Fair Robust Active Learning by Joint Inconsistency Supplementary Material

Tsung-Han Wu[1] Hung-Ting Su[1] Shang-Tse Chen[1] Winston H. Hsu[1,2]

[1]National Taiwan University [2]Mobile Drive Technology

1. Implementation Details

We introduce our computing infrastructure and training details in the supplementary material.

1.1. Computing Infrastructure

All experiments are conducted on an 8-core CPU personal computer with an NVIDIA RTX3090 GPU. The computational comparison shown in Tab. 4 in the main paper is evaluated on this machine.

1.2. Training Details

The overall framework of our proposed Joint INconsistency method (JIN) is provided in Algo. 1 in the main paper. Here we focus on the more detailed model training process, including the implementation of attack and defense methods as well as a conventional deep neural network pipeline.

Adversarial Attack and Defense. For all experiments, we utilize the python foolbox package [7] to achieve PGD-5 white-box adversarial attacks with maximum perturbation range $\ell_\infty = 4/255$ and step size $\alpha = 2/255$. We leverage the official TRADES loss [12] implementation[1] to realize the adversarial training with the same perturbation settings as the threat model and set the penalized term as $\beta = 6$.

It is important to note that our method is applicable to various adversarial training techniques, like PGD [5], since the phenomenon of joint inconsistency is widespread in all adversarial robust models [2, 9, 10, 11] and our method do not rely on any specific properties of the TRADES loss [12].

Deep Neural Network Pipeline. In the following, we elaborate on our deep neural network adversarial training and fine-tuning pipeline (corresponding to the "Adv-TRAIN" and "Adv-FINETUNE" in Algo. 1 of the main paper). For three datasets, we leverage the SGD optimizer to train our model with an initial learning rate of γ, a momentum of μ, and a weight decay λ. The batch size is set to B. Initially, we adversarially train the model for E_0 epoch. Then, for each active learning iteration, we adversarially fine-tune the model for E epochs. To make the whole training pipeline stable, we utilize the cosine annealing learning rate scheduler with a warm-up stage of initial E_w epochs in both the initialization and fine-tuning stage.

For the UTKFace dataset, we set $\gamma = 0.1$, $\mu = 0.9$, $\lambda =$ 2e-4, $E_0 = 100, E = 70$, $E_w = 10$, and $B = 32$. For the CINIC-10 dataset, we set $\gamma = 0.1$, $\mu = 0.9$, $\lambda =$ 2e-4, $E_0 = 110, E = 70$, $E_w = 10$ and $B = 64$. For the HAM-10000 dadtaset, we set $\gamma = 0.02$, $\mu = 0.9$, $\lambda =$ 2e-4, $E_0 = E = 50$, $E_w = 5$ and $B = 32$. Note that for a fair comparison, all models used in our experiments are not pre-trained.

Checkpoint Selection and Performance Evaluation. In fairness studies, a common tradeoff exists between average performance and fairness. This tradeoff can make it difficult to determine which model is the most representative, as changes in model weight can lead to minor positive or negative changes in performance or fairness. To ensure fair and reliable experimental results, we followed a protocol similar to prior related work and our baseline [8], recording results under fixed training epochs (when the model is converged) and repeating experiments three times to report their average. During evaluation, we prioritized fairness over average performance, as is consistent with mainstream fairness and robustness studies [6, 10].

2. Extensive Analyses and Results

2.1. Generalization on multiple sensitive groups

To validate the efficacy of various active selection methods on non-binary sensitive attributes, we further conduct experiments on the UTKFace gender prediction task and treat the four different races as sensitive groups (White, Black, Asian, and Indian). We use the same adversarial training protocol mentioned before with five active data selection rounds. The only difference is that we only randomly choose 10% D to initialize D_L and set merely 1% $|D|$ as the labeling budget for each round. The reason for using less labeled data is that the gender prediction task is

[1]https://github.com/yaodongyu/TRADES

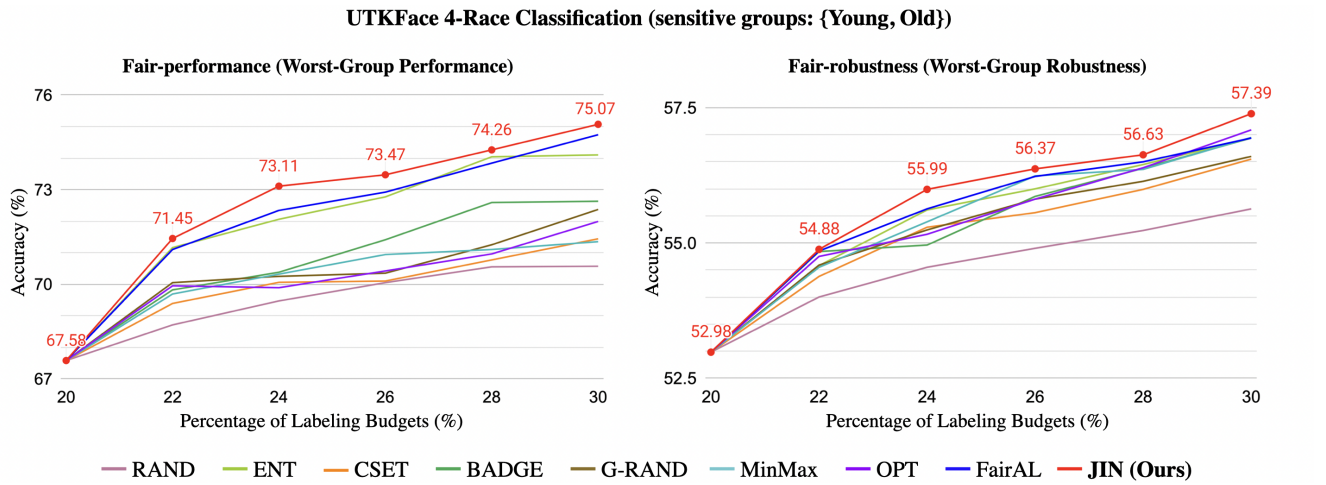

Figure 1. Active learning curves on the UTKFace dataset.

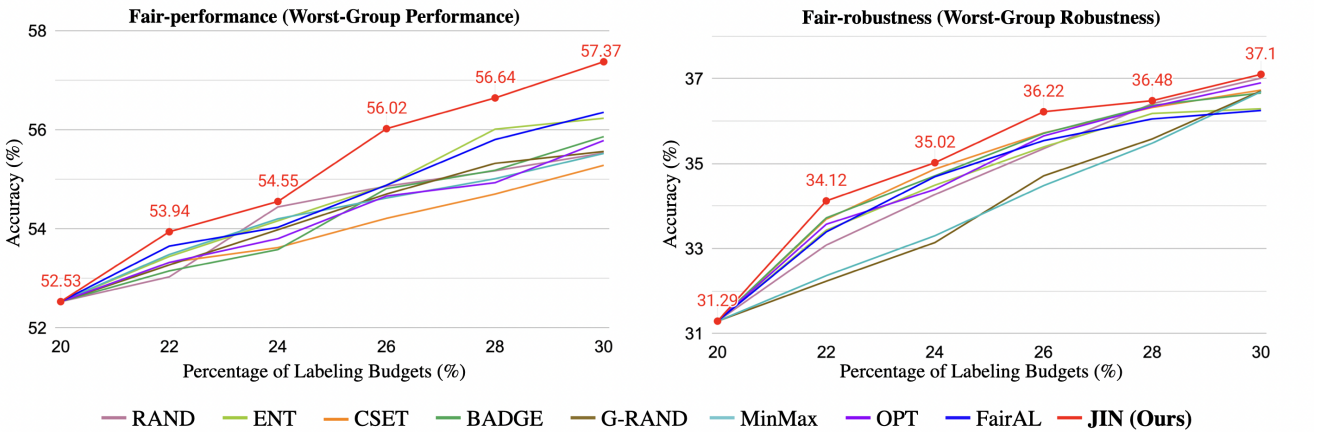

Figure 2. Active learning curves on the CINIC-10 dataset.

simpler than the 4-race prediction. As shown in Tab. 1, we outperform all baselines in both fair-performance and fair-robustness. The result demonstrates the generalization ability of our method.

	STD. Acc. (%)		Rob. Acc. (%)	
	Worst (↑)	Avg (↑)	Worst (↑)	Avg (↑)
Init. AT	77.74±0.66	81.03±0.33	67.61±0.21	70.84±0.36
RAND	78.57±0.31	82.34±0.10	69.14±0.30	72.34±0.14
ENT	80.70±0.59	84.01±0.10	69.79±0.14	72.67±0.21
G-RAND	81.08±0.22	82.95±0.31	70.38±0.23	**73.39±0.29**
FairAL	80.41±0.60	83.78±0.32	69.78±0.30	72.56±0.16
JIN	**82.77±0.27**	**84.96±0.25**	**70.56±0.13**	73.11±0.11

Table 1. **Comparison on the UTKFace gender prediction task under four ethnically sensitive groups.** Under the setting of multiple sensitive groups, our method still outperforms existing baselines in standard and robust minimax fairness.

2.2. Active Learning Curves

To enhance the credibility of our findings, we present the active learning curves for all tasks across multiple label budget combinations, illustrated in Fig. 1, 2, and 3. Our proposed JIN approach, leveraging the inconsistency properties in adversarial training, deliver the best fair-performance and fair-robustness across most label budget combinations.

We observed that among the baseline methods, FairAL and ENT perform better in datasets with severe data imbalance, such as UTKFace and Ham-10000. In contrast, data selection techniques that prioritize diversity, including RAND, CSET, and three group-aware methods, only perform better on the fair-robustness metric on a balanced dataset, CINIC-10. Notably, we found that group-aware data selection methods, including G-RAND, MinMax, and OPT, yielded very poor fairness results on HAM-10000, a highly data-imbalanced dataset. Even random selection or the initial adversarial training result performed better than

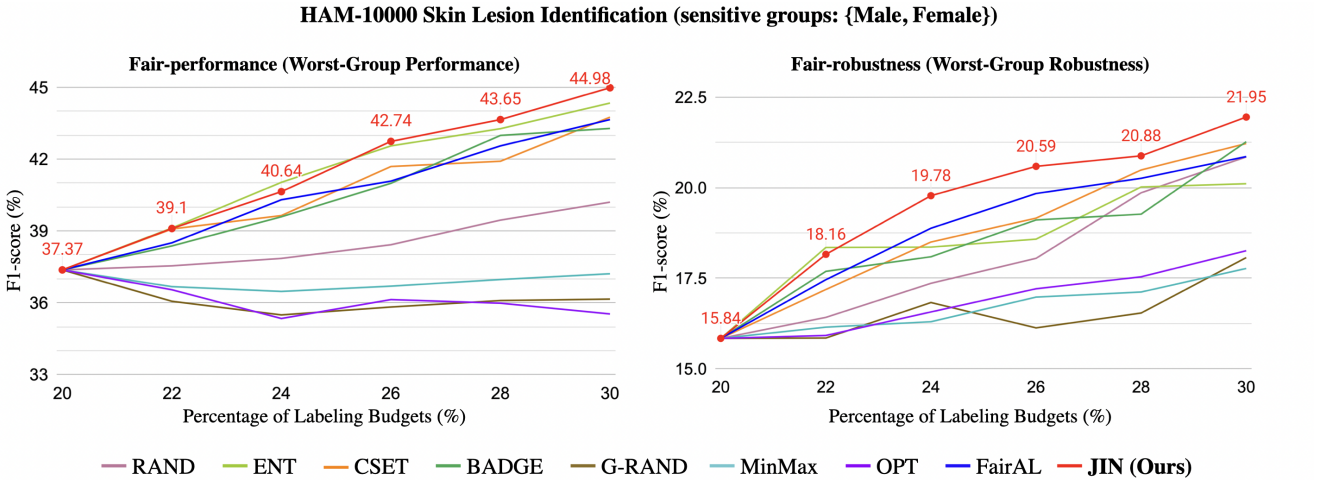

Figure 3. Active learning curves on the HAM-10000 dataset.

them. We infer that sampling bias can cause this negative impact. These results further underscore the limitations of the previous methods mentioned in the introduction section of the main manuscript.

2.3. Limitations and Future Work

We build on prior research on fairness [1, 3, 8] to address common biases through minimax fairness evaluation in discrete and limited groups, such as gender and race. Our JIN approach effectively reduces biases by primarily sampling from the worst-performing group. However, while minimax fairness is a widely-used evaluation, we acknowledge that this objective may have limitations [4]. Besides, we are open to exploring alternative budget distribution strategies among multiple groups, instead of solely sampling from a single worst group in our approach. Additionally, our approach is not directly applicable to continuous attributes, such as age, which we currently address by dividing them into bins for analysis. Future work will focus on exploring alternative budget distribution strategies and expanding the approach to handle continuous attributes or those without attribute information.

References

[1] Jacob D Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. Active sampling for min-max fairness. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, Sivan Sabato, et al., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 53–65. PMLR, 17–23 Jul 2022. 3

[2] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pages 854–863. PMLR, 2017. 1

[3] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021. 3

[4] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020. 3

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1

[6] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 466–477, 2021. 1

[7] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. 1

[8] Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. Adaptive sampling for minimax fair classification. *Advances in Neural Information Processing Systems*, 34:24535–24544, 2021. 1, 3

[9] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. 1

[10] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*, pages 11492–11501. PMLR, 2021. 1

[11] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019. 1

[12] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 09–15 Jun 2019. 1