Fair Robust Active Learning by Joint Inconsistency

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

Abstract

We introduce a new learning framework, ***Fair Robust Active Learning (FRAL)****, generalizing conventional active learning to fair and adversarial robust scenarios. This framework enables us to achieve fair-performance and fair-robustness with limited labeled data, which is essential for various annotation-expensive visual applications with safety-critical needs. However, existing fairness-aware data selection strategies face two challenges when applied to the FRAL framework: they are either ineffective under severe data imbalance or inefficient due to huge computations of adversarial training. To address these issues, we develop a novel Joint INconsistency (JIN) method that exploits prediction inconsistencies between benign and adversarial inputs and between standard and robust models. By leveraging these two types of easy-to-compute inconsistencies simultaneously, JIN can identify valuable samples that contribute more to fairness gains and class imbalance mitigation in both standard and adversarial robust settings. Extensive experiments on diverse datasets and sensitive groups demonstrate that our approach outperforms existing active data selection baselines, achieving fair-performance and fair-robustness under white-box PGD attacks.*

1. Introduction

While supervised deep learning methods have achieved remarkable success in a variety of computer vision tasks, the cost of labeling a large amount of data required for such a training paradigm is a huge burden. As a result, some utilize *active learning* (AL) techniques to achieve high performance by gradually selecting limited but valuable data for manual labeling [3, 13, 14, 19, 31, 40, 41].

Recently, in addition to reaching high performance, *fairness* and *robustness* have played increasingly vital roles in trustworthy visual applications. For example, a facial attribute recognition model is commonly used in biometric systems to protect the safety and confidentiality of individuals [35]. To ensure the safety and fairness of such systems, the model should neither exhibit low standard performance nor adversarial robustness against specific genders. Nonetheless, no existing work has explored the possibility of achieving this under limited annotations.

Observing this, we introduce a novel learning framework, **Fair Robust Active Learning (FRAL)**, generalizing conventional AL to fair and robust scenarios. By adopting this framework, we can attain *fair-performance* and *fair-robustness* while requiring only a small number of acquired labels. To elaborate, these two terms refer to the widely-used minimax fairness [11], which involves achieving the highest worst-case standard performance and adversarial robustness across sensitive groups. Fig. 1 illustrates the benefits of the FRAL framework for annotation-expensive visual applications with safety-critical needs.

Under the FRAL framework, due to the need to satisfy robustness constraints, existing fairness-aware active data selection methods would face two challenges. First, because performance disparities between classes are amplified under adversarial training [43, 44], existing methods [1, 34] may perform poorly on low-frequency classes of disadvantaged groups, resulting in poor fair-performance and fair-robustness. Second, as computational costs grow dramatically under adversarial training [32], some other approaches [2, 33] may suffer from expensive computations of measuring the expected fairness gain for each unlabeled sample during label acquisition.

To conquer these two problems, we propose a method called Joint INconsistency (JIN). Specifically, on benign data, disparities between classes and predicted errors increase from standard to robust models [37, 43, 44]. Similarly, for the robust model, incorrect outputs and class imbalance performance grow from benign to adversarial inputs [9, 43, 44, 46]. Hence, inspired by prior inconsistency-based active selection methods [42, 45], we leverage the two prediction disagreements to identify potential performance improvements and label imbalance mitigation. In practice, we estimate the worst-case group and then select the top-ranked samples in it for labeling at each active learning round. Unlike randomly drawing samples from the worst group [1, 34], we select more data in minor classes and thus work better under severe label imbalance. Also, by main-

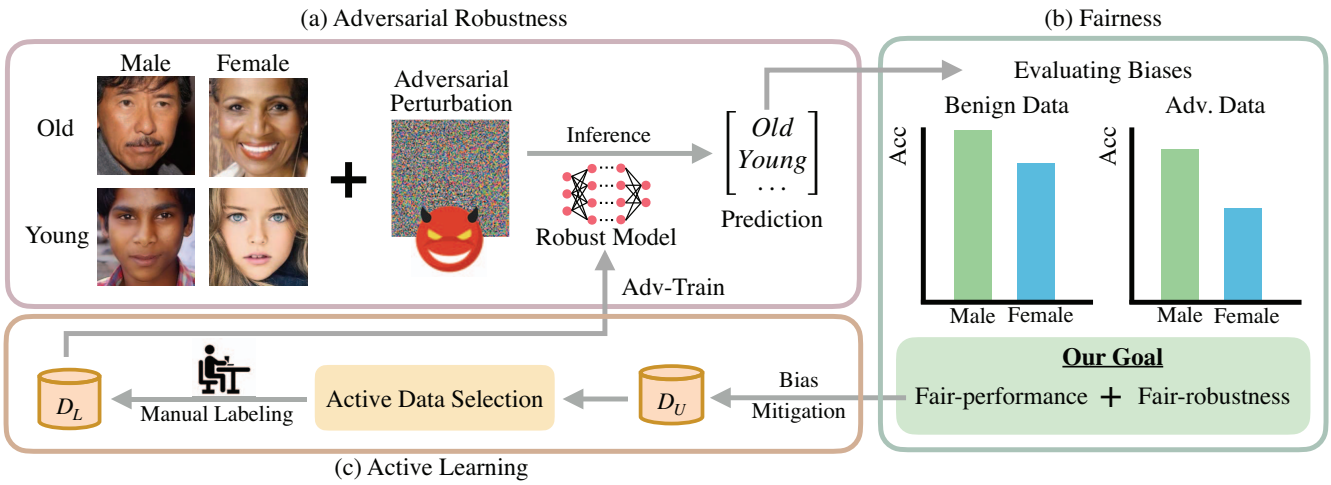

Figure 1. **A facial attribute recognition system based on our FRAL framework.** (a) In this system, a robust classifier predicting the young or the old is first adversarially trained with the current labeled dataset D_L against adversarial attacks. (b) However, the model may display biased performance on benign samples or unfair robustness on adversarially attacked images between male and female group. (c) Consequently, we iteratively acquire manual annotations of limited samples in the unlabeled pool D_U using a specific active data selection strategy. After being finetuned with these newly annotated data, the classifier can achieve better results in the underperforming (female) group, thus increasing both fair-performance and fair-robustness (return to (a)). With the FRAL framework, we can alleviate both performance and robustness discrimination with limited labeling effort.

taining a standard-trained model on a small labeled set instead of measuring the values of all unlabeled data by adversarial training [2, 33], we become computationally efficient.

We validate the efficacy of our method under white-box PGD attacks [22] in a wide variety of visual applications, including facial attribute recognition [48], object classification [10], and cell type identification [36]. Besides, we utilize several sensitive attributes in our experiments, including age, gender, and membership in a group. In various combinations, our JIN method outperforms existing active data selection approaches in both fair-performance and fair-robustness. Ablation studies prove the effectiveness of all our proposed components. Also, extensive experiments on different model backbones and multiple sensitive groups further demonstrate our generalization ability. To sum up, our contributions are listed as follows:

- We introduce a new learning framework, Fair Robust Active Learning (FRAL), practical for annotation-expensive applications with safety-critical needs.
- Under the FRAL framework, we design JIN, a novel data selection strategy, to solve the computation and class imbalance issue of prior fairness-aware methods.
- Our method surpasses existing active data selection baselines in both standard and robust fairness metrics under different experimental settings and datasets.

2. Related Work

Fairness in ML. Fairness is a fundamental problem in the field of ML. Many prior methods have pointed out that biases across sensitive groups are widely presented in ML models and datasets [18, 20, 27, 30]. Also, several debiasing training strategies [21, 24, 26, 38] are proposed to achieve fairness from different aspects, including making predictions independent of sensitive features [7, 12], yielding equal prediction odds on favored results [16], or maximizing the accuracy for disadvantaged groups [23]. Recently, some practices discussed fairness under adversarial attacked scenarios or via actively collecting a small amount of data. We elaborate on these methods in the followings.

Adversarial Robustness. Research on adversarial robustness can be roughly divided into attack and defense. Adversarial attacks aim to generate adversarial samples misclassified by ML models by adding the least perturbations to benign data, while defensive approaches seek to enhance model robustness against such attacks. Common adversarial scenarios are black-box and white-box threat models based on knowing all or nothing of the victim's ML model.

In the past few years, many classical attacks, such as FGSM [15] and PGD [22], produced adversarial examples by back-propagating loss functions. On the other hand, defensive methods utilized obfuscating gradients [4] or training with robust optimization [22, 46, 47] against attacks.

Recently, few approaches considered the intersection of adversarial robustness and fairness. Some discussed the bias between classes in adversarial training and proposed a training framework to mitigate this issue [44]; Others analyzed differences in robustness to adversarial samples between sensitive groups and developed a simple regularization method to address the problem [25].

Parallel to prior work of designing training algorithms, we delve into ways to achieve fair-performance and fair-robustness between groups via active data collection, which is more suitable for many real-world applications where manual labels are difficult to obtain.

Active Learning for Fairness. Conventional active learning aims to achieve high model performance by actively querying limited manual annotations. Common label acquisition strategies can be divided into two types: model uncertainty and data diversity. In the past, uncertainty-based methods [13, 14, 40, 41] collected data with the least model confidence for manual labeling, aiming to reduce model uncertainty after appending these data into training. Recently, some methods [3, 19, 31] increased the data diversity within a query batch, further improving the labeling efficiency.

More recently, several studies used active learning techniques to attain fair-performance. Some analyzed the efficacy of existing uncertainty-based active sampling methods under fairness evaluations [5]. Others designed fairness-aware data selection strategies by estimating expected unfairness reduction [2] or utilizing meta-learning [33]. Still others developed adaptive sampling policies specifically for fairness with theoretical foundations [1, 34]. However, these works merely focused on fairness without robust settings.

To the best of our knowledge, our FRAL framework is the first to target both fair-performance on benign data and fair-robustness to adversarial attacks. We have observed prior fairness-centric methods cannot afford the computations of adversarial training [2, 33] or struggle with fairness improvement under severe label imbalance [1, 34], leading to poor fair-performance and fair-robustness. To overcome these limitations, we propose an active selection method based on the properties of adversarial training, which sets our approach apart from all existing methods.

3. Method

3.1. Problem Definition

In this work, we introduce Fair Robust Active Learning (FRAL), a novel learning framework to reach fair-performance and fair-robustness between sensitive groups by actively acquiring limited labeled data. Specifically, let x and y represent an input sample and a target label; $z \in Z$ indicates a sensitive attribute or membership in a group. Taking the scenario in Fig. 1 as an example, x refers to a face image, y is a class in $Y = \{\text{Young}, \text{Old}\}$, and z is an attribute belonging to $Z = \{\text{Male}, \text{Female}\}$. Assume we have a training set D composed of a small labeled set $D_L = \{(x_i, y_i, z_i)\}_{i=1}^{N}$ and another unlabeled pool $D_U = \{(x_j, z_j)\}_{j=1}^{M}$. In our framework, we first train a robust deep learning model $M_R : x \to y$ with D_L by adversarial training. Then, once we observe performance or robustness discrimination between groups, we select a few samples in D_U with active selection strategies for manual labeling and further training.

In line with mainstream fairness studies [1, 34], we utilize the well-established minimax fairness [11] rather than predictive disparities between groups as the major objective to prevent achieving fairness by deliberately degrading the performance of dominant groups. In other words, all methods are required to achieve fairness by maximizing the standard performance and adversarial robustness of the least favorable group. Formally, given a white-box adversarial attack function $\mathcal{A}(x, \epsilon) \to \tilde{x}$ (ϵ is a pre-defined maximum perturbation range), a testing set D_T, and a robust deep learning model M_R, the fair-performance $\mathcal{F}_{per}$ and the fair-robustness $\mathcal{F}_{rob}$ are defined as the probability of correct predictions in the worst group as follows:

$$\begin{aligned} \mathcal{F}_{per} &= \min_{z \in Z} \mathbb{P}\{M_R(x) = y \mid (x, y, z) \in D_T^z\}, \\ \mathcal{F}_{rob} &= \min_{z \in Z} \mathbb{P}\{M_R(\mathcal{A}(x, \epsilon)) = y \mid (x, y, z) \in D_T^z\}, \end{aligned} \tag{1}$$

where D_T^z is a subset of D_T with the same attribute z.

Algorithm 1 Fair Robust Active Learning by JIN

Input: training set $D = \{D_L, D_U\}$, validation set D_V, adversarial attacks $\mathcal{A}$ with a perturbation range ϵ, maximum active rounds K, and labeling budgets B.
Output: An adversarial robust model M_R.
Init: $M_R \leftarrow$ Adv-TRAIN(D_L, ϵ)
for $k \leftarrow 1$ **to** K **do**
  $z^* \leftarrow$ EVAL$(M_R, \mathcal{A}, \epsilon, D_V)$ ▷ Get the worst group
  $X \leftarrow \{x \mid (x, z) \in D_U \ \wedge \ z = z^*\}$
  $I \leftarrow$ Get inconsistency scores for X via Eq. 2, 3, 4
  $X^* \leftarrow \{x \mid x \in X \wedge I_x$ is of the top-B values$\}$
  $Y^* \leftarrow$ Manual-Labeling(X^*)
  $D_L \leftarrow D_L \bigcup \{(x, y, z^*) \mid x \in X^*, y \in Y^*\}$
  $D_U \leftarrow D_U \setminus \{(x, z^*) \mid x \in X^*\}$
  $M_R \leftarrow$ Adv-FINETUNE(D_L, ϵ)
end for
return M_R

3.2. Overview

We propose a novel method called Joint INconsistency (JIN), general for various model architectures and adversarial training strategies under the FRAL framework. Algo. 1 shows the complete algorithm of our method, which consists of 3 main steps: (1) Model Initialization: Train the robust model M_R with the initial labeled set D_L by adversarial training. (2) Joint Inconsistency Sample Ranking: For each active learning round, estimate the worst-case group

and obtain joint inconsistency scores for all samples belonging to that group (Sec. 3.3). (3) Label Acquisition: Select top-ranked samples for manual labeling until running out of budgets, update D_L along with D_U accordingly, and fine-tune M_R to boost fairness objectives (Sec. 3.4).

3.3. Joint Inconsistency Sample Ranking

As stated in Sec. 1, prior fairness-aware selection methods either randomly sample data in the worst group [1, 34] or estimate expected fairness gain via meta-learning [33] or fine-tuning on all unlabeled data [2] for label acquisition. In adversarial training scenarios, however, the former suffer from severe data imbalance problems and the latter are confronted with an overwhelming computational burden. As a result, we design an efficient and effective sample ranking method via joint inconsistency to identify valuable samples for labeling, which is detailed below.

Worst group estimation. To simultaneously enhance both fair-performance ($\mathcal{F}_{per}$) and fair-robustness ($\mathcal{F}_{rob}$), we adopt a simple yet effective approach of acquiring valuable labeled data from the worst group for further training. This method builds upon the protocol in prior work [1,34], where we estimate the worst-case group using the validation set at the start of each active selection round. Specifically, we calculate the average of standard performance and adversarial robustness to identify the worst group. Then, we use our inconsistency scores for data selection, which will be explained in the following sections. We also provide additional discussion on this process in Sec. 4.3.

Inconsistency for fair-performance. To maximize the expected gain of $\mathcal{F}_{per}$ with limited labeling budgets, we select samples that have the highest inconsistency score between the robust model M_R and an auxiliary standard-trained model M_S. Inspired by prior active learning studies using disagreement between models as a data selection criterion [42,45], we hypothesize that prediction disagreement can indicate potential knowledge gain from annotation. As M_S has higher non-robust performance than M_R [37], samples with high disagreement are considered to be less confidently predicted by M_R. Therefore, additional labeling and training are needed to improve the standard performance. Specifically, assuming M_S and M_R are available, the performance inconsistency score I_x^{per} for an unlabeled sample x is defined as the prediction disagreement between M_S and M_R on benign data using the following equation:

$$I_x^{per} = D_{\text{KL}}(p(x, M_S) \,||\, p(x, M_R)), \tag{2}$$

where $p(x, M)$ indicates the predicted probability distribution of sample x from model M and $D_{\text{KL}}(\cdot||\cdot)$ means KL-divergence between the two distributions.

The motivation of Eq. 2 comes from theories in adversarial training. [37] proves that improving the robustness

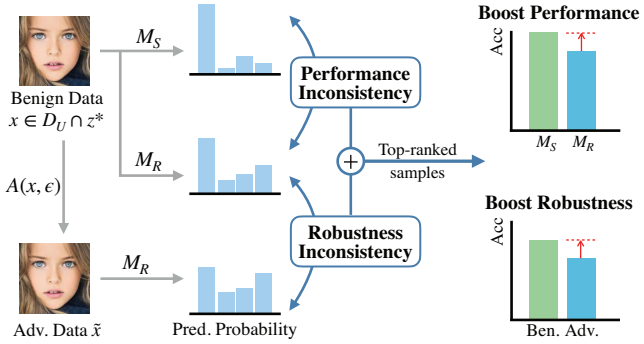

Figure 2. **The diagram of our proposed Joint INconsistency (JIN) method.** The performance inconsistency metric measures the divergence between the predicted probability distribution of benign data from the standard model M_S and the robust model M_R. Similarly, robustness inconsistency estimates the divergence of two distributions between benign and adversarial samples from the robust model M_R. By acquiring labels of limited samples belonging to the worst-case group with top-ranked inconsistency scores, the standard performance and adversarial robustness of the group can be enhanced, mitigating the unfairness issue.

of ML models would sacrifice performance on benign data, and [43, 44] observe that the issue of class-imbalanced performance on benign data becomes more severe under adversarial training. Based on these two studies, M_S has better performance and a milder class-imbalance problem than M_R on benign samples. Therefore, acquiring benign samples with utmost inconsistency between M_S and M_R could alleviate the performance drop of M_R. By calculating the inconsistency score with KL divergence between output distributions of the two models and selecting the top-ranked samples from the most unfavorable group z^*, we can improve the minimax fair-performance $\mathcal{F}_{per}$ and alleviate the class imbalance problem.

Our implementation is illustrated at the top of Fig. 2. For each active selection step, we first maintain an auxiliary model M_S by standard training with a small labeled set D_L. Then, all unlabeled benign samples are fed to M_S and M_R to obtain two different predicted probabilities. Lastly, we calculate the performance inconsistency scores I^{per} for all samples. The scores will be used in the label acquisition process, which we will cover later.

Inconsistency for fair-robustness. Similar to performance inconsistency, another inconsistency between benign and adversarial samples output by the same M_R is utilized to measure the expected improvement of $\mathcal{F}_{rob}$. Formally, the robustness inconsistency score I_x^{rob} of a sample x in the unlabeled set is defined as follows:

$$I_x^{rob} = D_{\text{KL}}(p(x, M_R) \,||\, p(\mathcal{A}(x, \epsilon), M_R)), \tag{3}$$

where $\mathcal{A}(x, \epsilon)$ is a white-box adversarial attack function identical to the definition in Sec. 3.1.

The intuition of Eq. 3 also stems from the properties of adversarial robustness. To begin with, model smoothness is considered to be highly correlated with adversarial robustness [9,46]. Moreover, under adversarial training, robust models generally exhibit larger performance disparities between classes given adversarial inputs than benign inputs [43,44]. Consequently, regardless of the data imbalance issue, the adversarial sensitive samples could be easily identified by measuring the benign and adversarial outputs from M_R and utilized to boost robustness. Specifically, in each active selection round, we use the KL distance of the two prediction distributions I^{rob} as another indicator to rank the samples from the worst group to enhance $\mathcal{F}_{rob}$. This process is illustrated at the bottom of Fig. 2.

So far, we have obtained two critical indicators, I^{per} and I^{rob}, that can identify the potential improvement of fair-peformance and fair-robustness. As our inconsistency metrics can address performance disparities between classes, we excel on severely class-imbalanced datasets compared to [1,34]. Besides, our method requires only multiple model inferences and low-cost standard training on a small labeled set rather than expensive adversarial finetuning on the unlabeled set. Therefore, it is much more computationally efficient than [2, 33].

3.4. Label Acquisition

As described above, the metrics I^{per} and I^{rob} are used to select samples that can boost fair-performance and fair-robustness respectively. Thus, to effectively maximize joint $\mathcal{F}_{per}$ and $\mathcal{F}_{rob}$ with the least manual annotations, we simply take the sum of these two scores as our final inconsistency metric for active data selection as follows:

$$I_x = N(I_x^{per}) + N(I_x^{rob}), \tag{4}$$

where $N(\cdot)$ is a standardization function that turns the value into an average of 0 and a standard deviation of 1.

After obtaining the score I for all samples from the worst group, we acquire the labels of top-ranked samples until running out of labeling budgets. Next, we append these labeled data into D_L and remove them from D_U. Finally, we fine-tune the robust model M_R with the updated D_L by adversarial training and proceed to the next round.

4. Experiments

4.1. Experimental Settings

Datasets. We use three different datasets in our experiments: UTKFace [48], CINIC-10 [10] and HAM-10000 [36]. For the UTKFace facial attribute dataset, we construct two sensitive groups (young and old) and perform the 4-race (White, Black, Asian, and Indian) classification task. For the CINIC-10 dataset, integrated by CIFAR-10 and partially synthesized ImageNet, we classify ten objects and treat membership in the two domains as sensitive attributes. For the HAM-10000 skin lesion dataset, we recognize seven cell types and use genders as sensitive attributes.

We directly use the official CINIC-10 data split in our experiments and construct the other two datasets ourselves due to the lack of such information. For the UTKFace, we first filter out face photos belonging to the "Other" race and put an age-related attribute tag on each photo based on the provided age metadata. Specifically, we evenly divide all faces into two groups (Young and Old) with a threshold of 30 years old. For the HAM-10000 dataset, we simply filter out repeated images and samples without a corresponding sex attribute. After the above preprocessing steps, we randomly split the two datasets into training and validation sets with a ratio of 7:3. We measure the effectiveness of all methods on the validation set.

Evaluation Metrics. We adopt minimax fairness as our primary fairness metric, which we refer to as fair-performance and fair-robustness in this paper (see Sec. 3.1 for a formal definition). For the adversarial robust scenario, we set the threat model as white-box PGD-5 attacks with maximum perturbation range $\ell_\infty = 4/255$ and step size $\alpha = 2/255$. To provide a comprehensive evaluation of fairness, we also include prediction disparities between the best and worst groups as an additional criterion [39]. Furthermore, we report the average group performance to investigate whether there is a significant trade-off between fairness and performance. For the UTKFace and CINIC-10 datasets, we use average accuracy to evaluate performance. However, due to the severe label imbalance in the HAM-10000 dataset, we adopt the F1-score as the performance evaluation metric, as suggested in prior work [29].

Training Protocol. We use MobileNetV2 [29] as the network backbone for both M_R and M_S owing to the great performance on adversarial training with high training and inference speed. In the model initialization and finetuning stage, we use the TRADES loss [46] with the same setting as our threat model to train our robust model. Note that in order to conquer serious label imbalance, we apply random oversampling when training models with HAM-10000 dataset. More training details and computing infrastructure are reported in the supplementary material.

As for the active learning setting, we perform five active data selection rounds ($K = 5$). For all three datasets, we randomly divide the training set D into 20% initial labeled set D_L plus 80% unlabeled set D_U as initialization. Then, the labeling budget B for each round is 2% of $|D|$.

Baselines. We compare our designed JIN method with eight active data selection baselines. For active learning baselines, in addition to random selection (RAND), we include three diverse and representative methods, including an uncertainty-based method (ENT [40]), a diversity ap-

Methods	UTKFace 4-Race Classification (sensitive groups: {Young, Old})						CINIC-10 Classification (sensitive groups: {CIFAR-10, ImageNet})					
	Standard Accuracy (%)			Robust Accuracy (%)			Standard Accuracy (%)			Robust Accuracy (%)		
	Worst (↑)	Disp (↓)	Avg (↑)	Worst (↑)	Disp (↓)	Avg (↑)	Worst (↑)	Disp (↓)	Avg (↑)	Worst (↑)	Disp (↓)	Avg (↑)
Init. AT	67.58±0.30	5.38±0.25	70.27±0.31	52.98±0.08	7.26±0.31	56.61±0.06	52.53±0.17	12.48±0.21	58.77±0.40	31.29±0.11	10.64±0.23	36.61±0.03
RAND	70.57±0.21	4.32±0.03	72.73±0.21	55.63±0.06	7.71±0.02	59.49±0.07	55.53±0.53	12.14±0.55	61.60±0.61	37.01±0.43	11.43±0.37	42.73±0.60
ENT	74.10±0.79	2.45±0.48	75.33±0.56	56.94±0.64	6.60±0.33	60.25±0.56	56.23±0.52	11.30±0.39	61.88±0.64	36.29±0.40	10.52±0.42	41.55±0.51
CSET	71.44±0.46	3.47±0.52	73.31±0.21	56.55±0.19	6.42±0.49	59.76±0.05	55.28±0.44	12.94±0.51	61.75±0.52	36.73±0.27	12.22±0.52	**42.74±0.39**
BADGE	72.63±0.20	3.53±0.23	74.31±0.13	56.94±0.40	6.07±0.20	59.98±0.50	55.86±0.38	11.96±0.44	61.84±0.37	36.66±0.30	11.04±0.38	42.18±0.29
G-RAND	72.37±0.32	2.15±0.26	73.45±0.23	56.60±0.04	6.07±0.33	59.63±0.13	55.56±0.43	**10.76±0.61**	60.94±0.66	36.71±0.35	10.02±0.41	41.72±0.59
MinMax	71.35±0.24	3.27±0.28	72.98±0.20	56.95±0.22	6.59±0.12	60.25±0.21	55.52±0.49	11.32±0.63	61.22±0.60	36.69±0.46	10.52±0.53	41.95±0.47
OPT	71.99±0.31	2.76±0.23	73.37±0.20	57.09±0.33	6.11±0.19	60.15±0.24	55.78±0.33	10.90±0.37	61.23±0.49	36.90±0.29	9.96±0.36	41.88±0.50
FairAL	74.74±0.31	2.20±0.13	**75.84±0.25**	56.94±0.16	6.64±0.17	**60.47±0.07**	56.35±0.45	10.98±0.44	61.84±0.58	36.25±0.29	10.40±0.33	41.45±0.37
JIN	**75.07±0.53**	**1.35±0.09**	75.74±0.49	**57.39±0.10**	**5.69±0.30**	60.10±0.25	**57.37±0.67**	11.16±0.52	**62.95±0.68**	**37.10±0.45**	**9.84±0.45**	42.02±0.48

Table 1. **Performance comparison with various active data selection methods on UTKFace and CINIC-10 datasets.** For both standard and adversarial robust settings, we report the results under three evaluation metrics, including minimax fairness, *i.e.*, highest worst group performance (Worst), performance disparity between the highest and the lowest group (Disp), and group average accuracy (Avg). To ensure the reliability of the experimental results, we conduct three experiments with different random seeds and report the average and standard deviation scores. The result shows that our JIN method achieves the highest standard and robust minimax fairness among all active data selection strategies. Also, our method obtains small performance disparity without incurring average performance loss in most cases. Detailed experimental results and extensive analyses are reported in the supplementary material.

proach (CSET [31]), and a hybrid strategy (BADGE [3]). Besides, we utilize four fairness-aware data selection baselines. They involve FairAL [2], a method leveraging expected fairness gain, and three adaptive sampling methods, including naive worst-group random selection (G-RAND), MinMax [1], and OPT [34]. Note that as FairAL focuses on demographic parity, we modify its source code to fit into our minimax fairness setting for fair comparisons. We do not include PANDA [33] in baselines due to unaffordable computations of meta-learning.

4.2. Main Results

Tab. 1 and Tab. 2 compare the effectiveness of various methods on three datasets. On the UTKFace and CINIC-10 datasets, the robust model favors the old group and samples belonging to CIFAR-10, respectively. On the HAM-10000 dataset, the robust model does not necessarily favor male or female group over several active learning rounds. Our proposed JIN method achieves the highest minimax fairness on all three datasets. Besides, in most cases, we deliver the lowest predictive disparity without degradation in average standard performance and adversarial robustness.

Performance comparison to active learning baselines. In most cases, our JIN method outperforms RAND, CSET, and BADGE by more than one standard deviation in two different fairness metrics with or without adversarial attacks. We observe that ENT obtains better fair-performance than three other active learning methods, which is identical to previous related research [5]. Still, our method reaches better fair-performance than ENT. Under the adversarial robust setting, our method achieves significant fairness advantages over ENT, including higher minimax fairness and lower predictive disparities. This indicates that in addition to selecting hard samples to boost fair-performance similar to ENT, our method can further identify adversarial sensitive samples in the least robust group for labeling.

Methods	HAM-10000 Skin Lesion Identification (sensitive groups: {Male, Female})					
	Standard F1-score (%)			Robust F1-score (%)		
	Worst (↑)	Disp (↓)	Avg (↑)	Worst (↑)	Disp (↓)	Avg (↑)
Init. AT	37.37±0.76	3.62±0.51	39.18±0.76	15.84±0.22	1.92±0.49	16.80±0.31
RAND	40.20±0.24	6.02±0.93	43.21±0.58	19.72±0.50	2.26±0.20	20.85±0.44
ENT	44.34±1.14	6.25±0.85	**47.46±1.54**	20.11±0.51	3.32±0.70	21.78±0.20
CSET	41.89±0.65	3.70±0.88	43.75±0.56	19.86±0.91	2.73±0.57	21.22±1.19
BADGE	43.28±1.00	3.52±0.46	45.04±0.83	20.07±0.22	2.39±0.39	21.27±0.29
G-RAND	36.15±1.37	3.46±0.61	37.88±1.67	16.65±0.36	2.85±0.89	18.07±0.66
MinMax	37.21±1.21	3.59±0.86	39.00±1.46	16.68±0.83	2.17±0.80	17.77±0.72
OPT	35.53±1.45	4.88±1.68	37.98±0.64	17.32±0.38	**1.88±0.38**	18.26±0.39
FairAL	43.65±0.99	3.53±0.77	45.42±0.68	19.64±0.54	2.44±0.81	20.86±0.83
JIN	**44.98±1.41**	**2.96±0.58**	46.46±1.48	**21.95±0.91**	2.28±0.66	**23.09±1.16**

Table 2. **Performance Comparison on the HAM-10000 dataset.** Similar to Tab. 1, our JIN method outperforms all baselines in standard and robust minimax fairness. Besides, under severe data imbalance, we observe three group-aware sampling methods (G-RAND, MinMax, OPT) cannot perform well. This issue is specifically discussed in Tab. 3.

Performance comparison to fairness-aware selection. Our JIN method outperforms group-aware adaptive sampling strategies (G-RAND, MinMax, OPT) by more than two standard deviations on fair-performance and by nearly one standard deviation on fair-robustness. Compared to the label-balanced CINIC-10 dataset, our method achieves greater benefits on two other datasets with label imbalance issue in terms of both fair-performance and fair-robustness. To investigate this issue, we analyze the correlation between label distribution and per-class performance on the HAM-10000 dataset in Tab. 3. Our JIN method acquires more samples from rare classes belonging to the worst group,

Methods	nv		mel		bkl		bcc		akiec	
	Class Freq (%)	F1-score (%) (STD / Rob)	Class Freq (%)	F1-score (%) (STD / Rob)	Class Freq (%)	F1-score (%) (STD / Rob)	Class Freq (%)	F1-score (%) (STD / Rob)	Class Freq (%)	F1-score (%) (STD / Rob)
Init. AT	80.11	91.22 / 84.05	4.10	19.04 / 3.44	8.12	41.17 / 10.53	3.30	30.76 / 11.69	2.74	24.06 / 0.00
RAND	80.07	92.41 / 86.67	4.12	23.81 / 4.17	8.16	40.90 / 12.12	3.51	43.51 / 13.67	2.54	19.36 / 0.00
ENT	69.84	**93.50 / 86.85**	5.71	39.13 / 11.11	12.36	35.00 / 9.71	4.65	39.22 / 14.67	5.88	31.82 / 14.26
G-RAND	80.31	92.68 / 86.32	4.04	20.12 / 2.38	8.01	40.87 / 7.79	3.34	27.61 / 14.78	2.52	25.64 / 0.00
FairAL	69.13	92.45 / 85.89	6.13	35.48 / 12.99	12.27	38.46 / 11.41	5.42	44.66 / 15.91	6.13	35.55 / 17.60
JIN	63.89	92.50 / 85.82	6.41	**41.67 / 22.22**	13.58	**42.00 / 14.67**	6.87	**45.28 / 16.67**	6.33	**36.73 / 19.67**

Table 3. **Correlation between class frequencies and per-class performance of the worst group on the HAM-10000 dataset.** We report the standard and robust F1-score for the top five high-frequency classes of the worst group. Compared with existing baselines, our JIN method is able to select more samples on less-frequent classes (mel, bkl, bcc, and akiec) and thus achieves significant improvement in standard and robust F1-score. As for the dominant class (nv), though selecting few samples, our method still achieves comparable results with prior approaches.

such as "mel" and "akiec", resulting in higher F1-scores for these classes compared to all other baselines, particularly G-RAND. Conversely, for the dominant class "nv", our method still achieves comparable results despite selecting fewer samples. The results show that existing adaptive sampling strategies under the FRAL framework cannot effectively handle datasets with uneven label distributions, but our proposed JIN method indeed addresses this situation.

We also compare our method with FairAL, which estimates the expected fairness increase per sample. Our results show that JIN outperforms FairAL in both fair-robustness and fair-performance. In terms of fair-performance, JIN outperforms FairAL by more than 1% in minimax fairness on the CINIC-10 and HAM-10000 datasets. Additionally, JIN achieves more than one standard deviation improvement in fair-robustness in all combinations compared to FairAL. This significant improvement indicates that our JIN method can effectively acquire more adversarially sensitive labeled samples for further training than FairAL.

Comparison on computational costs. We report the computation burdens of different active data selection methods in Tab. 4. Compared to ENT and G-RAND requiring low computations, our method takes more time because of maintaining a standard-trained auxiliary model. Despite this, our method achieves significantly better fair-performance and fair-robustness than the others. FairAL requires the most extended time, even more than twice as much initial adversarial training for data selection on the CINIC-10 dataset, making this approach impractical for real-world applications. This is mainly due to the large number of samples in the CINIC-10 dataset, which necessitates more adversarial finetuning to estimate potential fairness gain.

To sum up, our proposed JIN method leverages the distinct properties of adversarial training and achieves a great trade-off between fairness and computational costs. Among all active data selection baselines, we achieve the best fair-performance and fair-robustness using fewer than 30% of the initial adversarial training computations.

Methods	UTKFace	CINIC-10	HAM-10000
Init. AT	1h 4m 26s	1h 9m 31s	1h 22m 7s
ENT	14s	45s	12s
G-RAND	1m 5s	2m 17s	18s
FairAL	39m 47s	2h 21m 29s	19m 55s
JIN	10m 29s	19m 46s	15m 40s

Table 4. **Computations of different active selection strategies.** Referring to Tab. 1 and 2, we observe that the ENT and G-RAND methods require minimal computations but suffer from poor fairness outcomes. On the other hand, FairAL achieves better fair-performance, but at the cost of significant computational overhead. Our proposed JIN method outperforms all the above methods in terms of fair-performance and fair-robustness, while maintaining a reasonable computational cost.

4.3. Discussions

Below we first verify the efficacy of the proposed components, including two inconsistency scores and the process of selecting samples from a single group. Then, we show that our method is applicable to various deep neural networks. More discussions on multiple sensitive groups and potential limitations are left in the supplementary material.

Ablation studies. The efficacy of our proposed performance and robustness inconsistency metrics is shown in Tab. 5. A comparison of the first and second row indicates that the performance inconsistency metric enhances fair-performance and standard group average scores more than the robustness inconsistency metric. In contrast, the robustness inconsistency achieves better results in two adversarial robustness metrics. In the third row, where both metrics are used together, we observe a significant gain in average robustness and fair-robustness, but only a small drop in the standard performance counterpart. This suggests that using both metrics together may be the optimal strategy to achieve both fair-performance and fair-robustness under the FRAL framework. It is important to note that a model's performance on perturbed data (i.e., robustness) depends on its performance on benign data, but the reverse is not true. Therefore, using only the robustness inconsistency metric

can only achieve suboptimal robustness due to its poor standard performance as shown in the second row.

	STD. Acc. (%)		Rob. Acc. (%)	
	Worst (↑)	Avg (↑)	Worst (↑)	Avg (↑)
P	**75.18**±**0.47**	**75.84**±**0.27**	56.53±0.08	59.30±0.11
R	72.89±0.30	74.31±0.26	56.89±0.19	59.94±0.04
P+R	75.07±0.53	75.74±0.49	**57.39**±**0.10**	**60.10**±**0.25**

Table 5. **Ablation studies on the UTKFace dataset.** The letters P and R stand for performance and robustness inconsistency, respectively. Leveraging only robust inconsistency exhibits sub-optimal robustness due to its weak standard performance, while performance inconsistency boasts the best standard performance but lacks robustness. Using the two inconsistencies together strikes a balance by boosting both fair-performance and fair-robustness.

Discussion of worst group label acquisition. We discuss the process of acquiring labeled data from the worst group in our proposed JIN methods, which is a feature lacking in traditional active learning methods. As background, using a validation set is commonly accepted in mainstream fairness studies [1, 34, 44], and identifying the least favorable group is possible in safety-critical applications such as facial recognition systems [6]. Moreover, identifying the worst group through a validation set, rather than a training set, can be a better choice to prevent biases since adversarial training generally suffers from robust overfitting [8, 28]. In the following, we explore two critical issues related to allocating labeling budgets to a single worst group.

Firstly, we examine the rationale for allocating labeling budgets and identification to the worst group. Our empirical studies on three tasks show that the group with the least standard performance is of the worst adversarial robustness, consistent with prior studies on other datasets [25,44]. Thus, our method of identifying the single worst group for downstream active selection using averages of standard performance and adversarial robustness can contribute to both fair-performance and fair-robustness at the same time. We also find that primarily sampling from the single worst group is more effective than distributing budgets among multiple groups, which is in line with existing work [1, 34].

Secondly, we conduct additional experiments to provide a fair comparison between our approach and G-ENT, which extends the best traditional active learning method, ENT, to sample data only from the worst group, similar to JIN. The results in Tab. 6 show that G-ENT performs worse than our method and even worse than the original ENT in terms of fairness and average accuracy. This finding suggests that traditional active learning from the worst group alone may lead to suboptimal results, potentially due to sampling bias. Overall, the above literature discussion and further experiments justify our use of the validation set, the identification of the worst group, and the superiority of our JIN method over prior active learning baselines.

	STD. Acc. (%)		Rob. Acc. (%)	
	Worst (↑)	Avg (↑)	Worst (↑)	Avg (↑)
ENT	74.10±0.79	75.33±0.56	56.94±0.64	**60.25**±**0.56**
G-ENT	68.14±0.62	70.56±0.44	54.89±0.32	58.85±0.37
JIN	**75.07**±**0.53**	**75.74**±**0.49**	**57.39**±**0.10**	60.10±0.25

Table 6. **A fair comparison in sampling only from the worst group on the UTKFace dataset**. G-ENT not only performs worse than our method but even degrades a lot from the original ENT.

Generalization on various model architectures. We have tested various active data selection methods on ResNet18 [17] under the same training protocol to verify the generalization ability. As indicated in Tab. 7, our proposed approach outperforms all the representative baselines in both fair-performance and fair-robustness. Moreover, identical to Tab. 1, our method can select more adversarial sensitive samples from the worst group than FairAL, resulting in remarkably better fair-robustness. Overall, our results indicate that the JIN method is effective and applicable to various neural network backbones.

	STD. Acc. (%)		Rob. Acc. (%)	
	Worst (↑)	Avg (↑)	Worst (↑)	Avg (↑)
Init. AT	64.80±1.79	67.46±1.39	51.48±0.41	56.42±0.23
RAND	70.86±1.46	72.83±1.01	55.40±1.36	59.52±0.81
ENT	73.30±1.07	74.67±0.93	56.03±0.80	60.30±0.40
G-RAND	72.71±0.78	73.47±0.54	56.69±0.67	59.71±0.36
FairAL	74.28±0.60	75.41±0.35	56.80±0.46	**60.68**±**0.35**
JIN	**75.38**±**0.66**	**75.58**±**0.61**	**57.75**±**0.69**	60.42±0.25

Table 7. **Results on the UTKFace dataset using the ResNet18 model.** The experimental results in Tab. 1 and this table confirm that our method achieves advantages over multiple active selection baselines across model architectures.

5. Conclusion

We present FRAL, a novel learning framework designed to mitigate biases in safety-critical applications with limited labeling resources. Recognizing that previous data selection strategies have struggled with data imbalances and computational burdens within this framework, we introduce a elegant, effective, and efficient approach called the joint inconsistency (JIN) method to tackle these challenges. Through these advancements, we anticipate a new era of machine learning research for trustworthy visual applications.

Acknowledgement

This work was supported in part by the National Science and Technology Council, under Grant NSTC 111-2634-F-002-022 and MOST 110-2222-E-002-014-MY3, as well as Mobile Drive Technology Co., Ltd (MobileDrive). We are grateful to the National Center for High-performance Computing. We also thank Hsin-Ying Lee and Jhih-Ciang Wu for the kind suggestions on figures and paper writing.

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. 1, 3, 4, 5, 6, 8

[2] Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. Fair active learning. *Expert Systems with Applications*, 199:116981, 2022. 1, 2, 3, 4, 5, 6

[3] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2020. 1, 3, 6

[4] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018. 2

[5] Frédéric Branchaud-Charron, Parmida Atighehchian, Pau Rodríguez, Grace Abuhamad, and Alexandre Lacoste. Can active learning preemptively mitigate fairness issues? *arXiv preprint arXiv:2104.06879*, 2021. 3, 6

[6] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 8

[7] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21(2):277–292, 2010. 2

[8] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2021. 8

[9] 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, 5

[10] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018. 2, 5

[11] 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. 1, 3

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012. 2

[13] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 1, 3

[14] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192, 2017. 1, 3

[15] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, et al. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2

[16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, et al. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8

[18] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 2

[19] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, pages 7026–7037, 2019. 1, 3

[20] Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science*, 65(7):2966–2981, 2019. 2

[21] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018. 2

[22] 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. 2

[23] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR, 2020. 2

[24] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR, 2018. 2

[25] 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. 2, 8

[26] Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil. Taking advantage of multitask learning for fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 227–237, 2019. 2

[27] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased

performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019. 2

[28] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020. 8

[29] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 5

[30] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016. 2

[31] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 1, 3, 6

[32] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

[33] Amr Sharaf, Hal Daume III, and Renkun Ni. Promoting fairness in learned models by learning to active learn under parity constraints. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2149–2156, 2022. 1, 2, 3, 4, 5, 6

[34] 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, 4, 5, 6, 8

[35] Luke Stark. Facial recognition is the plutonium of ai. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3):50–55, 2019. 1

[36] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 2, 5

[37] 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, 4

[38] Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382. PMLR, 2019. 2

[39] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018. 5

[40] D. Wang and Y. Shang. A new active labeling method for deep learning. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 112–119, 2014. 1, 3, 5

[41] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016. 1, 3

[42] Ran Wang, Sam Kwong, and Degang Chen. Inconsistency-based active learning for support vector machines. *Pattern Recognition*, 45(10):3751–3767, 2012. 1, 4

[43] Wentao Wang, Han Xu, Xiaorui Liu, Yaxin Li, Bhavani Thuraisingham, and Jiliang Tang. Imbalanced adversarial training with reweighting. *arXiv preprint arXiv:2107.13639*, 2021. 1, 4, 5

[44] 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, 2, 4, 5, 8

[45] Weiping Yu, Sijie Zhu, Taojiannan Yang, and Chen Chen. Consistency-based active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3951–3960, 2022. 1, 4

[46] 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, 2, 5

[47] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, 2021. 2

[48] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017. 2, 5