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

Jiaee Cheong\textsuperscript{1,2,3}*, Sinan Kalkan\textsuperscript{2} and Hatice Gunes\textsuperscript{1}
\textsuperscript{1} University of Cambridge
\textsuperscript{2} Middle East Technical University
\textsuperscript{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, non-verbal 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 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-
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 a *causal interventional debiasing* (CID) using backdoor adjustment [Pearl, 2009] in order to achieve fairer representative 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 to address 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 an 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 equally. Some statistical constraints across groups while individual fairness metrics seek for similar individuals to be treated equally. 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] summarised the tools and frameworks available to mitigate bias in multimodal datasets. Bailey and Plumley, 2021; Cheong et al. [2023c] 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 $D = \{(x_i, y_i)\}$ for a supervised classification problem, where $x_i \in X$ 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 $x_i$ is composed of multiple modalities: i.e., $x_i = \{x_i^m \in X^m\}_{m}$, where $m$ can be e.g., “image” or “audio”. Each input $x_i$ 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 $f : X \rightarrow Y$. The function $f(\cdot; \theta)$ estimates the probabilities for all outcomes (classes) $P(Y|x_i)$. We use $P(y_i|x_i)$ to denote the predicted probability for the correct class, $y_i$, and $\hat{y}_i = \arg \max_{y \in Y} P(y|x_i)$ to denote the predicted class. Finally, the pre-Softmax activations, i.e., logits, will be denoted by $\phi_i = \phi(x_i; \theta)$.

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 $f$:

$$d_y(f(x_1), f(x_2)) \leq Ld_x(x_1, x_2), \quad \forall x_1, x_2 \in X.$$  

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:

$$\mathcal{M}_{indv}(x_i) = \left\| \hat{y}_i - \frac{1}{k} \sum_{x_j \in \text{kNN}(x_i)} \hat{y}_j \right\|,$$

where $\text{kNN}(x_i)$ denotes the $k$ nearest neighbours of $x_i$.

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 ($\mathcal{M}_{indv}^m - \text{Eq. 2}$) 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 Intervenational 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 $D = \{(x_i, y_i)\}$
2. **Output:** Networks for each modality ($f^m$) and RefNet. 

   # Module 1 - Causal Multimodal Intervenational Debiasing (CMID):
3. - Train each modality $f^m$ with causal debiasing (Eq. 9)
4. - Calculate $\mathcal{M}_{indv}^m$, individual fairness scores (Eq. 2) 

   # Module 2 - Referee Network:
5. - Train RefNet to maximize $F_RN$ with Cross-Entropy (Eq. 10)
This latent representation \( S \) is captured via an attention for gender \( s \) in the confounder dictionary, calculated using scaled dot-product attention [Vaswani et al., 2017]:

\[
\alpha_s = \text{Softmax} \left( \frac{(W_Q \phi(x)^T W_K r_s)}{\sqrt{d_m}} \right),
\]

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

### 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):

\[
\mathcal{M}_{indv}^m(x_i^m) = \left| \hat{y}_i^m - \frac{1}{K} \sum_{x_j^m \in \text{KNN}(x_i^m)} \hat{y}_j^m \right|.
\]

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

\[
P_{RN}(Y|x_i) = \text{Softmax} \left( FC([P^m; \mathcal{M}_{indv}^m(x_i^m)]) \right),
\]

where \( x_i = \{x_i^m \}_{m=1}^M \) 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 $m$:

$$L_{CID}(x^m_i, y_i) = L_{CE}(y_i, P^m(Y|do(X^m = x^m_i))). \tag{9}$$

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

$$L_{RN}(x_i, y_i) = L_{CE}(y_i, P_{RN}(Y|X = x_i)). \tag{10}$$

5 Experiment Setup and Details

5.1 Datasets

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 32 for D-Vlog as stated in Yoon 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.

5.3 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 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Modality</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-Vlog</td>
<td>AV</td>
<td>0.64</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>AV</td>
<td>0.64</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>AV</td>
<td>0.65</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>AV</td>
<td>0.65</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>AV</td>
<td>0.65</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>AV</td>
<td>0.66</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>AV</td>
<td>0.66</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>AV</td>
<td>0.66</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>AV</td>
<td>0.32</td>
<td>0.38</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>AV</td>
<td>0.60</td>
<td>0.43</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>AV</td>
<td>0.60</td>
<td>0.43</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>0.35</td>
<td>1.00</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>0.52</td>
<td>0.60</td>
<td>0.57</td>
</tr>
</tbody>
</table>

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.

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.
is noteworthy that according to the M in Section 7. Across the degree. We provide further insights to this phenomena fusion methods exacerbated bias compared to baseline. For < still shows bias against the minority group (values 6.2 Comparison with Other Fusion Methods

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.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Method</th>
<th>Classification</th>
<th>Group Fairness</th>
<th>Indiv. Fairness</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-Vlog</td>
<td>Unimodal</td>
<td>Audio Network</td>
<td>0.60</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Visual Network</td>
<td>0.57</td>
<td>0.62</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Baseline/AV</td>
<td>0.58</td>
<td>0.58</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>Fusion</td>
<td>Ensembles</td>
<td>0.59</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stacking</td>
<td>0.59</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>FairREFUSE</td>
<td>CID &amp; RefNet</td>
<td>0.61</td>
<td>0.82</td>
</tr>
<tr>
<td>DAIC-WOZ</td>
<td>Unimodal</td>
<td>Audio Network</td>
<td>0.54</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>Visual Network</td>
<td>0.58</td>
<td>0.62</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>Baseline/AV</td>
<td>0.56</td>
<td>0.52</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>Fusion</td>
<td>Ensembles</td>
<td>0.58</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stacking</td>
<td>0.53</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>FairREFUSE</td>
<td>CID &amp; RefNet</td>
<td>0.52</td>
<td>0.60</td>
</tr>
<tr>
<td>EDAIC</td>
<td>Unimodal</td>
<td>Audio Network</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>Visual Network</td>
<td>0.50</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Baseline/AV</td>
<td>0.50</td>
<td>0.48</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Fusion</td>
<td>Ensembles</td>
<td>0.54</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stacking</td>
<td>0.55</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>FairREFUSE</td>
<td>CID &amp; RefNet</td>
<td>0.56</td>
<td>0.62</td>
</tr>
</tbody>
</table>

6.2 Comparison with Other Fusion Methods

As our method can largely be considered a late 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 $M_{EOpp}$ and $M_{EOdd}$, all fusion methods improved fairness in favour of the minority group (values $> 1$) whereas $M_{EAcc}$ still shows bias against the minority group (values $< 1$). It is noteworthy that according to the M_SP measure, all the fusion methods exacerbated bias compared to baseline. For DAIC-WOZ, the M_SP values for ensembles, stacking and FairREFUSE are 284.78, 0, and 22.58 respectively whereas for E-DAIC, the 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 $M_{Indv}$, FairREFUSE also provided the fairest $M_{Indv}$ score which exceeds the $M_{Indv}$ 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 detection (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 prec. 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.

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
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 $M_{\text{Indiv}}$ 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 less males than females. 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 [Churaman 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 [Ciftci 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.

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.
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


