Towards Gender Fairness for Mental Health Prediction

Jiaee Cheong[1,3] , **Selim Kuzucu**[2] , **Sinan Kalkan**[2] and **Hatice Gunes**[1]
[1] University of Cambridge
[2] Middle East Technical University
[3] The Alan Turing Institute
{jiaee.cheong, hatice.gunes}@cl.cam.ac.uk, {selim.kuzucu, skalkan}@metu.edu.tr

Abstract

Mental health is becoming an increasingly prominent health challenge. Despite a plethora of studies analysing and mitigating bias for a variety of tasks such as face recognition and credit scoring, research on machine learning (ML) fairness for mental health has been sparse to date. In this work, we focus on *gender* bias in mental health and make the following contributions. First, we examine whether bias exists in existing mental health datasets and algorithms. Our experiments were conducted using Depresjon, Psykose and D-Vlog. We identify that both data and algorithmic bias exist. Second, we analyse strategies that can be deployed at the pre-processing, in-processing and post-processing stages to mitigate for bias and evaluate their effectiveness. Third, we investigate factors that impact the efficacy of existing bias mitigation strategies and outline recommendations to achieve greater gender fairness for mental health. Upon obtaining counter-intuitive results on D-Vlog dataset, we undertake further experiments and analyses, and provide practical suggestions to avoid hampering bias mitigation efforts in ML for mental health.

1 Introduction

Mental health disorders (MHDs) often impose serious burdens on individuals, families and society, and are becoming increasingly prevalent world-wide [Wang *et al.*, 2007; Mathers and Loncar, 2006]. Despite the severity of MHDs, there is currently no unique and effective clinical characterization of MHDs which makes their detection and the diagnosis difficult, time-consuming and subjective [Maj *et al.*, 2020]. Machine learning (ML) methods have been successfully applied to many real-world and health-related areas [Sendak *et al.*, 2020]. The natural extension of using ML for MHD analysis and detection has proven to be promising [Long *et al.*, 2022; He *et al.*, 2022; Zhang *et al.*, 2020].

On the other hand, ML bias is becoming an increasing source of concern [Buolamwini and Gebru, 2018; Barocas *et al.*, 2017]. Given the high stakes involved in MHD analysis and prediction, it is crucial to investigate and mitigate the ML biases present. A key challenge for fair ML in MHD analysis is the lack of publicly available datasets due to the sensitive nature of the problem setting. In order to evaluate and mitigate bias, sensitive attributes such as gender, ethnicity and age will be required. Even if such information were collected, they are often not made available in order to protect the subjects' privacy. Given the above, research in this area has been limited with only a handful of work investigating the problem of gender bias in ML methods when deployed on MHD applications [Bailey and Plumbley, 2021; Zanna *et al.*, 2022]. Bias mitigation can be conducted at either the pre-processing, in-processing or post-processing stage. Existing research only employed a single mitigation strategy and did not conduct fairness evaluation across different criteria. This is crucial as bias may be exacerbated if the incorrect criteria is chosen [Lee *et al.*, 2022]. Thus, the recommended practice is the inclusion of more fairness metrics [Hort *et al.*, 2022]. Moreover, it has been highlighted by [Pagano *et al.*, 2023] that the proliferation of fairness metrics and mitigation techniques has resulted in (i) a lack of direction about which is the appropriate metric and mitigation option and (ii) the need for use-case specific fairness research due to the different types and nature of bias given the use-case related specificity [Cheong *et al.*, 2023].

We hope to tackle the aforementioned gaps by addressing the following research questions (RQs). **RQ 1:** Is there gender bias in MHD models and non-lab-based datasets? If present, **what are the primary sources** of gender bias? **RQ 2: How effectively** can the gender bias be mitigated at the pre-processing, in-processing and post-processing stages? **RQ 3: What are the factors** that hamper gender bias mitigation for MHD? By doing so, we hope to address the real-world challenge of mental well-being and work towards reducing the gender bias present in existing mental health algorithms and datasets. These aims align with the United Nations Sustainable Development Goal (SDG) 3[1] and SDG 5[2] respectively. We obtain counter-intuitive results which prompted further experimentation and analysis. We observe several factors that might not only hamper bias mitigation efforts, but exacerbate the very bias we wish to mitigate. We highlight the source of such potential pitfalls and suggest workarounds to address the pertinent challenge of bias in MHD.

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

2 Related Work

Machine Learning for Mental Health. Recent advances in ML have prompted efforts to deploy deep learning methods for mental health analysis and detection [He *et al.*, 2022; Zhang *et al.*, 2020; Long *et al.*, 2022]. Existing works can generally be categorised by data modality. A line of work seeks to monitor mental health by monitoring physiological data such as heart-rate variability and respiratory rate [Yau *et al.*, 2022; Mundnich *et al.*, 2020] or motor activity data such as the intensity and duration of movement [Jakobsen *et al.*, 2020b; Garcia-Ceja *et al.*, 2018]. Another line of work seeks to detect and analyse an individual's mental health through the use of audio-visual sources [He *et al.*, 2022; Yoon *et al.*, 2022; Zhang *et al.*, 2020]. Audio-visual (AV) datasets typically include behavioural signals such as facial affect, body gestures and vocal intensity [DeVault *et al.*, 2014; Gratch *et al.*, 2014; He *et al.*, 2022]. The core motivation behind using these data is the findings that physiological and behavioural data can be used to distinguish an individual's affective or mental state. For instance, individuals struggling with bipolar depression typically display increased variability in activity [Scott *et al.*, 2017] and psycho-motor retardation which can manifest in the form of slower speech and response time [Yamamoto *et al.*, 2020].

Fair Machine Learning. The two main types of bias identified are dataset and algorithmic bias. Dataset bias can largely be understood as the bias stemming from the data whereas algorithmic bias can be understood as the bias that occurred during the algorithm training process [Mehrabi *et al.*, 2021]. To achieve ML fairness, we can mitigate bias at the pre-processing, in-processing or post-processing stages [Barocas *et al.*, 2017; Mehrabi *et al.*, 2021; Cheong *et al.*, 2021]. Pre-processing methods typically attempt to mitigate bias at the data-level by collecting or resampling datapoints in order to create a balanced dataset. In-processing methods mainly involve model-level interventions at different stages of the ML models or algorithms to mitigate bias. Post-processing methods chiefly modify the model output to achieve fairer predictions. There are a multitude of fairness metrics available to evaluate bias [Barocas *et al.*, 2017; Mehrabi *et al.*, 2021; Hort *et al.*, 2022]. Picking the appropriate fairness metric is important as it is used to *determine* the degree of bias present and to *evaluate* the effectiveness of bias mitigation strategies [Hort *et al.*, 2022]. [Hort *et al.*, 2022] also highlighted that 2.7 datasets were used per publication on average. In line with this, we utilise three datasets and conduct bias mitigation using methods from all three stages.

Gender Fairness for Mental Health Analysis. Research in gender fairness for machine-learning-based MHD analysis has been limited [Zanna *et al.*, 2022; Bailey and Plumbley, 2021]. [Zanna *et al.*, 2022] proposed an uncertainty-based loss re-weighting approach to address the bias present in the TILES dataset. [Bailey and Plumbley, 2021] demonstrated the effectiveness of using an existing bias mitigation method, data re-distribution, to mitigate the gender bias present in the DAIC-WOZ dataset. However, both existing works only focused on a single dataset, relied on self-reporting scores and only attempted bias mitigation at a single stage.

Comparative Summary. Our contributions can be summarised as follows. First, our work provides a comprehensive account of bias evaluation and mitigation across three publicly-available MHD datasets collected in the real world as datasets collected within a lab setting may not capture the naturalistic behaviour of individuals struggling with MHD in the real world [Huang *et al.*, 2020]. In addition, existing works chiefly rely on self-reported scores. In contrast, our work attempts to focus on patients who have been diagnosed by clinical experts and are likely to be prescribed with MHD medication. In Depresjon and Psykose, clinical experts diagnosed the patients whereas in D-Vlog, annotators were instructed to identify depressed vlogs based on statements which indicate that they are either suicidal or on MHD medication. Second, we attempt to identify the primary source of bias as opposed to solely applying convenient bias mitigation strategies. We hypothesise that this is an important factor to ensure that mitigation efforts work as intended. Third, we evaluate the effectiveness of popular fairness metrics with regards to *bias detection* and *mitigation strategy evaluation*. We utilise the four most popular fairness measures in order to ensure thorough experimentation and analysis. Table 1 summarises the differences between existing works and our work.

3 Methodology

We first formulate our research problem, and then introduce the mitigation methods and bias measures used.

3.1 Notation and Problem Definition

We approach MHD detection as a classification problem where we have a dataset D which consists of $\{(\mathbf{x}_i, y_i)\}_i$ values where $\mathbf{x}_i \in X$ is a tensor representing information (e.g. physiological signals, facial images) about an individual I and $y_i \in Y$ is the outcome (e.g. 1 depressed vs. 0 non-depressed) that we wish to predict. Each input $\mathbf{x}_i$ is associated (through an individual I) with a sensitive attribute $s(\mathbf{x}_i) \in S$ where $S = \{male, female\}$. This is a classification problem where we are interested in finding a parameterised function f with $f : X \to Y$. 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.

3.2 Bias Mitigation Methods

For all methods, we have chosen the popular methods according to [Hort *et al.*, 2022]. Further details about the specific architecture will be addressed in Section 4.

Pre-processing: Data Augmentation. Pre-processing methods typically attempt to mitigate bias prior to model training. We leverage a pre-processing Mixup data augmentation strategy proposed by [Zhang *et al.*, 2017]. The intuition behind the method is that it generates new samples by mixing up different features and their corresponding labels to prevent a learner from being too confident about the learned relationship between the features and their labels. A new training sample $(\mathbf{x}', y')$ is generated by $\mathbf{x}' = \lambda \mathbf{x}_i + (1 - \lambda)\, \mathbf{x}_j$ and $y' = \lambda y_i + (1 - \lambda)\, y_j$. This preserves the relation between the augmented data and the supervision signal. We do so for the minority group in order to obtained balanced samples across gender.

			Mitigation			Fairness Measures						
Study	Dataset	Data Type	Pre	In	Post	SP	EOpp	EOdd	EAcc	SB	L/R	ND
[Bailey, 2021]	DAIC-WOZ	AV	✓			✓				N	L	1
[Zanna, 2022]	TILES 2018	P		✓				✓		N	R	1
Ours	Depresjon, Psykose, D-Vlog	MA, AV	✓	✓	✓	✓	✓	✓	✓	Y	R	3

Table 1: Comparative Summary with Existing Work. Abbreviations: P: Physiological. AV: Audio-Visual MA: Motor Activity. L: Lab. R: Real world. SP: Statistical Parity. EOpp: Equality of Opportunity. DI: Disparate Impact. EOdd: Equalised Odds. EAcc: Equal Accuracy. SB: Source of Bias. ND: Number of Datasets. Y: Yes. N: No.

In-processing: Loss Re-Weighting. In-processing methods generally change or re-traine the ML model in a way that minimises bias. As an in-processing bias mitigation method, we utilize a popular and effective loss re-weighting approach as suggested by [Calders *et al.*, 2009]. This method works by penalizing mis-classified minority classes at a higher weighted proportion. More formally, the loss value for sample $\mathbf{x}_i$, i.e., $\mathcal{L}(\mathbf{x}_i, y_i; \theta)$, is multiplied by α_i, which is inversely proportional to the number of samples in group $s(\mathbf{x}_i)$. Thus, the algorithm is encouraged to pay more attention to the minority class which assists in bias mitigation.

Post-processing: Reject Option Classification (ROC). Post-processing algorithms typically modify the predicted labels to mitigate bias. We employ the ROC method suggested by [Kamiran *et al.*, 2012]. The intuition behind ROC is to re-classify the predictions of the minority group in favor of the minority group if predictions fall within a certain decision threshold region. If a sample $\mathbf{x}_i$ that falls in the "critical" region $1 - \tau \leq p(c|\mathbf{x}_i) \leq \tau$ where $0.5 \leq \tau \leq 1$, we classify $\mathbf{x}$ as c if $\mathbf{x}$ belongs to a minority group. Otherwise, i.e. when $p(c|\mathbf{x}_i) > \tau$, we accept the predicted output class c. In our experiments, we set $\tau = 0.6$ as suggested by Kamiran et al.

3.3 Prediction and Fairness Evaluation Measures

Throughout the paper, we use s_0 to denote the minority group. The minority group differs among datasets: Males are the minority in Depresjon and DVlog, whereas females are the minority in Psykose.

Prediction Measures. We use the commonly used measures, Accuracy ($\mathcal{M}_{Acc}$), Precision ($\mathcal{M}_P$), Recall ($\mathcal{M}_R$) and F1 ($\mathcal{M}_{F1}$), to evaluate prediction quality.

Fairness Measures. As fairness is a multifaceted challenge, existing fairness literature has highlighted the need for multiple metrics to characterize the ML bias present [Hort *et al.*, 2022; Cheong *et al.*, 2022b]. We highlight the most commonly used metrics [Hort *et al.*, 2022; Pessach and Shmueli, 2022] that we will adopt to evaluate our results and outline how each quantifies a different aspect of fairness:

- **Statistical Parity**, or demographic parity, is based purely on predicted outcome $\hat{Y}$ and independent of actual outcome Y:

$$\mathcal{M}_{SP} = \frac{P(\hat{Y}=1|s_0)}{P(\hat{Y}=1|s_1)}. \tag{1}$$

  According to this measure, in order for a classifier to be deemed fair, $P(\hat{Y}=1|s_1) = P(\hat{Y}=1|s_0)$.

- **Equal opportunity** states that both demographic groups s_0 and s_1 should have equal True Positive Rate (TPR).

$$\mathcal{M}_{EOpp} = \frac{P(\hat{Y}=1|Y=1,s_0)}{P(\hat{Y}=1|Y=1,s_1)}. \tag{2}$$

  According to this measure, in order for a classifier to be deemed fair, $P(\hat{Y}=1|Y=1,s_1) = P(\hat{Y}=1|Y=1,s_0)$.

- **Equalised odds** can be considered as a generalization of Equal Opportunity where the rates are not only equal for $Y=1$, but for all values of $Y \in \{1, ...k\}$, i.e.:

$$\mathcal{M}_{EOdd} = \frac{P(\hat{Y}=1|Y=i,s_0)}{P(\hat{Y}=1|Y=i,s_1)}. \tag{3}$$

  According to this measure, in order for a classifier to be deemed fair, $P(\hat{Y}=1|Y=i,s_1) = P(\hat{Y}=1|Y=i,s_0), \forall i \in \{1, ...k\}$.

- **Equal Accuracy** states that both subgroups s_0 and s_1 should have equal rates of accuracy.

$$\mathcal{M}_{EAcc} = \frac{\mathcal{M}_{ACC,s_0}}{\mathcal{M}_{ACC,s_1}}. \tag{4}$$

We have chosen the above fairness measures with careful deliberation. In addition to being the most commonly used metrics, the measures above cover all grounds ranging from the lenient Fairness through Unawareness (FTU) definitions to the stricter Equalised Odds framework [Hardt *et al.*, 2016]. As we see in Section 6, this can result in very different fairness outcomes which can have major implications on MHD resource allocation and patient prioritising. The potential consequences will be further discussed in Section 6. For all measures, we adopt the principle of disparate impact where a predictor or system is deemed to be fair when the performance or measure does not vary between the different demographic groups [Feldman *et al.*, 2015]. The ideal score of 1 indicates that both measures are equal for both groups and is henceforth considered "perfectly fair". For practical experimental purposes, we adopt the approach of existing literature which considers 0.80 and 1.20 as the acceptable lower and upper fairness bounds respectively [Zanna *et al.*, 2022]. Values which fall within this range are considered acceptably fair. Values which falls outside this range are considered unfair.

4 Experimental Setup

4.1 Datasets for MHD Analysis and Detection

In this study, we analyse two distinct types of datasets as summarized in Table 2. The dataset inclusion and exclusion

Dataset	MHD	Data Type	# Sbjcts	# Smpls	SA	Annotation	Class	Balanced
Depresjon	Depression	Motor Activity	55	693	Gender, Age	MADRS	Binary	Y
Psykose	Schizophrenia	Motor Activity	54	687	Gender, Age	DSM-IV, BPRS	Binary	N
D-Vlog	Depression	A + V	816	961	Gender	-	Binary	N

Table 2: Overview of the datasets included in this analysis. Abbreviations: A: Audio. V: Visual. MADRS: Montgomery–Åsberg Depression Rating Scale. DSM: Diagnostic and Statistical Manual of Mental Disorders. BPRS - Brief Psychiatric Rating Scale. SA: Sensitive Attributes. # Sbjcts: Number of subjects. # Smpls: Number of samples. Balanced: Balanced across gender. Y: Yes. N:No.

	Depresjon			Psykose			D-Vlog		
	Y_0	Y_1	T	Y_0	Y_1	T	Y_0	Y_1	T
M	150	160	310	NA	246	NA	140	182	322
F	252	131	383	NA	39	NA	266	373	639
T	402	291	693	402	285	687	405	555	961

Table 3: Dataset distribution and target attribute breakdown across datasets. Abbreviations: F: Female. M: Male. T: Total. Y_0: Control group. Y_1: MHD group. NA: Not available.

criteria is as follows. First, we restrict our analysis to publicly available datasets in order for our work to be comparable. Second, we focus on non-lab-based data, i.e. data that is collected in the real-world with fully naturalistic human behaviour. Third, we attempt to focus on patients who have been clinically diagnosed or are prescribed with MHD medication. We exclude datasets that chiefly rely on self-reported scores as well as datasets which were collected in a lab-based setting. Depresjon and Psykose consist of motor activity data whereas D-Vlog includes audio-visual recordings. With reference to Table 3, we see that each dataset has a different distribution breakdown across gender and MHD class. This is an important nuance which will impact our subsequent analysis and bias mitigation attempts.

Depresjon. Depresjon [Garcia-Ceja *et al.*, 2018] consists of motor activity data collected from 55 individuals. Data was recorded using an ActiGraph wristband. Out of the 55 participants, 23 individuals have been diagnosed with depression ($Y = 1$). This includes both unipolar and bipolar depression. This group was monitored for 291 days. 32 individuals do not have depression ($Y = 0$). This group was monitored for 402 days. The dataset contains 30 females and 25 males. This dataset contains another sensitive attribute age which is not used within our experiments. No standard train-test split protocol was provided by the dataset owners.

Psykose. Psykose [Jakobsen *et al.*, 2020b] consists of motor activity data collected from 54 individuals. Data was recorded using an ActiGraph wristband. Out of the 54 participants, 22 individuals have been diagnosed with schizophrenia ($Y = 1$). This group was monitored for 285 days. 32 individuals are in the control group ($Y = 0$). This group was monitored for 402 days. The dataset contains 23 females and 31 males. This dataset contains age as another sensitive attribute which is not used within our experiments. No standard train-test split protocol was provided by the dataset owners.

D-Vlog. D-Vlog [Yoon *et al.*, 2022] consists of vlog data collected from YouTube videos over a 13-months. The dataset contains 555 depressed and 406 non-depressed vlogs of 639 females and 322 males. The dataset owners provided a standard train-test split which we adhered to in our experiments.

4.2 ML Architecture

Depresjon and Psykose. As both Depresjon and Psykose are similar in size and collection method, we utilise the same experimental setup for both datasets. As the dataset owners have reported promising results with simpler ML models, we choose multi-layer perceptrons as an easy-to-use model for these datasets. For both, we utilise a 2-layer multi-layer perceptron with 3 neurons in the first layer with ReLU activation function and 2 neurons for the output layer. We utilise a Softmax function at the output layer to obtain class-wise prediction probabilities and minimize the Cross-entropy Loss.

D-Vlog. We utilise the *Depression Detector* architecture as proposed in [Yoon *et al.*, 2022]. It consists of two unimodal Transformer encoders and a multimodal Transformer encoder to incorporate the learned representations from both visual and acoustic features, followed by a global average pooling layer, a dropout layer, a a fully-connected layer and a Softmax layer. We minimize the Cross Entropy Loss. Readers can refer to [Yoon *et al.*, 2022] for further details on the architecture.

4.3 Implementation Details

Depresjon and Psykose. We train the networks using the Adam optimizer [Kingma and Ba, 2014] with a learning rate of 0.0001 and a batch size of 16 for 100 epochs (hyper-parameters are tuned using grid search). Depresjon and Psykose are both subject-dependent datasets. As each subject is tracked across several days, there will be several datapoints which belong to a single individual. As a result, we have opted to use a Leave-One-Subject-Out evaluation method which is aligned with existing literature [Garcia-Ceja *et al.*, 2018; Jakobsen *et al.*, 2020b; Jakobsen *et al.*, 2020a].

D-Vlog. We train the network using the Adam optimizer [Kingma and Ba, 2014] with a learning rate of 0.0002, a batch size of 32, a sequence length (t) of 596, for 50 epochs as stated in [Yoon *et al.*, 2022]. The dropout rate was not provided in the original work, and we empirically chose 0.1. We conduct two distinct sets of experiments for D-vlog. For the first set of experiments, for both the mixup and loss reweighting experiments, we assign weights to male and female samples as inversely proportional to their ratios in the training set. For the second of experiments, we assign weights to male and female samples directly proportional to their ratios in the training set, i.e. we generate twice as many males compared to females for the first experiments and twice as many females compared to males for the second experiments. Our rationale for the second set of experiments is to further examine whether the performance degradation on female samples can be overcome through assigning more weights to their loss or generating more female samples. For each of the experiments, we train and evaluate the networks three times with different seeds $(1, 2, 3)$ and report their average.

5 Results and Further Experiments

We discuss the results across all three datasets in relation to our research questions.

5.1 RQ1: What is the Bias Present?

Depresjon. With reference to Table 3, Depresjon is relatively balanced across gender with 310 males and 383 females. Males are considered the minority s_0 as we have less samples for them. Results for experiments on Depresjon are captured in Table 4. The baseline MLP model gives an accuracy of 0.72 (0.72) and an F1 score of 0.67 (0.71) which is consistent with existing work [Garcia-Ceja *et al.*, 2018] as highlighted within the parentheses. Two out of the four fairness measure used indicate that there is bias across gender. Both $\mathcal{M}_{SP}$ and $\mathcal{M}_{EOpp}$ give values within the 0.80 to 1.20 range. However, both $\mathcal{M}_{EOdd}$ and $\mathcal{M}_{EAcc}$ give a score of 0.47 and 0.72 respectively which indicate that the model is biased. **Answer to RQ 1:** Dataset bias is absent. Some metrics indicate that algorithmic bias may be present (F1.3[3]).

Psykose. Looking at Table 3, Psykose is imbalanced with 39 females in comparison with 246 males. For Psykose, females are the minority s_0. Our efforts are complicated by the fact that we do no have the sensitive attributes of those labeled with Y_0 which impacts our subsequent bias mitigation strategies. Looking at Table 4, the baseline MLP model gives a precision of 0.79 (0.80) and an F1 score of 0.79 (0.80) which is consistent with existing work [Jakobsen *et al.*, 2020b] as indicated within the parentheses. As we only have the sensitive attributes for Y_1, we are only able to evaluate bias across samples belonging to this category. Hence, the interpretation of the results may not be directly comparable to the other experiments. In addition, some of the fairness measures e.g. $\mathcal{M}_{EOdd}$ cannot be calculated as we are unable to calculate values such as the false positive rate according to sensitive attributes. Looking at the fairness of the baseline model, all values are close to 1, which implies that there is no algorithmic bias present. **Answer to RQ 1:** We note that dataset bias is present (F1.1) and algorithmic bias is absent (F1.2).

D-Vlog. With reference to Table 3, D-Vlog is imbalanced across gender as we have approximately twice as many samples for females compared to males. Here, males are considered the minority class s_0. The results for our experiments on DVlog are summarized in Table 4. The baseline method gives an accuracy of 0.64 and an F1 score of 0.69. For $\mathcal{M}_{SP}$, $\mathcal{M}_{EOpp}$ and $\mathcal{M}_{EAcc}$, scores are within the range 0.80-1.20, indicating a fair model, whereas $\mathcal{M}_{EOdd}$ gives 1.84 indicating strong algorithmic bias in favour of males. **Answer to RQ 1:** Results indicate that dataset bias is present (F1.1). Across algorithmic bias, different fairness measures give different outcomes. Three out of the four measures used indicated that the baseline model is acceptably fair.

5.2 RQ2: How Effective are the Bias Mitigation Strategies?

Depresjon. With reference to Table 4, we see that across most performance metrics, the pre- and in-processing results provide comparable outcomes. Overall, pre-processing methods seem to provide a slightly greater score improvement and outperform the other methods across accuracy, precision, recall and F1-measure. Across the fairness measures, for both the pre and in-processing methods, three out of the four fairness measures indicate that the model is fair. This is an improvement from the baseline. **Answer to RQ 2:** Different fairness measures report different outcomes (F2.1). The pre and in-processing methods provide more consistent bias mitigation compared to the post-processing method.

Psykose. Three of the fairness measures, $\mathcal{M}_{SP}$, $\mathcal{M}_{EOpp}$ and $\mathcal{M}_{EAcc}$, indicate that all models provide fairness results which are within the acceptable threshold 0.80-1.20. As we only have the sensitive attributes of the $Y = 1$ group, $\mathcal{M}_{SP}$ is inadvertently equal to $\mathcal{M}_{EOpp}$. Across all three mitigation strategies, the pre-processing method performs the best as it achieves a fairness score that is closest to 1. **Answer to RQ 2:** For Psykose, though all methods provide fairness scores within the fair range of $0.80 - 1.20$, both the pre- and in-processing methods give values that are closer to 1 whilst the post-processing mitigation strategy results in a slight bias towards the minority group. The pre-processing method provides the fairest result of all.

D-Vlog. Looking at the performance measures, the female-inclined pre-processing method and the female-inclined in-processing method yield the highest recall and F1-scores. The male-inclined pre-processing method have the edge in terms of accuracy. All results are relatively close and no single approach is best overall. Across fairness, different fairness metrics provide different results and even conflicting implications. To exemplify, if we were to evaluate fairness based on $\mathcal{M}_{SP}$, we would see that every single one of our approaches yields an improvement over the baseline, with the exception of the post-processing method ROC applied to the minority class males. Even more interestingly, the in-processing method yields a *perfectly fair* $\mathcal{M}_{SP} = 1.00$ score despite providing more weight to the majority female class. On the other hand, if we were to consider fairness solely based on $\mathcal{M}_{EOpp}$ or $\mathcal{M}_{EAcc}$, the baseline would be the fairest of all with $\mathcal{M}_{Opp} = 1.09$ and $\mathcal{M}_{EAcc} = 1.09$. Furthermore, one of the fairest approaches according to the $\mathcal{M}_{SP}$ metric, female-inclined pre-processing has $\mathcal{M}_{Opp} = 1.24$ and $\mathcal{M}_{EAcc} = 1.21$, which are both out of the $0.80 - 1.20$ range, indicating a strong bias in favour of males. **Answer to RQ 2:** The measures often contradict each other (F2.1). No method is consistently effective at mitigating bias and all seems to worsen the bias present (F2.2).

5.3 RQ3: What are the Factors that Hamper Gender Bias Mitigation in MHD?

Our results on D-Vlog highlight an important aspect which has yet to be considered within the fairness literature: In D-Vlog, males are considered the minority class. However,

[3] $Fx.y$ refers to the summary of findings in Table 5

	Depresjon				Psykose				D-Vlog						
Metrics	B	Pre	In	Post	B	Pre	In	Post	B	Pre (M)	Pre (F)	In (M)	In (F)	Post (M)	Post (F)
$\mathcal{M}_{Acc}$	0.72	0.77	0.76	0.74	0.77	0.75	0.78	0.78	0.64	0.66	0.65	0.65	0.65	0.64	0.64
$\mathcal{M}_{P}$	0.76	0.80	0.80	0.78	0.85	0.74	0.86	0.85	0.70	0.71	0.69	0.72	0.69	0.68	0.67
$\mathcal{M}_{R}$	0.67	0.74	0.73	0.71	0.73	0.74	0.74	0.74	0.69	0.71	0.73	0.65	0.73	0.71	0.73
$\mathcal{M}_{F1}$	0.67	0.74	0.73	0.71	0.73	0.74	0.75	0.74	0.69	0.70	0.71	0.68	0.71	0.69	0.70
$\mathcal{M}_{SP}$	$\underline{0.93}$	**1.22**	1.16	**1.30**	0.92	$\underline{0.97}$	0.95	1.20	0.92	1.02	1.01	0.95	$\underline{1.00}$	**1.24**	0.92
$\mathcal{M}_{EOpp}$	0.82	1.20	$\underline{1.11}$	1.20	0.92	$\underline{0.97}$	0.95	1.20	$\underline{1.09}$	**1.25**	**1.24**	1.18	1.19	**1.38**	1.16
$\mathcal{M}_{EOdd}$	**0.47**	$\underline{0.86}$	**0.67**	**0.61**	-	-	-	-	**1.84**	**2.13**	**2.09**	**2.45**	**1.87**	**$\underline{1.42}$**	**2.27**
$\mathcal{M}_{EAcc}$	**0.72**	$\underline{0.82}$	0.80	**0.79**	0.94	$\underline{1.03}$	1.04	1.19	$\underline{1.09}$	1.19	**1.21**	1.10	1.15	1.14	**1.21**
Consistency	2/4	3/4	3/4	1/4	3/3	3/3	3/3	3/3	3/4	2/4	1/4	3/4	2/4	1/4	2/4

Table 4: A comparison of the performance and fairness scores for the baseline model, the pre-processing data augmentation strategy, the in-processing loss reweighting and the post-processing ROC. Values in bold denote values that fall outside the acceptable range of 0.80-1.20. Underlined values highlight the best fairness score or values which are close to the ideal fair score of 1, i.e. within the 0.95-1.05 range. Abbreviations: B: Baseline. Pre: Pre-processing. In: In-processing. Post: Post-processing. M: Male. F: Female. **Consistency:** Highlights the number of fairness metrics which give values that fall within the acceptable fair threshold range. Additional results disaggragated across gender is available within the Appendix of the full paper[4].

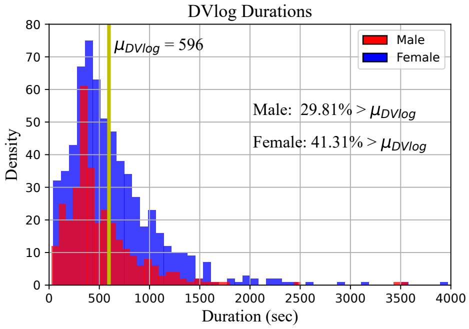

Figure 1: Male (red) and female (blue) vlog duration distribution in DVlog dataset. Yellow vertical line indicates the mean duration point. For females and males, 41.31% and 29.81% of the vlogs longer than the mean were truncated respectively.

despite having only approximately half the amount of samples as females, there is still strong bias *in favour of males* across most methods and fairness metrics. In addition, none of the bias mitigation methods seem effective as most perform poorly across many fairness metrics. In order to quantify the difficulty of depression detection for females vs. males, we measure the model's uncertainty in its predictions (i.e., predictive uncertainty). A model can be considered 'uncertain' in its predictions due to noise or lack of data. Therefore, a higher uncertainty value signifies a more difficult prediction problem. As there are twice as many female samples compared to males (Table 2), the expectation is to have a significantly lower uncertainty value for females. However, on average, we obtain similar uncertainty values of 0.645 and 0.649 for males and females, respectively (obtained using Deep Ensembles [Lakshminarayanan *et al.*, 2017] – see Appendix of the full paper[4] for details – F3.3). We identified two main potential factors for this counter-intuitive phenomena:

Factor 1: Data Pre-Processing. In [Yoon *et al.*, 2022], in order to achieve the same dimensionality for both visual and acoustic features, timesteps were either truncated or padded with zeros. $t = 596$ seconds was chosen as the truncation-padding boundary as it is the mean duration of all vlogs. This approach of truncating the endings inherently causes information loss for the vlogs that are longer than the mean duration whereas the shorter vlogs do not experience any loss as they are only padded with zeros. As illustrated in Figure 1, since female vlogs are significantly longer in duration compared to male vlogs on average, this approach might be a possible source of bias as it causes significantly more information loss for female vlogs compared to male vlogs (F3.1).

Factor 2: Gender Differences in Depression Manifestation and Diagnosis. Females and males tend to show different symptom profiles when depressed [Floyd, 1997; Barsky *et al.*, 2001; Ogrodniczuk and Oliffe, 2011]. Though existing research does not provide a conclusive indication of whether males or females are harder to diagnose, literature suggests that there are factors (e.g. physician bias, hormonal effects) which may make it more difficult to diagnose depression in females compared to males [Floyd, 1997; Barsky *et al.*, 2001] (F3.2).

Overall: Answer to RQ 3. Current bias mitigation techniques are inadequate to address the information loss due to data-preprocessing as well as the inherent gap in depression recognition difficulty across genders. This is evidenced by our experimental results. To mitigate bias, conventional mitigation methods dictate that we either balance the number of samples across class or apply e.g. a loss re-weighting proportional to the imbalance ratio. Despite doing so for the male minority class, we do not observe consistent performance and fairness improvement across metrics. To further illustrate our point, we replicated the same experiments in favour of females despite females being the majority class. As evidenced in Table 4, for all of the metrics other than $\mathcal{M}_{SP}$, this approach exacerbated the bias compared to the baseline.

6 Summary and Conclusion

Our findings indicate the presence of bias within both the datasets and models. For Psykose and D-Vlog, there is dataset

[4] https://www.repository.cam.ac.uk/handle/1810/349873

	Findings	Recommendations
RQ1	**F1.1:** Dataset and algorithmic bias are present. **F1.2:** Dataset bias, often, but does not always lead to biased outcomes and vice versa. **F1.3:** Algorithmic bias is present despite balanced samples.	**R1.1:** Ensure balanced dataset. **R1.2:** Train a separate model for each demographic group if samples are adequate. **R1.3:** Ensure appropriate experimentation methods, e.g. leave-one-subject-out for subject-dependent datasets. **R1.4:** Use dataset appropriate models e.g. simpler models for smaller datasets.
RQ2	**F2.1:** Fairness measures are not always consistent. **F2.2:** Mitigation worsens bias if it does not address root cause. **F2.3:** Pre- and in-processing methods are more consistently effective compared to the post-processing method.	**R2.1:** Use a range of fairness measures as they give different indication of bias mitigation effectiveness. **R2.2:** Rely on the stricter fairness measures for higher stake situations. **R2.3:** Employ appropriate mitigation strategies that address the root cause of bias. **R2.4:** Have a baseline result or expectation to compare against.
RQ3	**F3.1:** Inappropriate data pre-processing. **F3.2:** Gender difference in depression diagnosis. **F3.3:** Difficulty of the data impacts the algorithm's ability to learn appropriate representations.	**R3.1:** Conduct preliminary analysis to avoid removing important signals. **R3.2:** Work closely with clinical experts to devise better solutions that address gender differences in depression diagnosis. **R3.3:** Identify the source of difficulty between groups & if the discrepancy cannot be removed, employ methods that pay more attention to more difficult samples.

Table 5: Overview of our findings and recommendations on how to achieve machine learning fairness for mental health analysis.

bias in the form of dataset imbalance. However, there is still bias in the models even if we train the model on a more balanced dataset (e.g. Depresjon or the augmented balanced datasets). In addition to dataset imbalance, another key source of bias is the data pre-processing schema deployed. We see from our analysis with D-Vlog that inappropriate data pre-processing has potentially resulted in information loss for the female samples. This makes it harder for ML algorithms to detect depressed females which may have induced a bias against females compared to males. Another potential source of bias is the usage of inappropriate ML models or experimentation methods when analysing mental health data. For instance, MA data (e.g. Depresjon and Psykose) are very different from AV data (e.g. Depresjon): MA data is comparatively smaller than AV data. As such, using a simpler model e.g. a 2-layer MLP or a Linear SVM may be more appropriate for the former whereas a DNN would be appropriate for the latter. Using DNN for a small dataset may cause excessive memorisation which will likely fail to generalise and hence produce seemingly "biased" outcomes. In addition, for subject-dependent datasets (e.g. Depresjon and Psykose), it is important to ensure a subject-independent training procedure (e.g. Leave-One-Subject-Out). Otherwise, the model will produce overly optimistic results which will fail to generalise. Hence, biased results may be due to inappropriate model training and experimentation methods.

Our results indicate that bias can be mitigated if suitable mitigation methods were employed to address the root source of bias. For instance, looking at Psykose's results, if the problem is that of data imbalance, using bias mitigation strategies such as data augmentation and loss re-weighting work. However, if the source of bias is due to inappropriate data pre-processing (e.g. DVlog), regular bias mitigation may fail to work. Moreover, it is also important to take into account the potential shortcomings of bias mitigation strategies. For instance, the post-processing method ROC solely operates on the minority class by re-classifying them from $Y = 0$ to $Y = 1$ whenever they are in the critical region. This may risk inducing a bias in favour of the minority group and a bias against the majority group as evidenced by our results. To sum, bias cannot be sufficiently mitigated if the methods used do not adequately address the root cause of bias or if we do not account for the potential negative repercussions.

The findings above are critical as we see we can arrive at incorrect conclusions if the wrong fairness criteria or mitigation strategy is used. Taking D-Vlog for example, if we were solely to make decisions based on dataset bias, we may end up prioritising the minority group (i.e. males) when in fact the algorithm is biased against the majority group (i.e. females). In addition, some fairness measures are stricter than the others and some are true label dependent whereas some are not. It is important to have a good understanding of how these measures are computed to avoid arriving at misleading conclusions. For instance, the commonly used $\mathcal{M}_{EAcc}$ ascertains if a model is fair simply based on the ratio of its predictive accuracy between subgroups. Thus, it is possible to achieve a fairer score by reducing the performance accuracy on the majority group as exemplified in [Zanna *et al.*, 2022]. This is a sub-optimal solution and would not be ideal given the grave consequence associated with MHD disorder and analysis. We sum up our findings and outline a set of recommendations in Table 5 to address the foreseeable challenges in bias mitigation for MHD applications. The problem of bias in MHD analysis is multi-faceted and remains an open challenge. Overall, there are still many open-ended shortcomings that we have not been able to address in this paper. We hope that our results provide some much needed insights to the problem of bias in MHD analysis and will pave the way for several interesting future research directions [Cheong *et al.*, 2022a; Churamani *et al.*, 2023] to ensure that the technology developed for MHD analysis are fair and ethical for all.

Acknowledgements

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

References

[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.

[Barocas *et al.*, 2017] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NeurIPS Tutorial*, 1:2, 2017.

[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.

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[Calders *et al.*, 2009] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pages 13–18. IEEE, 2009.

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

[Cheong *et al.*, 2022b] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Counterfactual fairness for facial expression recognition. *2022 ECCV Workshop on Challenge on People Analysis (WCPA)*, 2022.

[Cheong *et al.*, 2023] 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*, pages 340–349, 2023.

[Churamani *et al.*, 2023] Nikhil Churamani, Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Towards causal replay for knowledge rehearsal in continual learning. *Proceedings of Machine Learning Research*, 208:1–8, 2023.

[DeVault *et al.*, 2014] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068, 2014.

[Feldman *et al.*, 2015] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[Floyd, 1997] Bonnie J. Floyd. Problems in accurate medical diagnosis of depression in female patients. *Social Science and Medicine*, 44(3):403–412, 1997.

[Garcia-Ceja *et al.*, 2018] Enrique Garcia-Ceja, Michael Riegler, Petter Jakobsen, Jim Tørresen, Tine Nordgreen, Ketil J Oedegaard, and Ole Bernt Fasmer. Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients. In *Proceedings of the 9th ACM multimedia systems conference*, pages 472–477, 2018.

[Gratch *et al.*, 2014] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The distress analysis interview corpus of human and computer interviews. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES, 2014.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[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.

[Huang *et al.*, 2020] Zhaocheng Huang, Julien Epps, Dale Joachim, Brian Stasak, James R Williamson, and Thomas F Quatieri. Domain adaptation for enhancing speech-based depression detection in natural environmental conditions using dilated cnns. In *INTERSPEECH*, pages 4561–4565, 2020.

[Jakobsen *et al.*, 2020a] Petter Jakobsen, Enrique Garcia-Ceja, Michael Riegler, Lena Antonsen Stabell, Tine Nordgreen, Jim Torresen, Ole Bernt Fasmer, and Ketil Joachim Oedegaard. Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls. *Plos one*, 15(8), 2020.

[Jakobsen *et al.*, 2020b] Petter Jakobsen, Enrique Garcia-Ceja, Lena Antonsen Stabell, Ketil Joachim Oedegaard, Jan Oystein Berle, Vajira Thambawita, Steven Alexander Hicks, Pål Halvorsen, Ole Bernt Fasmer, and Michael Alexander Riegler. Psykose: A motor activity database of patients with schizophrenia. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 303–308. IEEE, 2020.

[Kamiran *et al.*, 2012] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012.

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

[Lakshminarayanan *et al.*, 2017] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[Lee *et al.*, 2022] Joshua Lee, Yuheng Bu, Prasanna Sattigeri, Rameswar Panda, Gregory W Wornell, Leonid Karlinsky, and Rogerio Schmidt Feris. A maximal correlation framework for fair machine learning. *Entropy*, 24, 2022.

[Long *et al.*, 2022] Nannan Long, Yongxiang Lei, Lianhua Peng, Ping Xu, and Ping Mao. A scoping review on monitoring mental health using smart wearable devices. *Mathematical Biosciences and Engineering*, 19(8), 2022.

[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.

[Mathers and Loncar, 2006] Colin D Mathers and Dejan Loncar. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*, 3(11):e442, 2006.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[Mundnich *et al.*, 2020] Karel Mundnich, Brandon M. Booth, Michelle L'Hommedieu, Tiantian Feng, Benjamin Girault, Justin L'Hommedieu, Mackenzie Wildman, Sophia Skaaden, Amrutha Nadarajan, Jennifer L. Villatte, Tiago H. Falk, Kristina Lerman, Emilio Ferrara, and Shrikanth Narayanan. TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers. *Sci Data*, 7(354), 2020.

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

[Pagano *et al.*, 2023] Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1):15, 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.

[Scott *et al.*, 2017] Jan Scott, Greg Murray, Chantal Henry, Gunnar Morken, Elizabeth Scott, Jules Angst, Kathleen R Merikangas, and Ian B Hickie. Activation in bipolar disorders: a systematic review. *JAMA psychiatry*, 74(2):189–196, 2017.

[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 *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 99–109, 2020.

[Wang *et al.*, 2007] Philip S Wang, Sergio Aguilar-Gaxiola, Jordi Alonso, Matthias C Angermeyer, Guilherme Borges, Evelyn J Bromet, Ronny Bruffaerts, Giovanni De Girolamo, 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.

[Yamamoto *et al.*, 2020] Mao Yamamoto, Akihiro Takamiya, Kyosuke Sawada, Michitaka Yoshimura, Momoko Kitazawa, Kuo-ching Liang, Takanori Fujita, Masaru Mimura, and Taishiro Kishimoto. Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PloS one*, 15(9):e0238726, 2020.

[Yau *et al.*, 2022] Joanna C Yau, Benjamin Girault, Tiantian Feng, Karel Mundnich, Amrutha Nadarajan, Brandon M Booth, Emilio Ferrara, Kristina Lerman, Eric Hsieh, and Shrikanth Narayanan. Tiles-2019: A longitudinal physiologic and behavioral data set of medical residents in an intensive care unit. *Scientific Data*, 9(1):536, 2022.

[Yoon *et al.*, 2022] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. D-vlog: Multimodal vlog dataset for depression detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11), 2022.

[Zanna *et al.*, 2022] Khadija Zanna, Kusha Sridhar, Han Yu, and Akane Sano. Bias reducing multitask learning on mental health prediction. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2022.

[Zhang *et al.*, 2017] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2017.

[Zhang *et al.*, 2020] Ziheng Zhang, Weizhe Lin, Mingyu Liu, and Marwa Mahmoud. Multimodal deep learning framework for mental disorder recognition. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 344–350. IEEE, 2020.