Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on AI for Good

FairReFuse: Referee-Guided Fusion for Multimodal Causal Fairness in Depression Detection

Jiaee Cheong[1,2,3] *, **Sinan Kalkan**[2] and **Hatice Gunes**[1]
[1] University of Cambridge
[2] Middle East Technical University
[3] The Alan Turing Institute
{jc2208, hg410}@cam.ac.uk, skalkan@metu.edu.tr

Abstract

Machine learning (ML) bias in mental health detection and analysis is becoming an increasingly pertinent challenge. Despite promising efforts indicating that multimodal methods work better than unimodal methods, there is minimal work on multimodal fairness for depression detection. We propose a causal multimodal framework which consists of two modules. Module 1 performs causal interventional debiasing via backdoor adjustment for each modality to achieve group fairness. Module 2 adaptively fuses the different modalities using a referee-based individual fairness guided fusion mechanism to address individual fairness. We conduct experiments and ablation studies on three depression datasets, D-Vlog, DAIC-WOZ and E-DAIC, and show that our framework improves classification performance as well as group and individual fairness compared to existing approaches.

1 Introduction

Mental health disorders (MHDs) are becoming increasingly prevalent [Wang *et al.*, 2007]. Despite its severity, there is currently no effective clinical characterization of MHDs which makes their diagnosis difficult, time-consuming and subjective [Maj *et al.*, 2020]. A substantial body of literature focuses on depression detection using text mining [Dalal *et al.*, 2023]. However, as humans typically display and interpret affective states through a multitude of channels, nonverbal signals such as audio-visual cues [He *et al.*, 2022; Yoon *et al.*, 2022] are just as important for depression detection. Machine learning (ML) methods have been applied to many health-related areas [Sendak *et al.*, 2020]. The natural extension of using ML for multimodal non-verbal behavioural MHD analysis and detection has proven promising [Yoon *et al.*, 2022; Zheng *et al.*, 2023; Cheong *et al.*, 2022].

Concurrently, ML bias is becoming a growing source of concern [Bolukbasi *et al.*, 2016; Cheong *et al.*, 2021]. Given the high stakes involved in MHD analysis, it is crucial to investigate and mitigate the ML biases present. Research indicated the high prevalence of gender bias across a variety

*This work was undertaken while Jiaee Cheong was a visiting PhD student at the Middle East Technical University (METU).

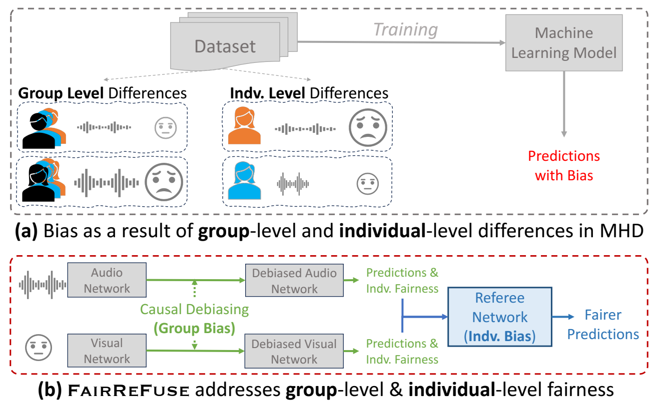

Figure 1: (a) Different individuals and groups manifest depression in different ways, causing vanilla approaches to have bias. (b) Our novel multimodal fusion approach targets both group and individual-level differences to obtain better group and individual fairness.

of tasks ranging from automated video interviews [Booth *et al.*, 2021] and image search [Feng and Shah, 2022]. However, research in gender fairness for MH has been limited with only a handful of studies investigating the problem of unimodal bias in ML methods when deployed on MHD applications [Bailey and Plumbley, 2021; Zanna *et al.*, 2022; Cheong *et al.*, 2023c]. None of the existing works have addressed gender fairness in MHD within a multimodal setting, despite the evidence that multimodal methods often work better than unimodal approaches in terms of predictive performance [Yoon *et al.*, 2022]. In order to address this gap, we have two main motivations in this paper:

Addressing Group Differences (M1). Literature indicates that females and males tend to show different behavioural symptoms when depressed [Barsky *et al.*, 2001; Ogrodniczuk and Oliffe, 2011]. As an example, as illustrated in Fig. 1, both males and females are expressing depressive symptoms. However, there are gender-specific latent representation differences in depression manifestation. For instance, certain acoustic features (e.g. MFCC) are only statistically significantly different between depressed and healthy males [Wang *et al.*, 2019]. On the other hand, compared to males, depressed females are more emotionally expressive and willing to reveal distress via behavioural cues [Barsky *et al.*, 2001; Hall *et al.*, 2000; Jansz and others, 2000]; i.e., group dif-

Study	Task	Modality	Approach		Evaluation			Fairness Measures				
			Causal	FF	GF	IF	ND	SP	EOpp	EOdd	EAcc	IF-Cons
Alasadi *et al.* [2020]	Cyberbullying Detection	VT	N	✓	✓	N	1	✓		✓		
Schmitz *et al.* [2022]	Emotion Detection	AVT	N	N	✓	N	1	✓	✓			
Yan *et al.* [2020]	Personality Assessment	AV	N	N	✓	N	1	✓			✓	
Kathan *et al.* [2022]	Humour Recognition	AV	N	N	✓	N	1	✓				
Chen *et al.* [2023]	Recommendation	AVT	N	N	✓	N	2		✓	✓		
Janghorbani *et al.* [2023]	Vision-Language Models	VT	N	N	✓	N	1					
Pena *et al.* [2023]	Automatic Recruitment	VT	N	N	✓	N	1	✓				
FAIRREFUSE	Depression	AV	✓	✓	✓	✓	3	✓	✓	✓	✓	✓

Table 1: Comparative Summary with existing Multimodal Fairness studies. Abbreviations (sorted): A: Audio. EAcc: Equal Accuracy. EOdd: Equalised Odds. EOpp: Equality of Opportunity. FF: Fairness-aware Fusion. IF-Cons: Individual Fairness (Consistency). N: No. ND: Number of Datasets. SP: Statistical Parity. T: Text. V: Visual.

ferences in depression manifestation. *Gap*: No existing ML for MHD detection approaches have considered this from a *causal* perspective. Regular depression detection models typically aim to approximate $P(Y|X)$ (X and Y denotes input and target variables respectively). $P(Y|X)$ may lead to bias since it may learn gender-specific representations that are not shared by new samples (Fig. 1). *Contribution*: This suggests that gender is a confounder which misleads a depression detection model to learn gender-specific latent representation in the training data, thus leading to prediction bias when tested on a subject of a different gender (Fig. 1). We adopt a causal approach as it provides a more principled way of representing and removing the effect of a confounder. Thus, we propose a method which approximates $P(Y|do(X))$ instead of $P(Y|X)$. The do-operation [Pearl, 2009] denotes intervening on X in order to remove the confounding relationship of gender on X. To achieve group-level gender fairness, we implement **causal interventional debiasing (CID)** using backdoor adjustment [Pearl, 2009] in order to achieve fairer representational learning for each modality.

Addressing Individual Differences (M2). Individual-level differences in depressive symptoms [Kendler *et al.*, 1994] are not accounted either in existing ML for MHD detection and analysis. *Contribution*: We propose to address this gap by taking into account individual-fairness when combining predictions across different modalities. First, we measure individual fairness using the **individual consistency scores** for each sample across the different modalities. Subsequently, we fuse the different modalities using a *referee network* that takes into account the individual fairness scores of each modality. To the best of our knowledge, we are the first to use and consider individual fairness in ML for MHD analysis.

Real-world Implication. Gender difference in depression manifestation has long been studied and recognised within fields such as medicine [Barsky *et al.*, 2001] and psychology [Hall *et al.*, 2000]. Anecdotal evidence have also often supported this view [Hall *et al.*, 2000]. However, existing ML research is unable to account for this innate group and individual subjectivity. We present the first attempt towards addressing this problem by motivating our proposed method, FAIRREFUSE, which builds on existing research on depression findings rooted in literature adjacent to traditional ML. To the best of our knowledge, ours is the first work that attempts to address the well-recognised **gender** and **individual** difference in depression manifestation. These aims align with the United Nations Sustainable Development Goal (SDG) 3[1] and SDG 5[2] respectively. The main contribution of this work is a dynamic referee-guided causal framework (FAIRREFUSE) that mitigates bias with **causal intervention** and **individual fairness-guided fusion**. We run experiments on three depression detection datasets, D-Vlog, DAIC-WOZ and E-DAIC. We demonstrate that our method was able to provide significant improvement in group and individual fairness across all datasets. The improvements are especially pronounced for D-Vlog. Results obtained on DAIC-WOZ and E-DAIC were better compared to the baseline and other existing methods. We identify **three key challenges:** dataset curation **(C1)**, appropriate evaluation **(C2)** and ethics and privacy **(C3)** as central topics that need to be tackled via collective efforts in order to promote real-world advancement in using ML to address the challenge of MHD in a fair and impactful manner.

Comparative Summary. In this work, we focus on multimodal gender fairness in MHD prediction on audio-visual datasets. To overcome the limitations in existing multimodal methods (see Table 1), we propose a referee network which dynamically learns how to fuse the different modalities. None of the current methods have combined this framework with causal intervention nor leveraged it to achieve multimodal fairness. Our work is distinct from existing work in several ways. First, we propose leveraging multimodal causal intervention to achieve multimodal fusion. Second, we use a individual fairness-guided referee network to adaptively learn the best way to fuse the different modalities. Third, we explore how the different modalities and fusion strategies impact gender fairness using both **group** and **individual** fairness measures to address the specific task of *depression detection*.

2 Related Work

Fairness in unimodal and multimodal ML. Fair ML can generally be categorised into group or individual fairness [Hort *et al.*, 2022]. Group fairness metrics typically enforce some statistical constraints across groups while individual fairness metrics seek for similar individuals to be treated similarly. Most existing works typically consider a unimodal setup which may not map to a multimodal setting. There has been minimal literature which examines ML fairness in the context of multiple modalities. Booth *et al.* [2021] demonstrated how using multiple modalities marginally improves prediction at the cost of reducing fairness for automated video

[1] "Ensure healthy lives and promote well-being for all at all ages."
[2] "Achieve gender equality and empower all women and girls."

interviews. Schmitz *et al.* [2022] studied how different multimodal approaches affect gender bias in emotion recognition. Janghorbani *et al.* [2023] presented a visual-textual benchmark dataset to assess the bias present in existing multimodal models. Mandhala *et al.* [2023] summarised the tools and frameworks available to mitigate bias in multimodal datasets. Pena *et al.* [2023] presented a new dataset of synthetic resumes to evaluate how multimodal ML is affected by demographic bias. All of the above studies did not propose any mitigation strategies. Kathan *et al.* [2022] proposed a weighted fusion approach to achieve fairness in audiovisual humour recognition. Yan *et al.* [2020] focused on adversarial bias mitigation for multimodal personality assessment. Alasadi *et al.* [2020] proposed a fairness-aware fusion framework for cyberbullying detection using a weighted approach. Chen *et al.* [2023] proposed a fairness-aware method for multimodal recommendations. No existing work has investigated fairness in multimodal *fusion* for depression detection.

Gender Fairness in ML for MHD. Only a few studies investigated gender fairness in ML-based MHD analysis [Zanna *et al.*, 2022; Bailey and Plumbley, 2021; Cheong *et al.*, 2023c]. Cheong *et al.* [2023d] proposed a data augmentation method to address the bias present within a small dataset of wellbeing coaching. Zanna *et al.* [2022] proposed an uncertainty-based approach to address the bias present in the TILES dataset. Bailey *et al.* [2021] used data re-distribution to mitigate the gender bias present in the DAIC-WOZ dataset. Cheong *et al.* [2023c] highlighted how existing bias mitigation methods do not fully address gender bias but did not propose any further mitigation strategies. Efforts have been partially hampered by the lack of datasets. Publicly available datasets are often in the form of extracted features to preserve the privacy of the subjects [Yoon *et al.*, 2022]. Research suggests that the extracted features may contain bias due to the underlying training data [Bolukbasi *et al.*, 2016; Garg *et al.*, 2018].

3 Preliminaries and Background

3.1 Notation and Problem Definition

We have a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i$ for a supervised classification problem, where $\mathbf{x}_i \in X$ is the input representing information about an individual $I_i \in \mathcal{I}$ and $y_i \in Y$ is the classification target (e.g., depressed vs. non-depressed). Distinct from conventional classification settings, each input $\mathbf{x}_i$ is composed of multiple modalities: i.e., $\mathbf{x}_i = \{\mathbf{x}_i^m \in X^m\}_m$, where m can be e.g., "image" or "audio". Each input $\mathbf{x}_i$ is associated (through an individual I_i) with a sensitive attribute $s_i \in S$ where, e.g., $S = \{\text{male}, \text{female}\}$. We are interested in finding a parameterised function $f : X \to Y$. The function $f(\cdot\,; \theta)$ estimates the probabilities for all outcomes (classes) $P(Y|\mathbf{x}_i)$. We use $P(y_i|\mathbf{x}_i)$ to denote the predicted probability for the correct class, y_i, and $\hat{y}_i \leftarrow \arg\max_y P(y|\mathbf{x}_i)$ to denote the predicted class. Finally, the pre-Softmax activations, i.e., logits, will be denoted by $\phi_i = \phi(\mathbf{x}_i; \theta)$.

3.2 Individual Fairness

Based on the principle of "similar individuals should have similar predictions", Dwork *et al.* [2012] defined *individual fairness* as the L-Lipschitz continuity of f:

$$d_y(f(\mathbf{x}_1), f(\mathbf{x}_2)) \leq L d_x(\mathbf{x}_1, \mathbf{x}_2), \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in X. \tag{1}$$

where the notion assumes suitable distance metrics $d_y(\cdot,\cdot)$ and $d_x(\cdot,\cdot)$ to be available for the predictions and the inputs respectively. Aligned with existing work [Zemel *et al.*, 2013; Mukherjee *et al.*, 2020], we use *consistency* as a measure of individual fairness. Concretely:

$$\mathcal{M}_{indv}(\mathbf{x}_i) = \left| \hat{y}_i - \frac{1}{k} \sum_{\mathbf{x}_j \in k\text{NN}(\mathbf{x}_i)} \hat{y}_j \right|, \tag{2}$$

where $k\text{NN}(\mathbf{x}_i)$ denotes the k nearest neighbours of $\mathbf{x}_i$.

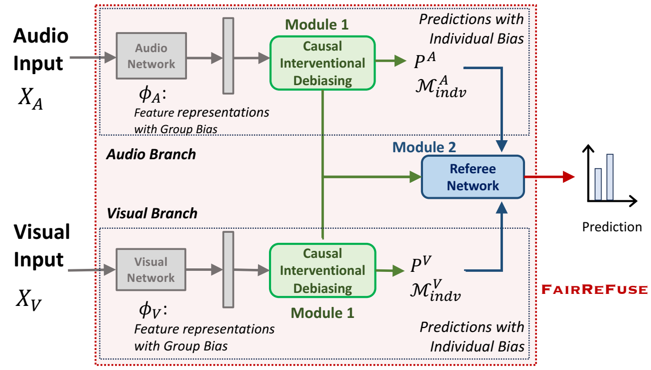

Figure 2: **Overview of FAIRREFUSE.** First, each modality network attempts to learn a fairer representation with the help of **Causal Interventional Debiasing** (Section 4.1). Second, the **Referee Network** (Section 4.2) learns to combine the predictions of the modalities dynamically using the individual fairness scores (Eqn. 2).

4 Proposed Method: FAIRREFUSE

We introduce FAIRREFUSE for fairer predictions in a multimodal classification setting. As outlined in Fig. 2, for each modality m, we employ causal intervention via back-door adjustment to remove the bias caused by the sensitive attributes. Then, individual fairness score of a sample ($\mathcal{M}^m_{indv}$ – Eq. 2) is used to dynamically fuse the predictions of each modality (Module 2). The pseudocode is shown in Algorithm 1. FAIRREFUSE mitigates bias with two novel modules:

Module 1: Causal Multimodal Interventional Debiasing for Group Fairness: Predictions made by individual modalities can have group-level biases (Fig. 1). To mitigate such modality-specific group-level bias, we adapt the unimodal work by Chen *et al.* [2022] to our multimodal setting.

Algorithm 1 FAIRREFUSE: a referee-guided fusion approach for multimodal causal fairness.

1: **Input:** Dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i$
2: **Output:** Networks for each modality (f^m) and RefNet. *# Module 1 - Causal Multimodal Interventional Debiasing (CMID):*
3: - Train each modality f^m with causal debiasing (Eq. 9)
4: - Calculate $\mathcal{M}^m_{indv}$, individual fairness scores (Eq. 2) *# Module 2 - Referee Network:*
5: - Train RefNet to maximize P_{RN} with Cross-Entropy (Eq. 10)

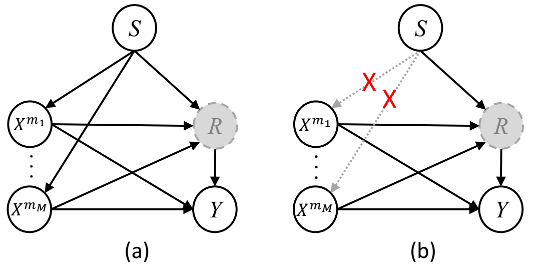

Figure 3: The causal model (a) and backdoor intervention across M modalities (b). We remove the spurious effects caused by the sensitive attributes S in order for the network to learn a debiased latent representation R independent of the sensitive attributes.

We attempt to remove the spurious confounding effect (see Fig. 3) of all modalities by intervening on the latent space of each modality, which is akin to attempting to remove the group-level modality-specific biases.

Module 2: Referee Network for Individual Fairness Guided Fusion for Individual Fairness: Module 1 addresses modality-specific group-level biases. As different modalities may have biases for different individuals (Fig. 1), we propose to estimate the modality-specific individual fairness (Eq. 2) and use these scores to fuse the different modalities. Note that, although Figs. 1 and 2 as well as the experiments focus on a bi-modal setting (with audio and vision), FAIRREFUSE can be applied to more than two modalities.

4.1 Module 1: CID

For each modality, we use causal intervention via back-door adjustment to remove the bias from the sensitive attributes.

Structural Causal Diagram

As shown in Fig. 3, there are four main variables involved. $\{X^{m_1}, ..., X^{m_M}\}$ represents the data input across M modalities, S represents the *sensitive attribute* associated with the data input, $R = \phi$ represents the latent representation learnt by the classifier and Y represents the probability outcome. The causal relationships between variables are as follows:

$S \rightarrow \{X^{m_1}, ..., X^{m_M}\}$: The sensitive attribute of the sample is bound to lead to attribute-specific features. For instance, in our setup where s represents gender, females are more likely to have higher pitched voice compared to males.

$\{X^{m_1}, ..., X^{m_M}\} \rightarrow R \leftarrow S$: R is the latent representation of X and this causal relationship is captured by $X \rightarrow R$. This latent representation R will also be determined by the sample's sensitive attribute S captured via $R \leftarrow S$.

$\{X^{m_1}, ..., X^{m_M}\} \rightarrow Y \leftarrow R$: The final outcome Y is affected by the input sample X and latent representation R as captured by $X \rightarrow Y$ and $Y \leftarrow R$ respectively.

Interventional Debiasing via Backdoor Adjustment

Since interventional debiasing is performed for each modality, to simplify notation, we drop m from all variables in this section. With reference to Fig. 3, for each modality, we attempt to model $P(Y|do(X))$ instead of $P(Y|X)$ in order to remove the bias introduced by the confounder sensitive attribute $s \in S$. Using backdoor adjustment [Pearl, 2009], we model $P(Y|do(X))$ by removing the causal relationship of $S \rightarrow X$ by marginalising over the confounder (S) for all modalities. For each modality, the debiased network can therefore be represented by:

$$P(Y|do(X = \mathbf{x})) = \sum_{s \in S} P(Y|X = \mathbf{x}, R = g(\mathbf{x}, s))P(s), \tag{3}$$

where $g(\mathbf{x}, s)$ is a function, as defined later in Eq. 5, which generates the latent representation R from X and s. We denote $P(Y|do(X = \mathbf{x}))$ of each modality m using P^m to simplify subsequent notation.

We adopt a similar approach to Chen *et al.* [2022] to estimate R from X and s. This involves a confounder dictionary, confounder attention and confounder priors. Given a confounder dictionary for each of the gender, male and female, and $r_s \in R$ is the prototype of gender s, there is one feature memory bank for each gender to store the latent features of training samples of a specific gender s. Thus,

$$r_s = \frac{1}{N_s} \sum_{\mathbf{x}_j \wedge s(\mathbf{x}_j) = s} \phi(\mathbf{x}_j), \tag{4}$$

where $\phi(\mathbf{x}_j)$ is the feature for a sample with sensitive attribute s and N_s is the number of samples for gender s. R is approximated as a weighted aggregation of all the prototypes of the specific gender s:

$$R = g(\mathbf{x}, s) = \sum_{s \in S} \alpha_s r_s P(s), \tag{5}$$

where $P(s)$ is the ratio of the number of gender s samples to the total number of training samples and α_s is the confounder attention for gender s in the confounder dictionary, calculated using scaled dot-product attention [Vaswani *et al.*, 2017]:

$$\alpha_s = \text{Softmax}\left(\frac{(W_Q \phi(\mathbf{x})^T (W_K r_s)}{\sqrt{d_m}}\right), \tag{6}$$

where $\phi(\mathbf{x})$ represents the extracted features for the current sample $\mathbf{x}$ and $W_Q \in \mathbb{R}^{d_m \times d_{in}}$ and $W_K \in \mathbb{R}^{d_m \times d_{in}}$ are weight parameters are learned.

4.2 Module 2: Referee Network for Individual Fairness Guided Multimodal Fusion

The Referee Network (Fig. 2) takes in P^m from each modality as features and attempts to learn to dynamically fuse the predictions using their individual fairness scores. We define individual fairness for a modality m by a simple extension of $\mathcal{M}_{indv}$ (Eq. 2):

$$\mathcal{M}^m_{indv}(\mathbf{x}^m_i) = \left| \hat{y}^m_i - \frac{1}{k} \sum_{\mathbf{x}^m_j \in k\text{NN}(\mathbf{x}^m_i)} \hat{y}^m_j \right|. \tag{7}$$

There are many ways to combine P^m and $\mathcal{M}^m_{indv}(\mathbf{x}^m_i)$, which can be explored through experimental analysis. We observe that a linear layer provides the best results:

$$P_{RN}(Y|\mathbf{x}_i) = \text{Softmax}\left(FC([P^m; \mathcal{M}^m_{indv}(\mathbf{x}^m_i)]_m)\right), \tag{8}$$

where $\mathbf{x}_i = \{\mathbf{x}^m_i\}_m$; $[;]_m$ represents the concatenation over M modalities and FC denotes a linear layer.

4.3 Loss Functions

We use the Cross Entropy loss while causally-debiasing each modality m:

$$\mathcal{L}_{CID}(\mathbf{x}_i^m, y_i) = \mathcal{L}_{CE}(y_i, P^m(Y|do(X^m = \mathbf{x}_i^m))). \quad (9)$$

The Referee Network (RefNet) is trained also with the Cross Entropy loss (with $\mathbf{x}_i = \{\mathbf{x}_i^m\}_m$):

$$\mathcal{L}_{RN}(\mathbf{x}_i, y_i) = \mathcal{L}_{CE}(y_i, P_{RN}(Y|X = \mathbf{x}_i)). \quad (10)$$

5 Experiment Setup and Details

5.1 Datasets

We performed our experiments using the following datasets.

	D-Vlog			DAIC-WoZ			E-DAIC		
	Y_0	Y_1	T	Y_0	Y_1	T	Y_0	Y_1	T
M	0.15	0.19	**0.34**	0.45	0.08	**0.53**	0.49	0.13	**0.62**
F	0.28	0.39	**0.66**	0.36	0.11	**0.47**	0.27	0.11	**0.38**
T	**0.42**	**0.58**	1.00	**0.81**	**0.19**	1.00	**0.76**	**0.24**	1.00

Table 2: Dataset distribution and target attribute breakdown across datasets by percentage. Abbreviations: F: Female. M: Male. T: Total. Y_0: Non-depressed. Y_1: Depressed. Red highlights imbalanced splits. Green denotes relatively balanced splits.

D-Vlog [Yoon *et al.*, 2022] consists of Youtube vlog data. **DAIC-WOZ** [Valstar *et al.*, 2016] and **E-DAIC** [Ringeval *et al.*, 2019] contain audio recordings, extracted visual features and transcripts. For all datasets, we work with the extracted features and followed the train-validate-test split provided. **Dataset Challenges:** We identify three dataset challenges: small dataset sample size **(C1a)**, class imbalance **(C1b)** and inconsistency in dataset distribution **(C1c)** which will be discussed in further detail in Section 7.

5.2 Implementation Details

We adopt Yoon *et al.*'s [2022] implementation to facilitate comparison. We train the model with the Adam optimizer [Kingma and Ba, 2014] at a learning rate of 0.0002 and a batch size of 32 for D-Vlog as stated in Yoon *et al.* For DAIC-WOZ and E-DAIC, we use a learning rate of 0.0005 and a batch size of 64. Weights are randomly initialised for all.

5.3 Baseline Models

For Module 1, we use the unimodal transformer encoder from Yoon *et al.* [2022] as the baseline. As Module 2 is a late fusion method, we compare RefNet against two other commonly used late fusion methods: (i) Ensemble method, where the final prediction is made according to the predicted class probability that is highest across all classifiers. (ii) Stacking method [Baltrušaitis *et al.*, 2018], where another classifier, a logistic regression model, is used to to provide the final classification. Our proposed method is most similar to a stacking classifier with the key difference that it weighs each modality according to the individual fairness (IF) scores and debiases the individual modalities before providing the final outcome.

5.4 Evaluation Protocols

Model Performance. We adopted the evaluation methods of existing work [Yoon *et al.*, 2022; Cheong *et al.*, 2023b] by using precision, recall and F1 to evaluate model performance.

	Method	Modality	Prec.	Rec.	F1
D-Vlog	Dosovitskiy *et al.* [2020]	AV	0.64	0.63	0.63
	Touvron *et al.* [2021]	AV	0.64	0.64	0.64
	Yin *et al.* [2022]	AV	0.65	0.64	0.65
	Wang *et al.* [2022]	AV	0.65	0.64	0.65
	Wu *et al.* [2022]	AV	0.65	0.64	0.65
	Zheng *et al.* [2023]	AV	**0.66**	0.64	0.65
	FAIRREFUSE	AV	0.61	**0.82**	**0.70**
DAIC-WoZ	Ma *et al.* [2016]	A	0.35	**1.00**	0.52
	Valstar *et al.* [2018]	A	0.32	0.86	0.46
	Williamson *et al.* [2016]	V	-	-	0.53
	Valstar *et al.* [2018]	V	0.60	0.43	0.50
	Valstar *et al.* [2018]	AV	**0.60**	0.43	0.50
	FAIRREFUSE	AV	0.52	0.60	**0.57**

Table 3: Comparison with other models which used extracted features. Best results highlighted in **bold**. Due to space constraints, the full table is made available within the Appendix of the full paper.

Group Fairness (GF). We use the most commonly used fairness metrics [Hort *et al.*, 2022; Pessach and Shmueli, 2022]: Statistical Parity ($\mathcal{M}_{SP}$), Equal Opportunity ($\mathcal{M}_{EOpp}$), Equalised Odds ($\mathcal{M}_{Odd}$) as well as Equal Accuracy ($\mathcal{M}_{EAcc}$) to evaluate group fairness. Specific formulations can be found in the appendix. We adopt the approach of existing work which considers 0.80 and 1.20 as the acceptable lower and upper fairness bounds [Zanna *et al.*, 2022]. Values outside this range are considered unfair.

Individual Fairness (IF). As there is no prior work which evaluates individual fairness for depression detection, we use *consistency* (Eq. 2) as a measure of individual fairness which aligns with existing work [Yurochkin and Sun, 2021; Mukherjee *et al.*, 2020; Zemel *et al.*, 2013].

6 Results

6.1 Baseline Performance Comparison

Table 3 presents our results compared against other methods which also worked with extracted features. We are starting with this comparison in order to demonstrate that FAIRREFUSE outperforms other methods which also worked with extracted features. D-Vlog only provides extracted data hence all the recent state-of-the art (SOTA) methods were implemented on extracted features. DAIC-WOZ and E-DAIC provided raw files in addition to the extracted features. As a result, most of the recent methods worked directly with the raw files in order to obtain better benchmark performance. We have chosen to only include methods which rely only on extracted features in order for our method to be comparable. As seen in Table 3, there is a general precision-recall trade-off across all methods hence more emphasis should be placed on the F1-score when evaluating performance results. We observe that our results are comparable and often outperform existing SOTA methods especially across the F1-score.

Summary: Despite the dataset challenges **(C1a-C1c)**, FAIRREFUSE still provides comparatively better results overall compared to existing methods for both datasets. This is significant as most of the recent studies which report higher accuracies typically work directly with the raw files. This may pose ethical and privacy concerns **(C3)** which will be discussed in further detail in Section 7.

	Approach	Method	Classification			Group Fairness				Indv. Fairness
			Prec.	Rec.	F1	$\mathcal{M}_{SP}$	$\mathcal{M}_{EOpp}$	$\mathcal{M}_{EOdd}$	$\mathcal{M}_{EAcc}$	$\mathcal{M}_{Indv}$
D-Vlog	Unimodal	Audio Network	0.60	0.58	0.63	1.07	0.71	0.59	0.82	0.54±0.22
		Visual Network	0.57	0.62	0.64	1.23	1.42	1.28	1.11	0.56±0.22
	Fusion	Baseline-AV	0.58	0.58	0.59	0.92	1.23	1.86	1.24	0.62±0.22
		Ensembles	0.59	0.83	0.69	1.20	0.97	0.95	**1.04**	0.61±0.23
		Stacking	0.59	**0.85**	0.69	1.33	0.98	**0.97**	1.07	0.60±0.25
	FAIRREFUSE	CID & RefNet	**0.61**	0.82	**0.70**	**1.03**	**1.02**	1.06	1.05	**0.80±0.24**
DAIC-WOZ	Unimodal	Audio Network	0.54	**0.66**	0.56	0.38	0.89	0.80	0.66	0.48±0.20
		Visual Network	**0.58**	0.62	0.53	**0.76**	0.81	0.83	0.66	0.56±0.24
	Fusion	Baseline-AV	0.56	0.52	0.53	0.75	0.88	0.77	**0.87**	0.56±0.20
		Ensembles	**0.58**	0.21	0.31	284.78	1.10	1.22	0.72	0.73±0.26
		Stacking	0.53	0.35	0.42	0.00	**1.05**	1.12	0.69	0.68±0.28
	FAIRREFUSE	CID & RefNet	0.52	0.60	**0.57**	22.58	1.08	**1.10**	0.74	**0.80±0.20**
EDAIC	Unimodal	Audio Network	0.50	0.51	0.50	0.53	0.64	0.70	0.74	0.52±0.21
		Visual Network	0.50	0.52	0.52	0.58	0.65	0.76	0.75	0.54±0.22
	Fusion	Baseline-AV	0.50	0.48	0.50	**0.84**	0.86	0.82	0.82	0.64±0.20
		Ensembles	0.54	0.19	0.29	281.78	1.09	1.21	0.69	0.68±0.22
		Stacking	0.55	0.37	0.44	0.00	1.10	1.24	0.72	0.70±0.23
	FAIRREFUSE	CID & RefNet	**0.56**	**0.62**	**0.60**	18.60	**1.05**	**1.11**	**0.88**	**0.78±0.22**

Table 4: A comparison of the performance and fairness across different **unimodal** and **multimodal fusion** methods and modalities where $k = 5$. Modalities. A: Audio. V:Visual. Best results are highlighted in **bold**.

6.2 Comparison with Other Fusion Methods

As our method can largely be considered a late fusion strategy, we compare FAIRREFUSE against other popular late fusion strategies: ensembles and stacking. With reference to Table 4, we see that our proposed method is comparable to or better than other fusion methods across most performance measures and especially the F1-score for all datasets. Across group fairness, our method generally ensures better group fairness for all datasets. For DAIC-WOZ and E-DAIC, across $\mathcal{M}_{EOpp}$ and $\mathcal{M}_{EOdd}$, all fusion methods improved fairness in favour of the minority group (values > 1) whereas $\mathcal{M}_{EAcc}$ still shows bias against the minority group (values < 1). It is noteworthy that according to the $\mathcal{M}_{SP}$ measure, all the fusion methods exacerbated bias compared to baseline. For DAIC-WOZ, the $\mathcal{M}_{SP}$ values for ensembles, stacking and FAIRREFUSE are 284.78, 0, and 22.58 respectively whereas for E-DAIC, the $\mathcal{M}_{SP}$ values are 281.78, 0 and 18.6 respectively. Ensembles and stacking severely exacerbated the bias whereas FAIRREFUSE lead to the same though to a less severe degree. We provide further insights to this phenomena in Section 7. Across $\mathcal{M}_{Indv}$, FAIRREFUSE also provided the fairest $\mathcal{M}_{Indv}$ score which exceeds the $\mathcal{M}_{Indv}$ scores of other methods for all datasets.

Summary: FAIRREFUSE generally out-performs other methods across most measures and achieves more consistent group and individual fairness compared to the other methods. The results are significant especially for D-Vlog. We also noted how certain fairness metrics are unsuitable for the task of depression detecion **(C2)**. We discuss **C2** as well as the dataset challenges **(C1a-C1c)** that may have impacted the results for DAIC-WOZ and E-DAIC in Section 7.

6.3 Ablation Studies

The Effects of Module 1: CID

With reference to Table 5, for the unimodal results, we see that the CID module was able to provide improvements across most metrics compared to the unimodal baselines for all datasets. For instance, for E-DAIC's audio modality, CID improved the prec., rec. and F1 from 0.50, 0.51, and 0.50 to 0.54, 0.57 and 0.56 respectively. The corresponding group fairness results also improved from 0.53, 0.64, 0.70 and 0.74 to 0.83, 0.74, 0.77 and 0.82. This trend is consistent for both modalities across all datasets. This suggests that for each modality, gender may have been a confounder as hypothesised and CID was effective in helping the model achieve group-level fairness across gender. Across the multimodal approaches, CID improves most metrics compared to baseline. For instance, for DAIC-WOZ, CID improved the prec., rec. and F1 from 0.56, 0.52 and 0.53 to 0.58, 0.59 and 0.59 respectively. The corresponding group fairness results also mostly improved from 0.75, 0.88, 0.77 and 0.87 to 0.85, 0.81, 0.83 and 1.23. The results are more pronounced for D-Vlog than DAIC-WOZ and E-DAIC.

Summary: CID is effective at improving both performance and fairness. It is most effective when the source of bias is the group difference in depression manifestation. This is distinct from the typical class imbalance problem. Despite females being the majority as evidenced in Table 2, there is still bias *against females* as seen from the baseline model in Table 4. CID which addresses the group difference is thus able to address this source of bias which existing bias mitigation methods were unable to (see Table 4 in [Cheong *et al.*, 2023c]).

The Effects of Module 2: RefNet

From Table 5, we see that RefNet improves performance and group fairness in addition to the increments provided by CID. For example, for D-Vlog, CID improves the rec. and F1 from 0.58 and 0.59 to 0.61 and 0.62 respectively. RefNet further improves the values to 0.82 and 0.70. For group fairness, we see RefNet improving beyond the CID results to achieve a near perfect group fairness score of 1.03, 1.02, 1.06 and 1.05. Across M_{Indv}, RefNet combined with CID consistently provides the fairest M_{Indv} score across all datasets. An analysis of the effects of k is within the Appendix of the full paper[3].

Summary: RefNet improves the classification and group fairness measures beyond the increment provided by the CID module. This effect is more pronounced for D-Vlog than it is for DAIC-WOZ and E-DAIC. This may be due to the fact that DAIC-WOZ and E-DAIC are much more challenging

[3] https://www.repository.cam.ac.uk/handle/1810/368887

		Method				Classification			Group Fairness				Indv. Fairness
		CID	MM	RN	Mod.	Prec.	Rec.	F1	$\mathcal{M}_{SP}$	$\mathcal{M}_{EOpp}$	$\mathcal{M}_{EOdd}$	$\mathcal{M}_{EAcc}$	$\mathcal{M}_{Indv}$
D-Vlog					A	0.60	0.58	0.63	1.07	0.71	0.59	0.82	0.54±0.22
		✓			A	**0.68**	0.72	0.69	0.97	1.12	0.78	1.08	0.61±0.21
					V	0.57	0.62	0.64	1.23	1.42	1.28	1.11	0.56±0.22
		✓			V	0.61	0.67	0.68	1.12	1.05	1.17	1.19	0.66±0.20
			✓		AV	0.58	0.58	0.59	0.92	1.23	1.86	1.24	0.62±0.22
		✓	✓		AV	0.63	0.61	0.62	1.29	1.11	1.18	1.09	0.60±0.19
	FAIRREFUSE	✓	✓	✓	AV	0.61	**0.82**	**0.70**	**1.03**	**1.02**	**1.06**	**1.05**	**0.80±0.24**
DAIC-WOZ					A	0.54	**0.66**	0.56	0.38	0.89	0.80	0.66	0.48±0.20
		✓			A	0.55	0.64	0.58	0.69	0.88	0.76	**0.90**	0.67±0.22
					V	0.58	0.62	0.53	0.76	0.81	0.83	0.66	0.56±0.24
		✓			V	**0.58**	0.63	**0.60**	0.73	0.76	0.75	1.12	0.71±0.24
			✓		AV	0.56	0.52	0.53	0.75	0.88	0.77	0.87	0.56±0.20
		✓	✓		AV	0.58	0.59	0.59	**0.85**	0.81	0.83	1.23	0.70±0.25
	FAIRREFUSE	✓	✓	✓	AV	0.52	0.60	0.57	22.58	**1.08**	**1.18**	0.74	**0.80±0.20**
E-DAIC					A	0.50	0.51	0.50	0.53	0.64	0.70	0.74	0.52±0.21
		✓			A	0.54	0.57	0.56	0.83	0.74	0.77	0.82	0.64±0.21
					V	0.50	0.52	0.52	0.58	0.65	0.76	0.75	0.54±0.22
		✓			V	0.53	0.58	0.58	**0.88**	0.81	0.86	0.80	0.70±0.22
			✓		AV	0.50	0.48	0.50	0.87	0.84	0.81	0.86	0.68±0.24
		✓	✓		AV	0.52	0.52	0.51	0.84	0.86	0.82	0.82	0.72±0.20
	FAIRREFUSE	✓	✓	✓	AV	**0.56**	**0.62**	**0.60**	18.60	**1.05**	**1.11**	**0.88**	**0.78±0.22**

Table 5: **Ablation Results:** Performance and Fairness Results of the ablation studies. A: Audio. V:Visual. CID represents Module 1: Causal Interventional Debiasing. MM represents Multi-modal. RN represents Module 2: Referee Network. Best results are highlighted in **bold**.

datasets due to **(C1a - C1c)**. This will be further discussed in Section 7. Our proposed method's efficacy is most effectively captured across the individual fairness measure $\mathcal{M}_{Indv}$ across all datasets. This suggests that RefNet is particularly effective at achieving individual-level fairness.

7 Conclusion and Discussion

Conclusion. We present a novel framework to achieve both group and individual level fairness for the task of depression detection. We focus specifically on extracted audio-visual data as this is less studied compared to text-based depression detection research. We show that both Module 1: CID and Module 2: RefNet were effective at improving ML performance and fairness. They are most effective when used together and are able to achieve good performance and fairness results without requiring access to the raw files. This respects the privacy and anonymity of the subjects. In addition, we highlight three key challenges as a call to the community to address the identified issues collectively. This is in tandem with the goal of addressing the real-world challenge of MHD in order to achieve social good for all.

Discussion. C1: Dataset Challenges. Compared to D-Vlog, the results seem less effective for DAIC-WOZ and E-DAIC. From Table 2, we see that there are significantly **less samples** (**C1a**) in DAIC-WOZ (185) and E-DAIC (268) compared to D-Vlog (961). Second, there is **class imbalance** (**C1b**) as seen in Table 2. D-Vlog is balanced across classes (Y_0: 0.42 vs Y_1: 0.58) but imbalanced across gender (M: 0.34 vs F: 0.66). DAIC-WOZ (Y_0: 0.81 vs Y_1: 0.19) and E-DAIC (Y_0: 0.76 vs Y_1: 0.24) are both imbalanced across classes. Third, Table 6 within the Appendix, suggest a significant **distribution shift** (**C1c**) between the training and testing set for DAIC-WOZ and E-DAIC. For instance, for DAIC-WOZ, the training set contains more males than females whereas the testing set contains more females than males. The training set contains more males of class Y_0 whereas the testing set contains more females of class Y_0. The smaller sample size (**C1a**), class imbalance (**C1b**) and inconsistency in dataset distribution (**C1c**) may have impacted the results for DAIC-WOZ and E-DAIC. Vabalas *et al.* [2019] demonstrated that small datasets and small sample sizes cause ML in MHD to be more vulnerable or sensitive to changes in data distribution. Our results support the hypothesis that this may have lead to biased outcomes. Future dataset owners can consider providing more samples with lesser class imbalance and more consistent data distribution as well as identifying the root cause of bias [Cheong *et al.*, 2023a] to mitigate this challenge.

C2: Inadequacy of Metrics. Moreover, existing fairness metrics are inadequate to deal with the small dataset challenge prevalent for depression detection. The small denominator resulting from the small sample size for DAIC-WOZ and E-DAIC inadvertently lead to massive numbers which cannot be interpreted without adequate context. Future work may consider proposing more appropriate fairness metrics or evaluation methods and adopting other approaches [Churamani *et al.*, 2023] which takes this challenge into account.

C3: Ethics and Privacy. We lack publicly available datasets due to the sensitive nature of the problem setting. Some MH datasets (e.g., the Turkish BD Corpus [Çiftçi *et al.*, 2018] and the Pittsburgh [Yang *et al.*, 2012]) which were previously publicly available for research purposes are no longer made available. Recent datasets have only released the extracted features due to privacy concerns. This necessitates the urgency to advance research practices that takes ethical concerns into consideration. The data collection procedure and proposed methods should respect subjects' privacy and anonymity and perform well across classification and fairness. Our work presents the first step towards that direction.

Limitations. We assume the availability of sensitive attribute labels, which is the common setting in the bias mitigation literature. It is possible to extend our framework to work without this assumption. We only evaluated our methods on three datasets with two modalities. Future work should consider experimenting on more datasets and adapting this approach to other modalities beyond audio-visual sources.

Acknowledgements

Open access: The authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. **Data access:** This study involved secondary analyses of existing datasets. All datasets are described and cited accordingly. **Funding:** J. Cheong is supported by the Alan Turing Institute doctoral studentship, the Leverhulme Trust and further acknowledges resource support from METU. H. Gunes' work is supported by the EPSRC/UKRI project ARoEq under grant ref. EP/R030782/1.

References

[Alasadi *et al.*, 2020] Jamal Alasadi, Ramanathan Arunachalam, Pradeep K. Atrey, and Vivek K. Singh. A fairness-aware fusion framework for multimodal cyberbullying detection. In *(BigMM)*, pages 166–173, 2020.

[Bailey and Plumbley, 2021] Andrew Bailey and Mark D Plumbley. Gender bias in depression detection using audio features. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 596–600. IEEE, 2021.

[Baltrušaitis *et al.*, 2018] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[Barsky *et al.*, 2001] Arthur J Barsky, Heli M Peekna, and Jonathan F Borus. Somatic symptom reporting in women and men. *Journal of general internal medicine*, 16(4):266–275, 2001.

[Bolukbasi *et al.*, 2016] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NeurIPS*, 29, 2016.

[Booth *et al.*, 2021] Brandon M Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K D'Mello. Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *ICMI*, 2021.

[Chen *et al.*, 2022] Yingjie Chen, Diqi Chen, Tao Wang, Yizhou Wang, and Yun Liang. Causal intervention for subject-deconfounded facial action unit recognition. In *AAAI*, pages 374–382, 2022.

[Chen *et al.*, 2023] Weixin Chen, Li Chen, Yongxin Ni, Yuhan Zhao, Fajie Yuan, and Yongfeng Zhang. Fmmrec: Fairness-aware multimodal recommendation. *arXiv preprint*, 2023.

[Cheong *et al.*, 2021] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. The hitchhiker's guide to bias and fairness in facial affective signal processing: Overview and techniques. *IEEE Signal Processing Magazine*, 38(6):39–49, 2021.

[Cheong *et al.*, 2022] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Causal fairness for affect recognition. *NeurIPS AFCP Workshop*, 2022.

[Cheong *et al.*, 2023a] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Causal structure learning of bias for fair affect recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[Cheong *et al.*, 2023b] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Counterfactual fairness for facial expression recognition. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, Proceedings, Part V*, pages 245–261. Springer, 2023.

[Cheong *et al.*, 2023c] Jiaee Cheong, Selim Kuzucu, Sinan Kalkan, and Hatice Gunes. Towards gender fairness for mental health prediction. In *IJCAI 2023*, 2023.

[Cheong *et al.*, 2023d] Jiaee Cheong, Micol Spitale, and Hatice Gunes. "it's not fair!" – fairness for a small dataset of multimodal dyadic mental well-being coaching. In *ACII 2023*, 2023.

[Churamani *et al.*, 2023] Nikhil Churamani, Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Towards causal replay for knowledge rehearsal in continual learning. In *PMLR*, pages 63–70, 2023.

[Çiftçi *et al.*, 2018] Elvan Çiftçi, Heysem Kaya, Hüseyin Güleç, and Albert Ali Salah. The turkish audio-visual bipolar disorder corpus. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6. IEEE, 2018.

[Dalal *et al.*, 2023] Sumit Dalal, Sarika Jain, and Mayank Dave. Early depression detection using textual cues from social data: A research agenda. In *IHIC*, pages 393–406. Springer, 2023.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[Dwork *et al.*, 2012] 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.

[Feng and Shah, 2022] Yunhe Feng and Chirag Shah. Has ceo gender bias really been fixed? adversarial attacking and improving gender fairness in image search. In *AAAI*, volume 36, pages 11882–11890, 2022.

[Garg *et al.*, 2018] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[Hall *et al.*, 2000] Judith A Hall, Jason D Carter, and Terrence G Horgan. Gender differences in nonverbal communication of emotion. *Gender and emotion: Social psychological perspectives*, pages 97–117, 2000.

[He *et al.*, 2022] Lang He, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo, Hongyu Wang, Songtao Ding, Zhongmin Wang, et al. Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, 80:56–86, 2022.

[Hort *et al.*, 2022] Max Hort, Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. Bias mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*, 2022.

[Janghorbani and De Melo, 2023] Sepehr Janghorbani and Gerard De Melo. Multi-modal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *EACL*, pages 1717–1727, 2023.

[Jansz and others, 2000] Jeroen Jansz et al. Masculine identity and restrictive emotionality. *Gender and emotion: Social psychological perspectives*, pages 166–186, 2000.

[Kathan *et al.*, 2022] Alexander Kathan, Shahin Amiriparian, Lukas Christ, Andreas Triantafyllopoulos, Niklas Müller, Andreas König, and Björn W Schuller. A personalised approach to audiovisual humour recognition and its individual-level fairness. In *MuSe '22'*, pages 29–36, 2022.

[Kendler *et al.*, 1994] Kenneth S Kendler, Ellen E Walters, Kim R Truett, Andrew C Heath, Michael C Neale, Nicholas G Martin, and Lindon J Eaves. Sources of individual differences in depressive symptoms: analysis of two samples of twins and their families. *The American Journal of Psychiatry*, 151(11), 1994.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, 2014.

[Ma *et al.*, 2016] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. Depaudionet: An efficient deep model for audio based depression classification. In *6th Intl. Workshop on audio/visual emotion challenge*, 2016.

[Maj *et al.*, 2020] Mario Maj, Dan J Stein, Gordon Parker, Mark Zimmerman, Giovanni A Fava, Marc De Hert, Koen Demyttenaere, Roger S McIntyre, Thomas Widiger, and Hans-Ulrich Wittchen. The clinical characterization of the adult patient with depression aimed at personalization of management. *World Psychiatry*, 19(3):269–293, 2020.

[Mandhala *et al.*, 2023] Venkata Naresh Mandhala, Debnath Bhattacharyya, and Divya Midhunchakkaravarthy. A novel study on tools and frameworks for mitigating bias in multimodal datasets. In *ICCIC*, pages 277–283. Springer, 2023.

[Mukherjee *et al.*, 2020] Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metrics from data. In *ICML*, pages 7097–7107. PMLR, 2020.

[Ogrodniczuk and Oliffe, 2011] John S Ogrodniczuk and John L Oliffe. Men and depression. *Canadian Family Physician*, 57(2):153–155, 2011.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge Uni. Press, 2009.

[Peña *et al.*, 2023] Alejandro Peña, Ignacio Serna, Aythami Morales, Julian Fierrez, Alfonso Ortega, Ainhoa Herrarte, Manuel Alcantara, and Javier Ortega-Garcia. Human-centric multimodal machine learning: Recent advances and testbed on ai-based recruitment. *SN Computer Science*, 4(5):434, 2023.

[Pessach and Shmueli, 2022] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[Ringeval *et al.*, 2019] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge. *AVEC 2019*, 2019.

[Schmitz *et al.*, 2022] Matheus Schmitz, Rehan Ahmed, and Jimi Cao. Bias and fairness on multimodal emotion detection algorithms. *arXiv preprint arXiv:2205.08383*, 2022.

[Sendak *et al.*, 2020] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. "the human body is a black box" supporting clinical decision-making with deep learning. In *FAccT*, 2020.

[Song *et al.*, 2018] Siyang Song, Linlin Shen, and Michel Valstar. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *FG 2018*, pages 158–165. IEEE, 2018.

[Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*. PMLR, 2021.

[Vabalas *et al.*, 2019] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J Casson. Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11), 2019.

[Valstar *et al.*, 2016] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Depression, mood, and emotion recognition workshop and challenge. In *Avec 2016*, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[Wang *et al.*, 2007] Philip S Wang, Sergio Aguilar-Gaxiola, Jordi Alonso, Matthias C Angermeyer, Guilherme Borges, Evelyn J Bromet, Ronny Bruffaerts, Ron De Graaf, Oye Gureje, et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the who world mental health surveys. *The Lancet*, 370(9590):841–850, 2007.

[Wang *et al.*, 2019] Jingying Wang, Lei Zhang, Tianli Liu, Wei Pan, Bin Hu, and Tingshao Zhu. Acoustic differences between healthy and depressed people: a cross-situation study. *BMC psychiatry*, 19:1–12, 2019.

[Wang *et al.*, 2022] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *CVPR*, 2022.

[Williamson *et al.*, 2016] James R Williamson, Elizabeth Godoy, Miriam Cha, Pooya Khorrami, Youngjune Gwon, Charlie Dagli, and Thomas F Quatieri. Detecting depression using vocal, facial and semantic communication cues. In *6th Intl. Workshop on Audio/Visual Emotion Challenge*, 2016.

[Wu *et al.*, 2022] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *ECCV*. Springer, 2022.

[Yan *et al.*, 2020] Shen Yan, Di Huang, and Mohammad Soleymani. Mitigating biases in multimodal personality assessment. In *ICMI*, pages 361–369, 2020.

[Yang *et al.*, 2012] Ying Yang, Catherine Fairbairn, and Jeffrey F Cohn. Detecting depression severity from vocal prosody. *IEEE transactions on affective computing*, 4(2):142–150, 2012.

[Yin *et al.*, 2022] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *CVPR*, 2022.

[Yoon *et al.*, 2022] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. D-vlog: Multimodal vlog dataset for depression detection. *AAAI*, 36(11), Jun. 2022.

[Yurochkin and Sun, 2021] Mikhail Yurochkin and Yuekai Sun. Sensei: Sensitive set invariance for enforcing individual fairness. In *International Conference on Learning Representations*, 2021.

[Zanna *et al.*, 2022] Khadija Zanna, Kusha Sridhar, Han Yu, and Akane Sano. Bias reducing multitask learning on mental health prediction. In *ACII*, pages 1–8. IEEE, 2022.

[Zemel *et al.*, 2013] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, pages 325–333. PMLR, 2013.

[Zheng *et al.*, 2023] Wenbo Zheng, Lan Yan, and Fei-Yue Wang. Two birds with one stone: Knowledge-embedded temporal convolutional transformer for depression detection and emotion recognition. *IEEE Transactions on Affective Computing*, 2023.