Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/73/20240392

In-depth exploration and potential improvements on learning fair classifiers with partially annotated datasets

Haoyu Han[1,*,#], Tsz Tung Wong[2,6,#], Senhei He[3,7,†], Muyin Cai[4,8,†], Shile Lei[5,9]

[1]School of Mechanical Engineering & Automation, Beihang University, Beijing, 100191, China
[2]Basis international School Guangzhou(G12), Guangzhou, 510663, China
[3]Quality School International of Shekou, Shenzhen, 518000, China
[4]Wuhan Britain-China School,Wuhan, 430034, China
[5]Orange County of America High School, Beijing,100000, China

*Corresponding author's e-mail: haoyuhan_howie@buaa.edu.cn
[6]natalie.wong14584-bigz@basischina.com
[7]2450366809@qq.com
[8]Caimuyin0109@163.com
[9]leishile21@163.com
[#]These authors are co-first authors
[†]These authors are co-second authors

Abstract. In recent years, fairness-aware learning has been increasingly investigated. Researchers are trying to train accurate but fair classifiers. Yet, most existing methods rely on a fully- annotated dataset, which is an unrealistic assumption, since majority of the sensitive attributes of data remained unlabelled. This paper thoroughly explores this problem, namely Fairness - Aware Learning on Partially Labeled Datasets (FAL-PL) and Confidence-based Group Label Assignment (CGL), which is an innovative attempt to address FAL-PL. We conduct experiments by altering the hyperparameter, epoch, and the parameter, group-label ratio of CGL and discover that this method's results are easily affected by slight changes in the epoch and group-label ratio. Such unstableness reveals CGL's lack of robustness. We propose 2 modifications to further enhance CGL – **1. Co- teaching Method for Classifier Training**: We use the co-teaching method, which employs two models for training. We create these models by tweaking parameters and epochs in the original CGL model. After training, we choose the better-performing classifier based on accuracy. **2. Reducing Impact of False Pseudo Labels**: We've noticed an issue with the CGL method – random false label assignments can lead to errors. When two outcomes have similar probabilities, CGL might assign the wrong label. To address this, we propose a new parameter, $\boldsymbol{w}$, based on Gini impurity. It measures similarity between probabilities and acts as a weight, minimizing the influence of unreliable labels during the training stage of final fair model $\boldsymbol{f}$.

Keywords: fairness-aware learning, sensitive attributes of data, co- teaching method, CGL.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/73/20240392

1. Introduction

As neural networks become more mature and developed, it becomes widely applied to real-world applications across various domains and has become pervasive in our daily lives [1,2]: facial recognition, object detection, video surveillance, social media content moderation and among others. It is inevitable that such a technology could make mistakes. However, certain social groups experience a larger proportion of these mistakes, e.g. females with darker skin tones have higher error rates that lead to more misidentification or false accusations [3]. The potential of these techniques to perpetuate biases and even exacerbate social discrimination is undoubtedly problematic and should be carefully addressed by making machine learning models more fair.

To enhance the fairness of current NNs, many existing methods implement various concepts and perspectives [4], but most rely on *group labels* and the assumption of the entire dataset being perfectly labeled. Such an assumption is unrealistic [5-8] because, in a dataset e.g. a collection of images of human faces, sensitive attributes like gender and ethnicity remain unlabeled or hidden for a large portion of the data due to privacy and regulatory considerations. Labelling all of them requires an unimaginably large amount of human labour and effort, making it too costly and resource-consuming to be a suitable solution. How to provide unbiased yet accurate labels for unlabelled data continues to be an ongoing question that *fairness-aware machine learning* constantly investigates [9].

Contribution. We name the overall problem: ***F**airness - **A**ware **L**earning on **P**artially **L**abeled Datasets (**FAL-PL**)*. Through a more nuanced and refined approach, we reproduced and thoroughly explored the original CGL method. In pursuit of better addressing *FAL-PL*, our paper's contribution can be concluded to 3 points:

1. Manipulation of Variables: We changed the amount of epochs that the group classifier g and fair model f is trained for. Rather than solely using the original quantity, 50 epochs, we recorded the results for 5, 10, 20 and 50 epochs. From this alteration, we identify that even the simplest hyperparameter (epoch) has a significant impact on both accuracy and fairness.
2. Flaws in CGL: We discovered that the epoch's impact is also deeply intertwined with the parameters within CGL (group-label ratio). This underscores the considerable sensitivity of CGL itself, revealing its lack of robustness and susceptibility to alteration in response to method adjustments.
3. Potential Modifications for Improvement: We aspire to make CGL more robust, and be able to perform well across a range of scenarios. Additionally, we aim for a weaker relationship between the selection of hyperparameters and the method's own settings. Based on above intuitions, we provide 2 adjustments that can be applied on CGL to strive for enhanced overall improvement. CGL randomly assigns labels for a portion of the data, leaving room for possible mistakes. Incorrect labels are then passed on, impacting the training of the final fair model f. We propose: **1) assign a weight** to each label according to its confidence when passed into the loss function of model f and **2) implement co-teaching**, and train group classifier $g1$ and $g2$ simultaneously, to create a more mature final model.

2. Related Work

Group Fairness Fair-Training. Various approaches for addressing Fairness-training have been developed over the years. The main focus of this paper is on group fairness which stresses the fairness of a model. Many of these methods fall under three categories: Pre-processing [10-13] in- processing, and post-processing. The difference between the three is the timing at which the fairness technique is injected. These methods however all lack considerations for when training datasets are not fully labeled. CGL creates an auxiliary group classifier that assigns random labels to replace predictions with confidence below a threshold value $\boldsymbol{\tau}$ while keeping the pseudo labels with higher confidence than the threshold. This approach is able to achieve satisfactory performance in both accuracy and fairness and enhances results when applied on top of existing fair-training methods.

Fairness with Noisy Labels. There has not been many studies on dealing with fairness in the case of imperfect sensitive attributes, e.g. noisy group labels. Issues that occur with the current work are that they either assess the difference in the case of proxy variables (e.g. surname) [14,15] or they purely focus on noisy group labels [16,17] which cannot be extended into the Fair-PG setting this paper is

DOI: 10.54254/2755-2721/73/20240392

based on. There have been works that involve fair training with no information on protected attributes. Such works include a distributionally robust optimization (DRO) based approach [18],using an adversary to locate high-loss regions and re-weight them[19]. At the same time, there also exist de-biasing methods which attempt to solve the bias problem with no labels denoting bias. CGL has a key advantage over the brought-up methods which is CGL considers scenarios with partially annotated group labels.

Semi-Supervised Learning. The core of semi-supervised learning (SSL) is to learn a model with a few labeled samples and many unlabelled samples. The aim is to use the partially annotated training set to predict future attribute labels as accurately as possible, it is unclear if the predicted label can be directly applied to achieve group fairness in the test set. Additionally, the recent SSL methods are unable to be directly plugged in into the fairness problem as they seek augmentation methods instead of the pseudo- labeling strategy CGL takes. On top of that, CGL is similar to another approach UPS. This strategy withdraws pseudo-labels for those with low confidence labeled as opposed to replacing them with random labels. Throughout a series of experiments performed [CGL] it has been proven that by providing random labels, CGL outperforms the withdrawing label strategy UPS.

3. Method

This section formally defines the overall problem that this paper attempts to address and the fairness metric used to quantify and measure the results. Also, it explains the CGL method which this paper is built upon.

3.1. Problem Formulation & Fairness Metric

Under the FAL-PL scenario, there is an input feature $\boldsymbol{x}$ with $\boldsymbol{n}$ images, where $x_k \in \boldsymbol{x} \subset \mathbb{R}^d$. It is divided into a group-labelled set $D_l = \{X_k, S_k\}_{k=1}^{n_l}$ with representing the number of images and $\boldsymbol{S}_k \in \{0, 1, \ldots, n_s - 1\}$ representing sensitive attributes, e.g. gender, skin tone. There are target labels $\{\boldsymbol{y}_k\}$, with $\boldsymbol{y}_k \in \{0, 1, \ldots, n_y - 1\}$. In this paper, we generalize the sensitive attributes and target labels into binary variables, making $\boldsymbol{n}_s = \boldsymbol{n}_y = \boldsymbol{2}$. Due to the existence of unlabelled data, our goal is to train a group classifier g that precisely assigns pseudo labels and finally, to create a fair classifier $\boldsymbol{f}$, that can accurately predict target label $\boldsymbol{y}$ according to an input $\boldsymbol{x}$ while remaining unbiased against any sensitive attribute $\boldsymbol{S}$.

Following the CGL method [20], our method takes on the same fairness metric, **Equal**ity of **Opp**ortunity (EqualOpp) [21]. This metric centers around ensuring that individuals qualifying for an outcome have an equal chance of being correctly classified for that outcome. For example, when predicting whether a person committed a crime among a group of criminals, the result should remain unaffected by sensitive attributes like gender and ethnicity. EqualOpp is met when $\forall S_0, S_1 \in S, y \in Y, P(Y = y|S = S_0, Y = y) = P(Y = y|S = S_1, Y = y)$, meaning that the probability of an input feature being classified as y is the same, regardless of sensitive attribute being S_0 or S_1. The level of unfairness can be measured by a slight modification of the fairness metric, which we denote as **D**isparity of **E**qual**O**pp (DEO).

$$\Delta(f, P, y) = \max_{S_0, S_1} |P[f(x) = y|S = S_0, Y = y] - P[f(x) = y|S = S_1, Y = y]|$$

By taking the maximum (ΔM) and average (ΔA) of this value across different attributes, we develop a more robust measurement of overall fairness by explicitly demonstrating the worst, most unfair scenario and the average level of unfairness.

$$\Delta M(f, P) = \max_{y} \Delta(f, P, y), \Delta A(f, P) = \frac{1}{|Y|} \sum_{y \in Y} \Delta(f, P, y)$$

DOI: 10.54254/2755-2721/73/20240392

3.2. Overall Process

Our proposed method uses the CGL method as a foundation. According to the CGL method, a dataset is first separated in to 2 portions, one labeled group and one unlabelled group. Within the labeled group, it is partitioned into a training set (**labeled dataset 1**) and a validation set (**labeled dataset 2**). A group classifier $\boldsymbol{g}$ is trained on these two sets of labeled data. When dealing with the pseudo labels that classifier $\boldsymbol{g}$ generates, the method makes uses of a hyperparameter $\boldsymbol{\tau}$. $\boldsymbol{\tau}$. determines whether a pseudo label is kept or replaced by a random label: if the pseudo label's confidence level is larger than $\boldsymbol{\tau}$. it is kept; otherwise a label is randomly assigned according to probability. In the end, all unlabelled data becomes labeled, and the fair model $\boldsymbol{f}$. can be trained on this fully labeled dataset with another base fair-training method.

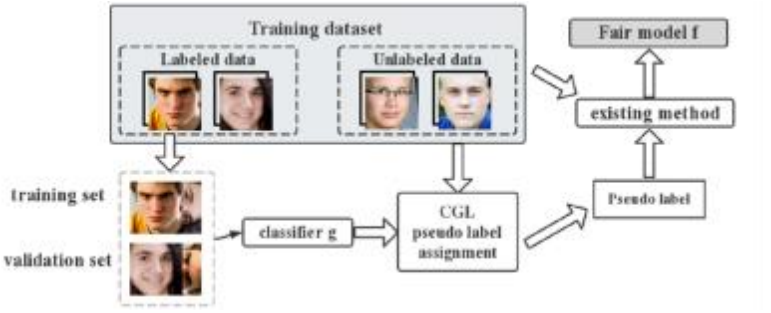


Figure 1. Overview of the CGL method.

After successfully reproducing the CGL method, we manipulate the amount of epochs executed. Concurrently, we vary the group label ratios to observe the optimal results and compare with the original CGL results. This reveals CGL's internal potential for better overall performance.

4. Experiment

In this section, we first specify our experimental settings and present the results. We next systematically analyse the correlation among performance (accuracy and fairness), hyper-parameters, and parameters in CGL to support our claim on the necessity of robustness improvement.

4.1. Experimental Setting

4.1.1. Dataset

CelebA. A dataset containing over 200,000 celebrity face images. Each is annotated with various attributes such as gender, age, presence of glasses, facial expression, and more. These annotations make the dataset valuable for tasks like facial recognition, attribute recognition, and generative modelling.

UTKFace. A dataset that contains over 20,000 face images, each annotated with labels for age, gender, and ethnicity. The dataset covers a wide range of ages, genders, and ethnicities, making it valuable for training and evaluating models that can predict these attributes from facial images.

UCI Adult. A tabular dataset where information is organized into rows and columns: each row corresponds to a specific observation or record; each column represents a different attribute or feature of that observation. It contains information about individuals from the U.S. Census, including attributes such as age, education, marital status, occupation, and income level.

In this paper, we focus on the **UCI Adult** dataset since tabular data is less explored in the CGL paper and more efficient to be analyzed. Notice that CelebA and UTKFace is more complicated than UCI Adult, and we believe the issue we discover in UCI Adult should be more severe on them.

DOI: 10.54254/2755-2721/73/20240392

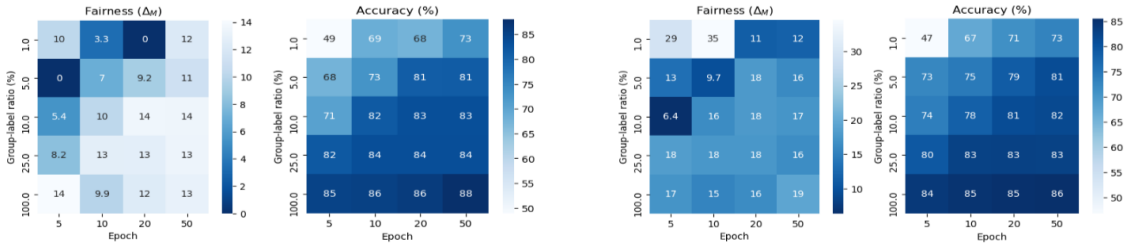
4.1.2. Base Fair-Training Methods

We employ 2 state-of-the-art in-processing methods: *MMD-based Fair Distillation* (MFD) and *FairHSIC* to train the final fair model $\boldsymbol{f}$ .Both methods use additional fairness-aware regularization terms as the relaxed version of the targeted fairness criteria.

4.2. Implementation Details

Our experiments were conducted on the UIC Adult dataset, with the baseline methods MFD and FairHSIC. Similar to the CGL experiment, we trained the group classifier $\boldsymbol{g}$ on different levels of initial group-label ratio (1%, 5%, 10%, 25%, 100%), yielding different accuracies and fairness. For each group-label ratio, we also altered the amount of epochs executed (5, 10, 20, 50), to examine how $\boldsymbol{g}$ performs differently when trained for a different amount of cycles. For each scenario, we record and plot out the fairness and accuracy, accordingly.

4.3. FINAL RESULTS



(a)FairHSIC Results (b)MFD Results

Figure 2. Results on UIC Adult. The accuracy (%) fairness (M) of each scenario of two CGL-based methods under varying group ratio levels (sv) and different training epochs. Darker colour means better results. Under different sv, the performance fluctuates differently under varying training epoch, indicating the strong correlation between sv and epoch.

4.3.1. Study On The Results

Fairness. A discernible pattern illustrating the connection between fairness and the number of epochs employed is not evident, primarily due to the presence of outlier data and contradictory trends that emerge as the epoch count escalates. Sometimes a downward trend is presented, while for the other times, and more commonly, an slight upward trend is shown. Similarly, no clear trend is exemplified between the amount of epochs and the group-label ratio. Interestingly, the results at epoch = 20 seem to outperform the ones when epoch is set to be 50, underscoring CGL's capability of achieving better fairness when it is trained for a different amount of epochs.

Accuracy. As the amount of epochs ran increase, there tends to be an upward trend in accuracy regardless of group-label ratio and method. Notably, this effect is accentuated when the group-label ratio is lower, resulting in a more pronounced accuracy boost. Overall, the highest accuracy is achieved when the group-label ratio is highest and epoch amount largest.

Correlation. In both cases, a slight modification in the hyper-parameter, epoch, or the parameter, group-label ratio results in rapid changes in the accuracy and fairness. This implies that CGL is actually more sensitive than their paper suggests. In fact, we notice a strong correlation between the effects of hyper-parameters (e.g., epochs) and parameters in CGL (e.g., group ratio levels) on the accuracy and fairness of the trained models, which impedes the direct application of CGL on downstream tasks. To make the mode more easily applied to different problems, it is necessary to enhance its robustness through a better design of the method.

DOI: 10.54254/2755-2721/73/20240392

5. Future Work

We take the existing CGL approach and modify the variables. After manipulating the number of epochs executed and the group labeled ratio, we observe optimal results when compared to the original CGL method. We think CGL has the potential for even better performance. We suggest introducing parameter $\boldsymbol{w}$ as a weight to the pseudo label and integrating it into the outermost part of the loss function. Co-teaching may also be a useful addition to training the model for a better result.

5.1. Overall Process

We implement co-teaching to collaboratively train group classifier $\boldsymbol{g1}$ and $\boldsymbol{g2}$, in order to result in a more thoroughly developed and mature classifier $\boldsymbol{g}$ that can provide labels more fairly and precisely. Also, we remove the initial threshold $\boldsymbol{\tau}$ that determines whether a pseudo label is remained or replaced by a random label; instead we use a parameter $\boldsymbol{w}$ to assigned to each label prior inputting them into the training of final fair model $\boldsymbol{w}$.

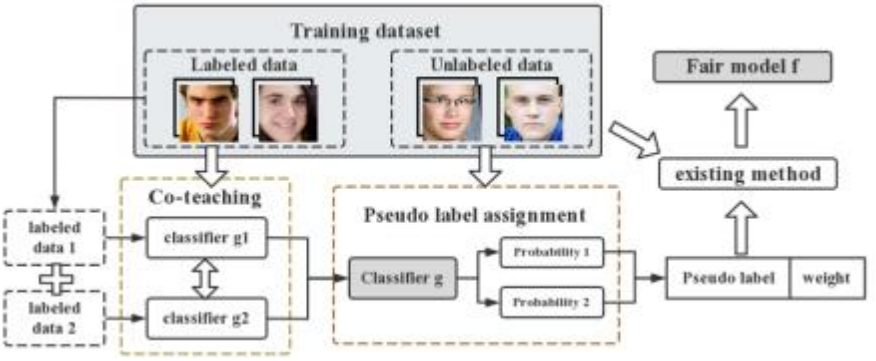


Figure 3. Overview of our proposed method.

5.1.1. Confidence -Weighted Labels

CGL's reliance on randomly assigned labels is rather risky, incorrect labels continue to have negative impacts when inputted into the final training stage of fair model $\boldsymbol{f}$. Our approach is to give up on assigning random labels when the confidence level of a pseudo label is below the threshold and to construct a new parameter w by the two probability values returned from the classifier and normalize it so that its value is between 0 and 1.

The calculation formula for parameter $\boldsymbol{w}$ is: $w = 1 - 2 * [2 * p1 * p2]$, where "p1" represents the probability of one prediction and " p2" represents the probability of the other ($p1 + p2 = 1$). It should be noted that the parameter "w" signifies the distance between the two probabilities. When "w" is closer to 1, the difference between the two probabilities is greater (for instance, one being 0.9 and the other being 0.1). This indicates that the classifier has higher confidence in the prediction with the larger probability. Conversely, when $\boldsymbol{w}$ approaches 0 (e.g., 0.5 and 0.5), the two probabilities are closer together, showing how the classifier is unable to differentiate effectively between the two predictions.

This mathematical method we chose is called the **G**ini **I**mpurity (GI). It must be noted that GI is a very "sensitive" parameter. In our solution, a probability pair of (0.999, 0.001) corresponds to a GI of about 0.996, while a probability pair of (0.9, 0.1) corresponds to a GI that quickly drops to 0.64. Such a characteristic of GI helps achieve our goal of accurately assigning false labels since the reduced confidence in the results predicted by the classifier is drastically amplified on the parameter $\boldsymbol{w}$.

During the training of the subsequent fair model $\boldsymbol{f}$, this weight can be multiplied in the outermost layer of its loss function. Consequently, predictions with lower confidence (indicating challenges in distinguishing original images due to unclear features or diminished image quality) will exert a minor influence on the equitable model's training.

DOI: 10.54254/2755-2721/73/20240392

5.1.2. Co-Teaching

Co-teaching [22] is a deep learning paradigm that was specially developed to combat noisy labels. The core essence of the method involves training two deep neural networks at the same time and allowing them to teach each other. In each batch of data, the network chooses the low-loss sample to be useful knowledge. Subsequently, the two networks will communicate with one another and backpropagate the data chosen by the other network then update itself. We believe the same can be done for CGL. By replacing g and g' as the two models we can train model g into a better version of itself. As a result, when the new dataset provided by the resultant model g is inputted into fair classifier f, we perceive a fairer and more accurate result.

6. Conclusion

With use of a confidence threshold, CGL effectively and accurately assigns pseudo labels to the unlabeled portion of a dataset, so later a classifier can be trained on this fully-labeled dataset, nurturing outcomes with higher quality. Yet, through a series of comprehensive experiments, we investigate the impact of varying hyperparameters, epochs, and the critical parameter, group-label ratio, on the performance of CGL. Our findings reveal that CGL's results are susceptible to slight changes in these parameters, indicating a lack of robustness in the method. It becomes evident that ensuring the stability of fairness-aware learning in partially labeled datasets is a complex challenge.

To enhance the robustness of CGL, we propose two significant modifications.

First, we introduce the co-teaching method for classifier training, employing two models with tweaked parameters and epochs. Following training, we select the superior classifier based on accuracy, thus enhancing the reliability and stability of the learning process.

Second, our modification addresses a notable issue in the CGL method – the potential impact of random false label assignments, which can lead to errors. When two outcomes have similar probabilities, CGL may assign the incorrect label. To mitigate this issue, we propose a novel parameter, denoted as w which relies on Gini impurity to gauge the similarity between probabilities. This parameter serves as a weighting factor, diminishing the influence of unreliable labels during the training phase, ultimately contributing to the robustness and accuracy of the final fair model f.

Ultimately, the goal of this research is to train fair and accurate classifiers, contributing to the advancement of social equity and inclusiveness. By addressing the challenges associated with partially labeled datasets and improving the robustness of fairness-aware learning methods, we hope to make meaningful strides towards creating more equitable and inclusive machine learning models.

Acknowledgement

Haoyu Han and Tsz Tung Wong contributed equally to this work and should be considered co-first authors.

Senhei He and Muyin Cai contributed equally to this work and should be considered co-second authors.

References

[1] Aditya K. Menon Alex Lamy, Ziyuan Zhong and Nakul Verma. Noise-tolerant fair classification. In Adv. Neural Inform. Process. Syst., volume 32, 2019.

[2] Neha Sharma, Reecha Sharma, Neeru Jindal, Machine Learning and Deep Learning Applications-A Vision, Global Transitions Proceedings, Volume 2, Issue 1, 2021.

[3] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conf. Fairness, Accountability and Transparency, pages 77–91. PMLR, 2018.

[4] Kehrenberg Thomas, Chen Zexun, Quadrianto Novi, "Tuning Fairness by Balancing Target Labels", Frontiers in Artificial Intelligence, Volume 3, 2020.

[5] Neha Sharma, Reecha Sharma, Neeru Jindal, Machine Learning and Deep Learning Applications-A Vision, Global Transitions Proceedings, Volume 2, Issue 1, 2021.

DOI: 10.54254/2755-2721/73/20240392

[6] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In IEEE Conf. Comput. Vis. Pattern Recog., pages 12115– 12124, 2021.
[7] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5810–5818, 2017.
[8] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Int. Conf. Comput. Vis., pages 3730–3738, 2015.
[9] T. Kamishima, S. Akaho and J. Sakuma, "Fairness-aware Learning through Regularization Approach," 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, 2011.
[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Conf. Innov. Theor. Compu. Scien., 2012.
[11] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In Int. Conf. Learn. Represent., 2017.
[12] Robert Geirhos, Jorn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11):665–673, 2020.
[13] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Int. Conf. Mach. Learn., pages 1050– 1059. PMLR, 2016.
[14] Clare Garvie. The perpetual line-up: Unregulated police face recognition in America. Georgetown Law, Center on Privacy & Technology, 2016.
[15] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conf. Fairness, Accountability and Transparency, pages 77–91. PMLR, 2018.
[16] Aditya K. Menon Alex Lamy, Ziyuan Zhong and Nakul Verma. Noise-tolerant fair classification. In Adv. Neural Inform. Process. Syst., volume 32, 2019.
[17] T. Kamishima, S. Akaho and J. Sakuma, "Fairness-aware Learning through Regularization Approach," 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, 2011.
[18] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In Int. Conf. Mach. Learn., 2019.
[19] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. Adv. Neural Inform. Process. Syst., 33:728–740, 2020.
[20] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In IEEE Conf. Comput. Vis. Pattern Recog., pages 12115– 12124, 2021.
[21] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Conf. Innov. Theor. Compu. Scien., 2012.
[22] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. Adv. Neural Inform. Process. Syst., 34, 2021.