Gender Bias in Depression Detection Using Audio Features

Andrew Bailey, Mark D. Plumbley

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, U.K.
{andrew.bailey, m.plumbley}@surrey.ac.uk

Abstract—Depression is a large-scale mental health problem and a challenging area for machine learning researchers in detection of depression. Datasets such as Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) have been created to aid research in this area. However, on top of the challenges inherent in accurately detecting depression, biases in datasets may result in skewed classification performance. In this paper we examine gender bias in the DAIC-WOZ dataset. We show that gender biases in DAIC-WOZ can lead to an overreporting of performance. By different concepts from Fair Machine Learning, such as data re-distribution, and using raw audio features, we can mitigate against the harmful effects of bias.

Index Terms—Depression Detection, Fair Machine Learning, Bias, Deep Learning

I. INTRODUCTION

Depression is a mental health disorder that causes severe symptoms for individuals. Typical emotional symptoms include lasting feelings of unhappiness, hopelessness, and a lack of interest or enjoyment. The speech of individuals with depression has been characterised as lifeless or flat, and delivered in a monotone fashion [1]. Depression is a common and costly mental health disorder. McManus et al. [2] reported in 2014 that 3.8% of adults in England suffered from depression and, according to the OECD [3], the cost of mental health in the UK was around 4% of GDP (€106bn) in 2015.

Machine learning techniques have successfully been applied to many health-related areas [4], [5] and therefore have the potential to improve depression detection. The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset [6] was designed to facilitate research into depression detection and was released as part of the 2016 Audio-Visual Emotion Challenge (AVEC) [7]. DAIC-WOZ contains interviews from 189 participants: 57 diagnosed with post-traumatic stress disorder or depression, and 132 who are not.

The ground truth labels in DAIC-WOZ were obtained through a depression diagnosis questionnaire, the Patient Health Questionnaire 8 (PHQ-8) [8], which consists of 8 questions. Every participant in the DAIC-WOZ answered the 8 questions by assigning an integer score in the range $0-3$. Once all the questions were completed, the scores were added together to give a depression rating, r out of 24, where $r < 10$ is classified as not depressed and $r \geq 10$ is classified as depressed.

Traditional machine learning algorithms such as Support Vector Machines (SVM) were used by the AVEC 2016 [7]

organisers as an audio baseline classification system. Since then, deep learning methods have also been applied to DAIC-WOZ including Convolutional Neural Networks (CNN) [9]–[11], Fully Connected Deep Neural Networks [12], and Recurrent Neural Networks (RNN) [10], [11], along with hand-crafted features, including the spectrogram [11] and the mel-spectrogram [10].

However, when machine learning is used to tackle health-related problems, such as depression detection, considerations must be taken to ensure that these models do not incorporate bias [13]. Bias can result in subjects with protected characteristics, such as race or gender, being unfairly penalised. This can occur due to many factors such as poor data collection standards [14] or the data processing in machine learning models [14].

Fair Machine Learning (Fair ML) [15] is an area of research that explores the idea of fair, unbiased classification, in an attempt to work towards a fairer society through the use of machine learning. Fair ML techniques have been used in healthcare domains such as the paralinguistics community to detect dementia [16].

We show in this paper that the DAIC-WOZ dataset contains gender bias and that this bias can negatively affect the resulting accuracy of machine learning models using audio features. We also find that deep learning models based on raw audio are more robust to gender bias than ones based on other common hand-crafted features, such as mel-spectrogram.

Our paper is organised as follows. Section II explores the background on the DAIC-WOZ dataset and Fair ML, Section III discusses the baseline model architecture, Section IV explores our proposed method, Section V explores the bias problem and how to mitigate against it and Section VI concludes the paper.

II. DAIC-WOZ, DEPRESSION DETECTION, AND FAIRNESS

A. DAIC-WOZ Dataset

To gather the data for the DAIC-WOZ dataset, interviews were conducted between each participant and a virtual interviewer ("Ellie") controlled by a researcher in another room [17]. The audio and facial features of the participants were recorded (Figure 1). The interviews range from 7 minutes to 35 minutes.

For Machine Learning purposes, the DAIC-WOZ dataset is divided into a training set of 107 files (76 non-depressed (ND) interviews and 31 depressed (D) interviews), a validation set
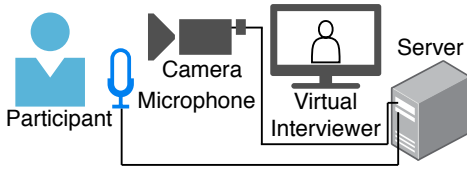
Fig. 1. A typical interview scenario where a participant converses with virtual interviewer on a TV screen. The camera records the facial movements and the microphone records the audio of the session.

TABLE I
SUBDIVISIONS OF TRAINING FILES (GENDER AND CLASS) OF DAIC-WOZ.

	ND	D	Total (ND+D)
Female (F)	27 (25%)	17 (16%)	44 (41%)
Male (M)	49 (46%)	14 (13%)	63 (59%)
Total (F+M)	76 (71%)	31 (29%)	107 (100%)

of 35 files (23 ND interviews and 12 D interviews), and a test set of 47 files. The distribution of data for the test set is hidden due to the AVEC competition [7].

To maintain the privacy of the participants, the raw visual data is not released. Instead, visual features based on the OpenFace framework [18] and the FACET toolbox [19] were extracted. Raw audio files, sampled at 16 kHz, are provided.

B. Gender Bias in DAIC-WOZ

As previous researchers have noted, we can see from Table I, that there is a high-level class imbalance of roughly 2:5 in terms of D:ND interviews in the training set.

In addition, we also see a gender bias in Table I that has not been explicitly acknowledged by previous researchers. While the ratio of D:ND females is roughly 5:8, the ratio of D:ND males in the dataset is 2:7. Thus the frequency of individuals with depression in the dataset differs between genders, such that $p(D \mid g = \text{f}) > p(D \mid g = \text{m})$ where g is gender, f is female and m is male. We also see a similar class imbalance as well as gender bias in the validation set in Table II.

C. Training on Separate Gender Subsets

Based on research indicating that depressive symptoms can differ depending on gender [20], some authors split DAIC-WOZ into two gendered subsets, one containing the female interviews and one containing the male interviews [12], [21], [22]. These authors [12], [21], [22] did not report awareness of the gender bias.

While these authors reported improved performance detecting depression on DAIC-WOZ compared to not splitting the dataset, deep learning benefits from large amounts of data, therefore training two separate models on subsets of the larger training data may may lead to reduced performance.

In addition, gender differences in speech signals are well known [23]. Huang et al. [24] proposed to reduce gender dependency in speech signals by applying z-normalisation to every audio signal based on gender according to, $\tilde{x} = (x-\mu)/\sigma$, where x is the input feature, μ and σ are the mean and standard

TABLE II
SUBDIVISIONS OF VALIDATION FILES (GENDER AND CLASS) OF
DAIC-WOZ. TEST SET DETAILS ARE NOT RELEASED FOR COMPETITION
PURPOSES.

	ND	D	Total (ND+D)
Female (F)	12 (34%)	7 (20%)	19 (54%)
Male (M)	11 (31%)	5 (14%)	16 (46%)
Total (F+M)	23 (65%)	12 (34%)	35 (100%)

deviation of the features in the DAIC-WOZ training set split by gender and $\tilde{x}$ is the normalised feature. Huang et al. [24] did not compare results before and after gender normalisation and did not comment on the larger gender bias in the dataset.

In this paper, we will discuss ways to deal with gender bias in the training set without creating two gendered subsets. To do this, let us explore some recent work on fair machine learning to resolve gender bias.

D. Fair Machine Learning

In Fair Machine Learning (Fair ML), we are interested in making sure that protected characteristics, such as race or gender, are treated equally. One Fair ML approach, *statistical parity* [25], is designed to ensure that an acceptance rate, such as admission to a University programme, is the same regardless of the protected characteristics of an individual, i.e.:

$$\mathbb{P}\left[R = 1 \mid A = a\right] = \mathbb{P}\left[R = 1 \mid A = b\right] \quad (1)$$

where R is the score and A is a protected characteristic.

Another principle that Fair ML follows, known as *sufficiency*, says a target label $Y = 1$ is independent of a protected characteristic, A, given some score $R = r$, i.e.:

$$\mathbb{P}\left[Y = 1 \mid R = r, \ A = a\right] = \mathbb{P}\left[Y = 1 \mid R = r, \ A = b\right] \quad (2)$$

An example of *sufficiency* would be falsely diagnosing individuals of cancer equally across protected characteristics (equal false positive rate).

Krawczyk [26] mentions several approaches for ensuring an unbiased classifier, including data generation techniques such as data sub-sampling, the process of equalising the number of examples from each class in the dataset by randomly selecting a portion of examples from the majority classes.

For more examples of fairness criteria and alternative definitions, additional work can be found at [27]–[29].

In this paper, we will determine the effects of gender bias in the DAIC-WOZ dataset on a deep learing model and apply data sub-sampling techniques from Krawczyk [26] to tackle this bias.

III. DEPAUDIONET

Our model will be based on "DepAudioNet" of Ma et al. [10], which was one of the best models in AVEC 2016 to utilise deep learning with audio-only processing for depression detection.
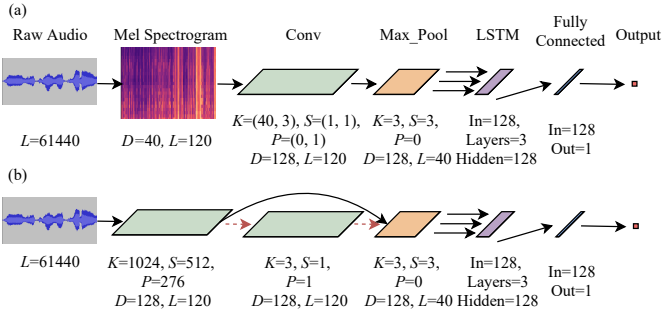
Fig. 2. Network architectures, D is depth (number of feature channels), L is length, K is the kernel size, S is the stride length, and P is the zero padding. (a) shows the DepAudioNet architecture (after [10]); (b) shows our raw audio model where the dotted red arrow shows an alternative network architecture with an additional convolutional filter.

A. File Processing

DepAudioNet [10] extracts mel-spectrogram features (mel filter bank with 40 frequency bins) using a Hanning window of length $w = 1024$, and hop size $h = 512$. Next, the features from each audio signal are normalised according to the z-normalisation ($\tilde{x} = (x - \mu)/\sigma$) where $\tilde{x}$ is the normalised feature, x is the feature from the input audio signal, μ and σ are the mean and standard deviation calculated based on the features from each individual audio signal.

Ma et al. [10] randomly crop every normalised feature to the same size as the shortest audio signal. To account for the DAIC-WOZ D:ND class imbalance (ND > D), they then randomly sub-sample the ND features to train with an equal number of D and ND examples. They then split the resulting feature into multiple temporal segments of size $N_{\text{seg}} = 120$ to be used as input for DepAudioNet.

B. Model Architecture

Figure 2(a) shows the overall DepAudioNet architecture. The source code for DepAudioNet is not available and so we reproduced the model through details given in the paper [10] using reasonable assumptions for unspecified aspects.

We experimented on the validation data to set the initial learning rate (LR) and rate of decay of the LR (λ) as no details in [10] were given. We initialised the LR to 0.001 and decayed the LR by a factor of 0.9 every λ_{epoch} equal to 2 or 3. We used the Adam optimiser, the F1-Score to measure the accuracy of the model, and the Binary Cross Entropy Loss to update the network weights.

C. DepAudioNet Reproduction Results

As confirmation of successful reproduction, our reproduction achieved comparable results to those reported in [10]. As we can see from Table III, the performance of our reproductions with $\lambda = 2$ and $\lambda = 3$ are slightly better than the performance reported by Ma et al. [10].

IV. PROPOSED RAW AUDIO MODEL

We hypothesise that features such as the mel spectrogram used in DepAudioNet may emphasise pitch or vocal tract

Model	λ	C	F1 (ND)	F1 (D)	F1 avg
DepAudioNet [10]	–	1	.700	.520	.610
DepAudioNet*	2	1	.732	.522	.627
DepAudioNet*	3	1	.740	**.539**	.634
Raw Audio	2	1	.707	.501	.603
Raw Audio	3	1	.707	.531	.619
Raw Audio	2	2	.766	.531	.648
Raw Audio	**3**	**2**	**.796**	.520	**.658**

information that may be a proxy for gender, which could result in utilising the gender bias of DAIC-WOZ. To test the robustness of the mel-spectrogram features against the gender bias of DAIC-WOZ, we compared our reproduced DepAudioNet with an alternative raw audio model, Figure 2(b).

Our proposed raw audio model follows the pre-processing outlined in Section III-A except that we use raw audio as our input instead of extracting a mel-spectrogram. The input size of the raw audio features is calculated according to $r = N_{\text{seg}} \times h$.

The kernel of the convolutional filter for our raw audio model matched the window and hop size from the mel spectrogram analysis from Section III-A, $\{K = 1024, \ S = 512, \ P = 276\}$, where K is the kernel size, S is the stride length, and P is the zero padding. These values were chosen so that the output dimensions of the convolutional filter would match those used by Ma et al. [10]. We also experimented with an additional convolutional filter (shown by red dotted arrow in Figure 2(b)) added after the first with parameters $\{K = 3, \ S = 1, \ P = 1\}$ (a 1D version of the convolutional filter in DepAudioNet shown in Figure 2(a)).

A. Results of Proposed Raw Audio Model

From Table III, our proposed raw audio model with a single convolutional filter performed comparably, within $\pm 1\%$, to the results reported by Ma et al. [10]. When we added an extra convolutional layer, our proposed raw audio model surpassed the performance of Ma et al. [10] and our reproduced DepAudioNet models.

In the next section we will evaluate the effect of the gender bias of the DAIC-WOZ and show how using raw audio as input adds robustness against this type of bias.

V. IMPACT OF GENDER BIAS

We saw in Table I (Section II-A) that the number of training examples varies according to class (D:ND) and gender (female/male). While Ma et al. [10] addressed the (D:ND) class imbalance by sub-sampling the ND examples, they did not report on gender bias.

Following the data re-distribution approach of Krawczyk [26] we performed data sub-sampling to have equal data in all quadrants of Table I. We split the training data into the four quadrants in Table I: (f, D), (f, ND), (m, D), (m, ND).

TABLE IV
RESULTS ON THE VALIDATION SET OF OUR REPRODUCED AND PROPOSED MODELS SHOWING THE EFFECT OF GENDER BALANCING BY UTILISING THE SUB-SAMPLING DATA RE-DISTRIBUTION METHOD FROM [26]. DEPAUDIONET* IS OUR REPRODUCED DEPAUDIONET MODEL BASED ON MA ET AL. [10], λ IS THE RATE OF DECAY OF THE LR IN EPOCHS AND C IS THE NUMBER OF CONVOLUTIONAL FILTERS IN THE MODEL.

Row	Model	λ	C	Gender Balance	Female			Male			F1 Total Average	Difference (%)
					F1 avg	F1 (ND)	F1 (D)	F1 avg	F1 (ND)	F1 (D)		
1	DepAudioNet*	2	1	N	.628	.637	.619	.539	.800	.279	.627	−14.0
2	DepAudioNet*	2	1	Y	.547	.696	.397	.497	.730	.264	.539	
3	DepAudioNet*	3	1	N	.665	.688	.641	.527	.777	.277	.634	−13.4
4	DepAudioNet*	3	1	Y	.573	.697	.449	.514	.703	.324	.549	
5	Raw Audio	2	1	N	.546	.555	.537	.618	.819	.417	.604	+0.82
6	Raw Audio	2	1	Y	.608	.747	.470	.606	.710	.502	.609	
7	Raw Audio	3	1	N	.526	.505	.547	.659	.846	.471	.619	+3.43
8	Raw Audio	3	1	Y	.613	.808	.418	.660	.766	.555	.641	
9	Raw Audio	2	2	N	.627	.672	.582	.614	.841	.387	.648	−4.78
10	Raw Audio	2	2	Y	.600	.782	.419	.635	.763	.507	.617	
11	Raw Audio	3	2	N	.679	.743	.615	.553	.840	.267	.658	−4.26
12	Raw Audio	3	2	Y	.586	.711	.462	.668	.737	.598	.630	

We then randomly sub-sampled the audio signals from every quadrant with respect to the quadrant with the fewest examples, in this case 14 audio signals from depressed males (m, D). Following this process, our training data equally represents every quadrant found in Table I and ensures independence [25] from gender, such that $p(D \mid g = \text{f}) = p(D \mid g = \text{m})$ from the perspective of the training data.

A. Results of Gender Balance

The results after training the models following gender balance are shown in Table IV. The performance of DepAudioNet* (our reproduction of Ma et al. [10]) (rows 1 and 3 respectively) drops by 14.0% and 13.4% respectively after we employ gender balance (rows 2 and 4 respectively) from .627 to .539 (row 1 to row 2), and .634 to .549 in (row 3 to row 4).

Our results using the proposed raw audio model vary. While using one convolutional filter (rows 5 and 7 respectively), we see a performance boost of 0.82% and 3.43% respectively (rows 6 and 8 respectively) after applying gender balance. But the addition of the second convolutional filter (rows 9 and 11 respectively) shows a drop in performance of 4.78% and 4.36% respectively after gender balance (rows 10 and 12 respectively), albeit less than the drop found using DepAudioNet*.

We can see (rows 1 to 2 and rows 3 to 4) that the (f, D) performance of DepAudioNet* drops more than our raw audio model (rows 5 to 6, 7 to 8, 9 to 10, and 11 to 12).

The performance drop of DepAudioNet* suggests that DepAudioNet* may use gender-based information. Finally, we can see that performance on the minority class (m, D) increases across all models (with the exception of rows 1 to 2). Our proposed raw audio model shows superior results to DepAudioNet* in terms of overall performance and performance difference after applying gender balancing. These results suggest that raw audio may be more robust to gender bias than using the mel-spectrogram for this task.

We recognise the limitations of this study as we have only tested the robustness of the mel-spectrogram and raw audio on one dataset and one model architecture. However, as a proof of concept our results suggest that further analysis would be beneficial to concretely determine the effect of gender bias on various datasets and input features.

VI. CONCLUSION

We have studied potential gender bias in the DAIC-WOZ dataset. In Section V we removed gender bias through data re-distribution by balancing the training data quadrants from Table I. Data re-distribution helps the classifier to learn in an unbiased manner by ensuring independence from gender with respect to the training data.

We showed in Table IV that leaving the dataset unaltered can result in overreporting performance due to the gender bias (shown by performance drops after applying gender balancing techniques), especially when using the mel-spectrogram.

We found that using raw audio can provide a classification that is more robust to gender bias than the mel-spectrogram as shown by the performance changes before and after gender balancing. Using raw audio can also outperform results using the mel-spectrogram despite the network architecture being optimised for the mel-spectrogram and not raw audio.

In future work, it would be interesting to explore other datasets which exhibit gender bias, and the robustness of different features against dataset bias. In addition, it would be useful to examine the robustness of other popular hand-crafted features such as the Mel-Frequency Cepstral Coefficients (MFCC). We would like to determine the efficacy of approaches such as gender normalisation on the gendered subsets of DAIC-WOZ [24] to tackle the gender bias of DAIC-WOZ as well.

Our code is available for download[1].

[1]https://github.com/adbailey1/DepAudioNet_reproduction

REFERENCES

[1] S. Scherer et al., "Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 59–73, January 2016.

[2] S. McManus et al., "Mental Health and Wellbeing in England: Adult Psychiatric Morbidity Survey 2014," *Leeds: NHS Digital*, 2016.

[3] OECD and European Union, *Health at a Glance: Europe 2018: State of Health in the EU Cycle*, OECD Publishing, November 2018.

[4] M. Sendak et al., "The human body is a black box: Supporting clinical decision-making with deep learning," in *Proc. FAccT*, Barcelona, January 2020, pp. 99–109.

[5] A. Peimankar and S. Puthusserypady, "An ensemble of deep recurrent neural networks for p-wave detection in electrocardiogram," in *Proc. ICASSP*, Brighton, May 2019, pp. 1284–1288.

[6] J. Gratch et al., "The distress analysis interview corpus of human and computer interviews," in *Proc. LREC*, Reykjavik, May 2014, pp. 3123–3128.

[7] M. Valstar et al., "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. AVEC*, Amsterdam, October 2016, p. 3–10.

[8] K. Kroenke et al., "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[9] S. P. Dubagunta, B. Vlasenko, and M. Magimai-Doss, "Learning voice source related information for depression detection," in *Proc. ICASSP*, Brighton, May 2019, pp. 6525–6529.

[10] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "DepAudioNet: An efficient deep model for audio based depression classification," in *Proc. AVEC*, Amsterdam, October 2016, pp. 35–42.

[11] A. Mdhaffar et al., "DL4DED: Deep learning for depressive episode detection on mobile devices," in *How AI Impacts Urban Living and Public Health*, 2019, pp. 109–121.

[12] L. Yang et al., "Multimodal measurement of depression using deep learning models," in *Proc. AVEC*, Mountain View, California, October 2017, pp. 53–59.

[13] U. Hebert-Johnson, M. Kim, O. Reingold, and G. Rothblum, "Multicalibration: Calibration for the (computationally-identifiable) masses," in *ICML*, Stockholm, June 2018, pp. 1939–1948.

[14] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky, "Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy," in *Proc. FAccT*, New York, January 2020.

[15] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "An empirical study of rich subgroup fairness for machine learning," in *Proc. FAccT '19*, Atlanta, October 2019, FAccT, pp. 100–109.

[16] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, "Multimodal inductive transfer learning for detection of Alzheimer's dementia and its severity," in *Proc. Interspeech*, Shanghai, October 2020, pp. 2212–2216.

[17] D. DeVault et al., "SimSensei kiosk: A virtual human interviewer for healthcare decision support," in *AAMAS*, Paris, May 2014.

[18] T. Baltrušaitis, P. Robinson, and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *WACV*, Lake Placid, May 2016.

[19] G. Littlewort et al., "The computer expression recognition toolbox (CERT)," in *Face and Gesture*, Barcelona, March 2011, pp. 298–305.

[20] F. Honig, A. Batliner, E. Noth, S. Schnieder, and J. Krajewski, "Automatic modelling of depressed speech: Relevant features and relevance of gender," September 2014.

[21] N. Cummins, B. Vlasenko, H. Sagha, and B. Schuller, "Enhancing speech-based depression detection through gender dependent vowel-level formant features," in *AIME*, Vienna, June 2017, pp. 209–214.

[22] B. Vlasenko, H. Sagha, N. Cummins, and B. Schuller, "Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition," in *Proc. Interspeech*, Stockholm, August 2017, pp. 3266–3270.

[23] K. Ishikawa, J. MacAuslan, and S. Boyce, "Toward clinical application of landmark-based speech analysis: Landmark expression in normal adult speech," vol. 142, no. 5, pp. EL441–EL447, November 2017.

[24] Z. Huang, J. Epps, D. Joachim, and V. Sethu, "Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 435–448, 2020.

[25] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*, fairmlbook.org, 2019, http://www.fairmlbook.org.

[26] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[27] L. T. Liu, M. Simchowitz, and M. Hardt, "The implicit fairness criterion of unconstrained learning," in *Proc. MLR*, 2019, vol. 97, pp. 4051–4060.

[28] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," vol. 5, no. 2, pp. 153–163.

[29] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," Red Hook, NY, USA, Proc. NIPS'16, pp. 3323–3331.