ACHIEVING FAIR SPEECH EMOTION RECOGNITION VIA PERCEPTUAL FAIRNESS

Woan-Shiuan Chien, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan

ABSTRACT

Speech emotion recognition (SER) is a key technological module to be integrated into many voice-based solutions. One of the unique fairness issues in SER is caused by the inherently biased emotion perception given by the raters as ground truth labels. Mitigating rater biases are at core for SER to move toward optimizing both recognition and fairness performance. In this work, we proposed a *two-stage* framework, which produces debiased representations by using a fairness constraint adversarial framework in the first stage. Then, users are endued with the right to toggle between specified gender-wise perceptions on-demand after the gender-wise perceptual learning in the second stage. We further evaluate our results on two important fairness metrics to show that the distributions and predictions across different gender are fair.

***Index Terms*—** speech emotion recognition, rater bias, fair representation, perceptual fairness

1. INTRODUCTION

Emotion AI is a fast-developing contemporary technology, and the inclusion of a speech emotion recognition (SER) module enables machines to intelligently interact with humans [1, 2]. Due to the natural subjectivity in emotion perception and idiosyncratic factors of emotion production, the modern SER systems built on data-driven machine learning methods suffer from the issue of unfairness (biasedness). Being such a fundamental and computable internal attribute of humans, which has a tremendous opportunity permeating different applications, the biases introduced in the SER system not only put the system in doubt but also potentially create undesirable effects for users. As a result, deploying fair SER systems has emerged as a critical issue in modern society.

A fair SER system is a complex technology that is multifaceted. Firstly, deploying a SER system can be abstracted as having two components: *technology enablers* and *service providers*. As a developer of algorithms, it enables the SER technology but may not necessarily be the one that provides services, and those who provide services need to consider use case scenario from the user perspective. For *technology enablers*, they are tasked with developing algorithms, which render their decisions fair from the biases [3, 4]. While considerable works have proposed several algorithms for fair models in other ML applications, not many works for fair SER. In the context of SER, most works concentrate on mitigating subject-wise biases. For example, Gorrostieta et al. [5] proposed a gender and age de-biased SER using an adversarial invariant strategy. Due to subjectivity in the emotion labeling that is used as ground truth [6, 7], perceptual fairness (rater biases) is a uniquely critical factor though it has received much less attention. Studies [8, 9] have stated that gender differs in their sensitivity to emotion. Swerts et al. [10] also pointed out that females experience emotion more intensively than males. In this work, we investigate fairness from the angle of gender difference when rating emotional speech.

Furthermore, most of the approaches for achieving fairness are based solely on the concept that the model should produce in-distinguishable (debiased) representations or outcomes. While these are promising first steps, we argue that to make a SER fair, the technology should make users "feel fair" as well. That is, from the angle of a *service provider*, it should provide transparent outcomes that make users know what is learned about a person based on SER and have the opportunity to toggle between results. In fact, Grill et al. [11] stated that people associate transparency with fairness believing that it will lead to "fairer" outcomes. Schoeffer et al. [12] also observed that different amounts of information significantly influence user's perceived fairness. Accordingly, we deem that a fair SER system not only contains de-biasing module but also satisfies transparency criterion for end users. In this work, users should also be able to toggle from perceptually biased-free SER to a male/female viewpoint.

To achieve the two goals in realizing fair SER mentioned above: debiasing representation and transparency of gender-wise rater perspective, we propose a *two-stage* learning model. The first stage (*technology enablers*) provides fair representation to ensure that the SER model is unbiased with respect to rater's gender. It is done by explicitly reducing the distributional distances between different rater's gender groups. The second stage (*service providers*) can further provide users the ability to toggle between multiple transparent outputs by taking the debiased representation to move toward a gender-specific space when learning to output an emotion label. We evaluate our fair SER system with two important fairness metrics: statistical parity and consistency score to examine the trade-off between fairness and recognition performance.

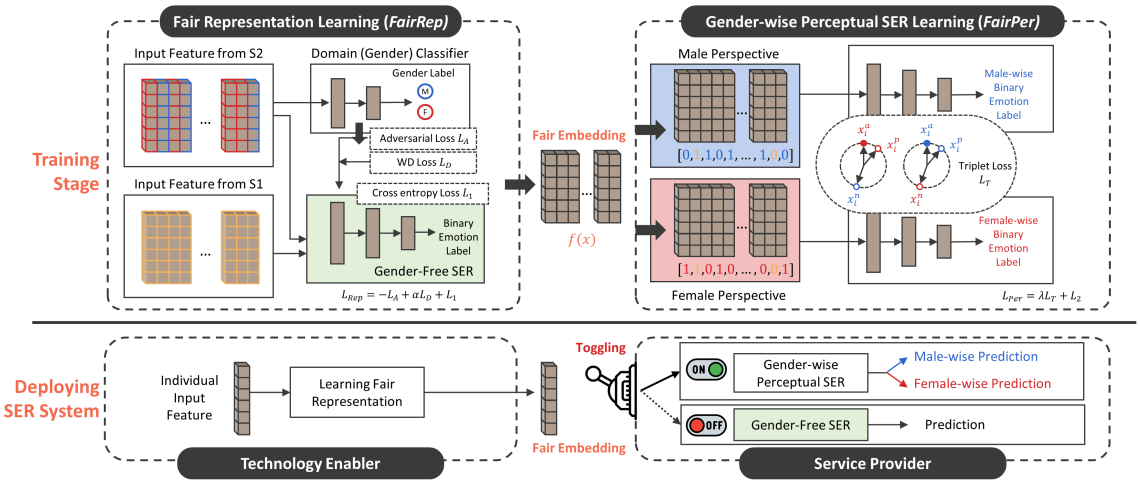

Fig. 1: Overview of the fair speech emotion recognition (SER) architecture using our proposed *two-stage* framework. The first stage is used to train for a fair representation, then utilize the fair representation to train for the gender-wise perceptual predictions through the second stage. For the corresponding application scenarios, once the technology enabler provides a fair embedding from the *FairRep*, the service provider provides users the ability to toggle options on-demand.

2. RESEARCH METHODOLOGY

2.1. Database

The IEMOCAP dataset [6] is a popular SER benchmark that contains five dyadic spoken interaction sessions in total and two actors (one male and one female) are included in each session. There are six unique raters (two males and four females) that provide emotion ratings. Each utterance is annotated by at least three raters on the primary emotions. The consensus labels are obtained with the plurality rule for primary emotions. Similar to most conventional SER studies, we focus our study on samples with emotional labels belonging to the five categorical emotion labels: Neutral, Happiness, Anger, Sadness and Frustration.

2.1.1. Rater's Perceptions

In this section, we show the unfair nature of labeling from gender-wise perceptions. The five major categorical emotion utterances contain 7362 utterances in total. First, we use all data to calculate the matching percentage between the male-/female- raters' consensus on each utterance and the voted ground truth. We further examine those samples where males and females raters have different emotion ratings (Non-Consensus Data). Table 1 shows these matching percentages between ground truth labels and different gender-rater splits. The number shows that there indeed exists a big perceptual difference between *Male* and *Female*. For instance, males disagree more with the ground truth label of Frustration compared to females, while they agree much more on all other emotions. In this work, we split our dataset into 2 subsets for further experiments: **S1** (the gender-wise perceptually-**unbiased** set), both males and females have identical emotion perceptions to the ground truth labels (a total of 3038 samples included). **S2** (the gender-wise perceptually-**biased** set), the ground truth labels have either the same emotion perceptions as males or females (a total of 4342 samples).

Table 1: Preliminary analyses on males and females consensus as compared to the voted ground truth.

Label	All Data			Non-Consensus Data		
Emotion	Samples	Male (%)	Female (%)	Samples	Male (%)	Female (%)
Overall	7362	80.66	**59.85**	4324	67.72	**32.28**
Neu.	1706	90.04	**34.82**	1323	87.30	**12.70**
Hap.	1633	91.73	**80.96**	446	69.73	**30.27**
Ang.	1099	90.81	**50.77**	628	85.03	**14.97**
Sad.	1080	85.83	**53.80**	628	77.55	**22.45**
Fru.	1844	**29.18**	75.70	1299	**33.95**	66.05

2.2. Computational Framework

In this work, we propose a *two-stage* framework to achieve fair SER. We extract the acoustic features as input, then derive perceptually-fair representation in the first stage. On the basis of fair embedding, we carry out on-demand emotion recognition that can toggle between the specified gender-specific viewpoints in the second stage (Fig. 1).

2.2.1. Acoustic Features

We extract 512-dimensional latent vq-wav2vec vectors as the acoustic features [15] using the fairseq tool [16] to embed information from raw audio. The pre-trained audio encoder could directly project the raw waveform into the latent space. All the features are speaker-wise z-normalized.

2.2.2. Stage 1: Fair Representation Learning

In the first stage, we propose *a fairness constraint adversarial framework* to learn the fair representations. Intuitively, our goal is the direct elimination of gender information, thus

Table 2: A summary of the experimental recognition results on *Fair Representation Learning* by F1 score (%) and the fairness performance by statistical parity score. The underlined datasets indicate the performance we are interested in. The bold numbers represent the corresponding best performance. The performance in 'All' dataset provides the "Gender-Free SER" prediction, which is mentioned in Fig.1.

Metrics	Recognition Performance															Statistical Parity Score (ideal=0)				
Emotion	Neutral			Happiness			Anger			Sadness			Frustration			Neu.	Hap.	Ang.	Sad.	Fru.
Dataset	All	S1	S2	All	S1	S2	All	S1	S2	All	S1	S2	All	S1	S2			S2		
DNN	77.73	**79.07**	66.12	70.00	**73.75**	62.73	76.44	**80.54**	73.28	82.28	**88.54**	78.68	63.54	67.61	61.30	0.650	0.428	0.389	0.169	0.626
LFR [13]	67.89	73.64	65.17	66.01	70.77	61.22	70.92	70.56	68.11	76.74	80.39	76.28	70.54	60.43	62.30	0.350	0.142	0.194	0.069	0.493
NRL [14]	68.55	75.12	66.84	69.46	72.10	65.65	70.76	71.40	69.81	78.74	82.01	77.98	64.32	65.02	64.88	0.332	0.134	0.197	**0.068**	0.452
FairRep_G	66.89	77.62	63.58	64.48	70.88	64.86	76.28	78.20	73.90	80.15	76.33	79.58	67.54	66.72	62.68	0.342	0.128	0.197	0.106	0.472
FairRep_F	68.82	76.46	65.46	66.46	71.66	63.84	78.26	79.00	76.69	78.26	79.00	77.44	69.02	66.61	63.68	0.350	0.136	0.208	0.104	0.467
FairRep	68.80	77.10	63.25	65.14	72.78	62.83	75.68	78.66	70.15	76.84	80.00	76.28	70.22	**66.82**	67.43	**0.350**	**0.126**	**0.189**	0.088	**0.448**

learning unbiased representation latent embedding, i.e., without gender-wise perspective. Specifically, we impose the constraint that the conditional distributions given the gender attributes are identical across the feature space. The process is similar to domain-invariant learning proposed by [17]. In this case, we train a domain (gender) classifier on the S2 (the biased set) with a cross-entropy loss. With the aim of mitigating the gender-wise perceptual biases in the representation, this loss L_A is subtracted from the emotion detection network.

We further impose fairness constraints on the distribution of instances in the feature space, such that fairness metrics are better met. Dwork et al. [**?**, 18] showed that statistical parity and consistency score, e.g., two important fairness metrics, can be jointly achieved if the Wasserstein Distance (WD) between two groups is minimized. Hence, by explicitly designing the WD as an optimization objective, we can ensure our classification results are fair when using the learned features. To calculate the loss L_D, we measure the distance D_W [14] between male's features and female's features. Hence, the parameters of this network are trained by minimizing the following loss function:

$$L_{\text{Rep}} = L_1 - L_{\text{A}} + \alpha L_{\text{D}}, \tag{1}$$

The proposed loss L_{Rep} comprises three parts: L_1 is the standard cross-entropy loss for evaluating the emotion classification performance on both S1 and S2; L_{A} is the gender information loss term, evaluating how much gender information is in the embedding; L_{D} measures how close the distributions over groups of the rater gender attribute are in the latent space. The hyperparameter α controls the trade-off of the system. The model in the first stage we defined as *FariRep* in the rest of the paper.

2.2.3. Stage 2: Gender-wise Perceptual SER Learning

In the second stage, our goal is to provide users the ability to toggle the recognition outcomes on-demand between gender-specified perception and gender-free perception. For gender-free SER, the predictions can be derived directly from the output in the first stage. While for gender-specified perceptual SER, we first extract the fair embedding from the last dense layer of *FairRep*. Every embedding will be assigned one label from male, one label from female and the corresponding gender index. Algorithmically, inspired from [19], we integrate triplet constraints to enforce that the feature space should be bounded according to the given gender index.

In a triplet-loss network, the inputs are a batch of triplet units $\langle x_i^a, x_i^p, x_i^n \rangle$ where x_i^a and x_i^p belong to the same rater-gender while x_i^a and x_i^n refer to different gender. Let $f(x)$ denotes the feature embedding of input x which embeds the representation into a d-dimensional Euclidean space. For a training triplet $\langle x_i^a, x_i^p, x_i^n \rangle$, the conditional triplet loss L_T being minimized is defined in [19]. As shown in Figure 1, we construct layers of fully-connected network as triplet-loss embedding layers before feeding into the deep neural network. The complete triplet-loss embedded network is optimized using the following total loss function,

$$L_{\text{Per}} = L_2 + \lambda L_{\text{T}}, \tag{2}$$

where L_2 is the standard cross entropy to the target gender-specified emotion label, and λ refers to the weighting between the two losses.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup

We design 2 experiments to evaluate the effectiveness of our proposed method. In order to further understand the trade-off of our proposed model between fairness and recognition performances, we present the recognition metrics along with two different fairness metrics chosen for each experiment. **Exp 1**: In the first stage, our target emotion label is voted ground truth, and the emotion recognition performance on S1 is evaluated by the weighted F1-score to demonstrate that the recognition performance on the consensus data is not affected by fairness operations. The fairness metric, statistical parity score (ideal value=0) [20], is evaluated on S2 dataset between different rater's gender and our predictions to demonstrate that our model has no preference for one category of the gender attribute over the other. **Exp 2**: In the second stage, it stands on a known "fair" ground while providing options of specific gender-wise perceptions on-demand. Thus all experiments are conducted by using the fair representation derived in *Stage 1*. the target emotion label is male raters' consensus or female raters' consensus, and we evaluate the gender-wise emotion recognition performance by weighted F1-score, i.e. using all male perspectives as the target label, and the

Table 3: A summary of the recognition results on *Gender-wise Perceptual SER Learning* by F1 (%) and the fairness performance by consistency score. 'M' and 'F' represents evaluating all data with male/female labels from male/female perspectives.

Metrics	Recognition Performance										Consistency Score (ideal=1)				
Emotion	Neutral		Happiness		Anger		Sadness		Frustration		Neu.	Hap.	Ang.	Sad.	Fru.
Perspective	M	F	M	F	M	F	M	F	M	F	Original - 0.579				
DNN	67.68	62.18	67.33	61.10	75.52	66.52	82.04	75.90	62.17	64.96	0.537	0.552	0.583	0.552	0.568
LFR [13]	65.23	63.44	64.00	60.71	69.22	68.15	77.06	72.41	61.17	65.02	0.685	0.698	0.722	0.735	0.706
NRL [14]	67.25	65.51	66.12	63.02	69.98	68.33	78.43	75.62	64.05	63.72	0.786	0.783	0.790	0.762	0.741
FairRep_G	66.32	64.04	67.57	62.22	77.33	70.16	73.40	75.62	63.00	60.77	0.791	0.772	0.801	0.766	0.743
FairRep_F	65.10	64.66	64.57	63.86	76.98	78.03	78.80	76.69	62.06	64.16	0.788	0.761	0.762	0.733	0.752
FairRep	68.26	75.82	66.73	68.15	78.36	75.54	77.97	77.25	75.88	77.98	0.802	0.777	0.809	0.739	0.760
FairPer	**80.83**	**86.75**	**73.64**	**75.89**	**80.46**	**88.30**	**79.16**	**80.96**	**85.16**	**77.42**	**0.853**	**0.827**	**0.822**	**0.797**	**0.840**

same for females. Since in the second stage, we define fairness as the same rater-gender samples should be close to each other (biased-free indicates there is no unwanted viewpoint in our prediction in this case), we use the fairness metric of consistency score (ideal=1) [13], which evaluates the consistency between the embedding and gender attribute within a k-nearest neighbor set ($k = 20$).

We use DNN as the simplest vanilla baseline, i.e., directly using all features to train a classifier with three dense layers. The compared fair representations methods include: LFR [13] directly assigned the instances to certain prototypes as latent representations; NRL [14] can be regarded as a SOTA framework for fair representation learning with fairness constraints. Our abated method, FairRep_F does not consider rater-gender domain knowledge, which is similar to the NRL framework, while FairRep_G does not consider fairness constraints. For all experiments, a session-independent cross-validation scheme is applied. we set the learning rate and decaying factor at 0.001, the drop out is set to 0.2. Hyperparameters λ and α are grid-searched among [0.01, 0.001]. Batch size is fixed as 32, the max epoch is 500, and the optimizer is Adam.

3.2. Experimental Results and Analyses

We examine the results of our *two-stage* framework for each stage (Exp 1: *Stage 1* & Exp 2: *Stage 2*). First, Table 2 summarizes the results on the first stage. Since we impose fairness constraints, our method is necessarily bound to drop slightly in terms of recognition performances overall as compared to methods without optimizing for fairness, i.e, DNN. However, the advantage of our *FairRep* is twofold: 1) it better satisfies statistical parity metrics than methods without consideration of fairness; 2) it suffers the least performance drop on the consensus dataset (S1). Along with the statistical parity score, several observations can be made. Without a fair mechanism, the SER model results in obviously unfair prediction (biased toward certain gender's perception of emotion). While in the ablated method of FairRep_G, gender differences are eliminated to some extent. Since it does not consider the fairness constraints in the learning, there may still be indirect bias inclusion [21] due to other attributes related to "gender". This is also evident when examining the statistical parity score. Furthermore, FairRep_F does not directly eliminate gender differences in domain knowledge resulting in retaining gender information. Consequently, from the two ablated results, both domain invariant and distance distribution matching are necessary and would affect the learned model performances, as measured in terms of both fairness metrics and discriminative power. We observe our proposed method, *FairRep*, only suffers a slight drop of 1.97% for Neutral, 0.97% for Happiness, 1.88% for Anger and 0.79% for Frustration in F1-Score as compared to DNN. However, it can achieve a competitive fairness performance on statistical parity score. This result depicts that *FairRep* leads to a more fair SER that can handle both the consensus data (S1) and gender-biased data (S2).

Table 3 summarizes the results of the second stage. Our proposed method, *FairPer*, offers more precise gender-wise perceptual predictions for different rater's gender. This is evident in the high recognition performances across all gender-wise emotion detection tasks. An interesting finding is that male versus female performance patterns reflect the differences in gender-wise perception observed in Table 1. Through our model, the emotion category of which there is more disagreement between male or female to the voted ground truth (shown in Table 1), the gender class would achieve higher performances. For example, males perceptions do not match well to the voted ground truth on Frustration label, yet our model can accommodate specifics to male perspectives, reaching 85.16% in F1 score. The higher consistency score in our model further reinforces that learned representation in the second stage would clustered gender-specified samples closer together providing a more consistent gender-specific viewpoint. Collectively, these findings provide evidence that our model conforms to gender perspectives, and could better provide more humane and "fair" experiences due to its flexibility to toggle between male, female, or gender-free recognition.

4. CONCLUSION

In this work, we propose a *two-stage* framework, which produces fair representations by using a fairness constraint adversarial framework in the first stage. Then, users are given abilities to toggle the specified gender-wise perceptions on-demand in the second stage. Our results reveal interesting insights: 1) our proposed model alleviates the drop in performance when validating on unbiased data; 2) the fairness metrics indicate that our model addresses the unfair issue. The immediate further work will explore and address the fairness issues in SER from an angle of joint subject-rater biases.

5. REFERENCES

[1] Simran Kaur and Richa Sharma, "Emotion ai: integrating emotional intelligence with artificial intelligence in the digital workplace," in *Innovations in Information and Communication Technologies (IICT-2020)*, pp. 337–343. Springer, 2021.

[2] Shrikanth Narayanan and Panayiotis G Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[3] Shikha Verma, "Weapons of math destruction: How big data increases inequality and threatens democracy," *Vikalpa*, vol. 44, no. 2, pp. 97–98, 2019.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine bias," in *Ethics of Data and Analytics*, pp. 254–264. Auerbach Publications, 2016.

[5] Cristina Gorrostieta, Reza Lotfian, Kye Taylor, Richard Brutti, and John Kane, "Gender de-biasing in speech emotion recognition.," in *INTERSPEECH*, 2019, pp. 2823–2827.

[6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[7] Soroosh Mariooryad, Reza Lotfian, and Carlos Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] Leslie R Brody, "Gender differences in emotional development: A review of theories and research," *Journal of personality*, vol. 53, no. 2, pp. 102–149, 1985.

[9] Jacob Miguel Vigil, "A socio-relational framework of sex differences in the expression of emotion," *Behavioral and Brain Sciences*, vol. 32, no. 5, pp. 375–390, 2009.

[10] Marc Swerts and Emiel Krahmer, "Gender-related differences in the production and perception of emotion," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[11] Gabriel Grill and Nazanin Andalibi, "Attitudes and folk theories of data subjects on transparency and accuracy in emotion recognition," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–35, 2022.

[12] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski, ""there is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making," New York, NY, USA, 2022, FAccT '22, p. 1616–1628, Association for Computing Machinery.

[13] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[14] Rui Feng, Yang Yang, Yuehan Lyu, Chenhao Tan, Yizhou Sun, and Chunping Wang, "Learning fair representations via an adversarial framework," *arXiv preprint arXiv:1904.13341*, 2019.

[15] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2019.

[16] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 48–53.

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[18] 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*, 2012, pp. 214–226.

[19] Andreas Veit, Serge Belongie, and Theofanis Karaletsos, "Conditional similarity networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 830–838.

[20] Ke Yang and Julia Stoyanovich, "Measuring fairness in ranked outputs," in *Proceedings of the 29th international conference on scientific and statistical database management*, 2017, pp. 1–6.

[21] Indre Zliobaite, "A survey on measuring indirect discrimination in machine learning," *arXiv preprint arXiv:1511.00148*, 2015.