.73 97.35 / 54.12	55.59
East-Asian Female ($\epsilon = 0.9$)	✓	78.65 95.96 / 61.33	48.82
East-Asian Female ($\epsilon = 0.9$)	✗	81.30 87.99 / 74.60	84.78
Black ($\epsilon = 0.9$)	✓	74.41 96.72 / 52.10	53.87
Black ($\epsilon = 0.9$)	✗	71.35 99.05 / 43.66	44.08
Male ($\epsilon = 0.6$)	✓	78.11 86.33 / 69.89	80.96
Male ($\epsilon = 0.6$)	✗	79.40 93.50 / 65.29	69.83
Female ($\epsilon = 0.9$)	✓	76.19 97.86 / 54.52	55.81
Female ($\epsilon = 0.9$)	✗	78.86 95.84 / 61.89	64.58

FlexMatch	PB	Acc [%] (M / F)	SR[%]
FF	✓	83.36 89.62 / 77.10	86.03
FF	✗	79.18 67.75 / 90.61	74.77
Gender-Balanced	✓	67.49 94.93 / 40.05	42.19
East-Asian Female	✓	71.87 83.44 / 60.30	72.27
East-Asian Female	✗	56.04 76.07 / 36.01	47.37
Male	✓	62.92 52.61 / 73.21	71.86
Male	✗	54.16 19.02 / 89.30	21.28

Pseudo-Labeling with PB

- ❖ Diverse & balanced unlabeled data (**FairFace**)➔ boosts accuracy + fairness
- ❖ Biased & aligned subgroups (**East-Asian Female, Female**) ➔ reinforces model bias, hurts fairness
- ❖ Biased subgroups (**Male, Black**)➔spreads representation, counteracts bias

Conclusion

- **Simple & effective fairness boost ➔** PB works strongest with diverse, *balanced or moderately biased unlabeled* datasets
  - **+6.53%** accuracy and **+44.17% SR** over the baseline
- **Label-free ➔**scalable to real-world applications or different classification problems
- Limitations: less effective with *severely biased unlabeled* data

Full paper here!