U-Fair: Uncertainty-based Multimodal Multitask Learning for Fairer Depression Detection

Jiaee Cheong* jc2208@cam.ac.uk
University of Cambridge & the Alan Turing Institute, United Kingdom.

Aditya Bangar adityavb21@iitk.ac.in
Indian Institute of Technology, Kanpur, India.

Sinan Kalkan skalkan@metu.edu.tr
Dept. of Comp. Engineering and ROMER Center for Robotics and AI, Middle East Technical University (METU), Turkiye.

Hatice Gunes hg410@cam.ac.uk
University of Cambridge, United Kingdom.

Abstract

Machine learning bias in mental health is becoming an increasingly pertinent challenge. Despite promising efforts indicating that multitask approaches often work better than unitask approaches, there is minimal work investigating the impact of multitask learning on performance and fairness in depression detection nor leveraged it to achieve fairer prediction outcomes. In this work, we undertake a systematic investigation of using a multitask approach to improve performance and fairness for depression detection. We propose a novel gender-based task-reweighting method using uncertainty grounded in how the PHQ-8 questionnaire is structured. Our results indicate that, although a multitask approach improves performance and fairness compared to a unitask approach, the results are not always consistent and we see evidence of negative transfer and a reduction in the Pareto frontier, which is concerning given the high-stake healthcare setting. Our proposed approach of gender-based reweighting with uncertainty improves performance and fairness and alleviates both challenges to a certain extent. Our findings on each PHQ-8 subitem *task difficulty* are also in agreement with the largest study conducted on the PHQ-8 subitem *discrimination capacity*, thus providing the very first tangible evidence linking ML findings with large-scale empirical population studies conducted on the PHQ-8.

* This work was undertaken while Jiaee Cheong was a visiting PhD student at METU.

1. Introduction

Mental health disorders (MHDs) are becoming increasingly prevalent world-wide (Wang et al., 2007) 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; Xu et al., 2020; Cheong et al., 2021, 2022, 2023a). Given the high stakes involved in MHD analysis and prediction, it is crucial to investigate and mitigate the ML biases present. A substantial amount of literature has indicated that adopting a multitask learning (MTL) approach towards depression detection demonstrated significant improvement across classification-based performances (Li et al., 2022; Zhang et al., 2020). Most of the existing work rely on the standardised and commonly used eight-item Patient Health Questionnaire depression scale (PHQ-8) (Kroenke et al., 2009) to obtain the ground-truth labels on whether a subject is considered depressed. A crucial observation is that in order to arrive at the final classification (depressed vs non-depressed), a clinician has to first obtain the scores of each of the PHQ-8 sub-criterion and then sum them up to arrive at the final binary classification (depressed vs non-depressed). Details on how the final score is derived from the PHQ-8 questionnaire can be found in Section 3.1.

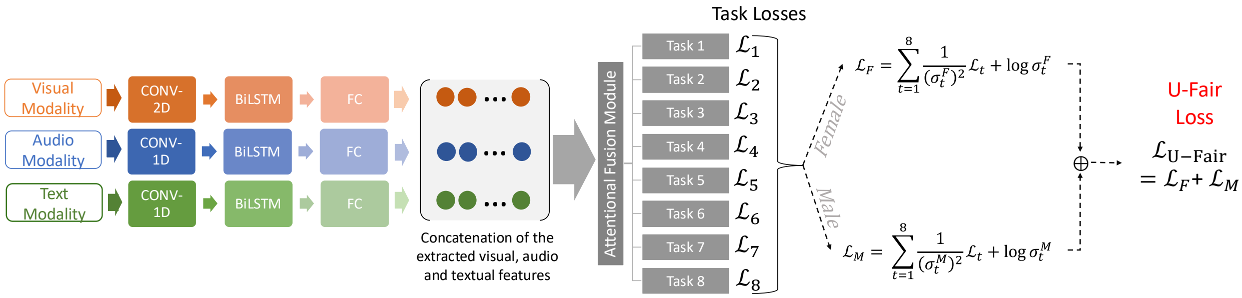


Figure 1: Our proposed method is rooted in the observation that each gender may have different PHQ-8 distributions and different levels of task difficulty across the t_1 to t_8 tasks. We propose accounting for this gender difference in PHQ-8 distributions via U-Fair.

Moreover, each gender may display different PHQ-8 task distribution which may results in different PHQ-8 distribution and variance. Although investigation on the relationship between the PHQ-8 and gender has been explored in other fields such as psychiatry (Thibodeau and Asmundson, 2014; Vetter et al., 2013; Leung et al., 2020), this has not been investigated nor accounted for in any of the existing ML for depression detection methods. Moreover, existing work has demonstrated the risk of a fairness-accuracy trade-off (Pleiss et al., 2017) and how mainstream MTL objectives might not correlate well with fairness goals (Wang et al., 2021b). No work has investigated how a MTL approach impacts performance across fairness for the task of *depression detection*.

In addition, prior works have demonstrated the intricate relationship between ML bias and uncertainty (Mehta et al., 2023; Tahir et al., 2023; Kaiser et al., 2022; Kuzucu et al., 2024). Uncertainty broadly refers to confidence in predictions. Within ML research, two types of uncertainty are commonly studied: data (or aleatoric) and model (or epistemic) uncertainties. Aleatoric uncertainty refers to the inherent randomness in the experimental outcome whereas epistemic uncertainty can be attributed to a lack of knowledge (Gal, 2016). A particularly relevant theme is that ML bias can be attributed to uncertainty in some models or datasets (Kuzucu et al., 2024) and that taking into account uncertainty as a bias mitigation strategy has proven effective (Tahir et al., 2023; Kaiser et al., 2022). A growing body of literature has also highlighted the importance of taking uncertainty into account within a range of tasks (Naik et al., 2024; Han et al., 2024; Baltaci et al., 2023; Cetinkaya et al., 2024) and healthcare settings (Grote and Keeling, 2022; Chua et al., 2023). Motivated by the above and the importance of a clinician-centred approach towards building relevant ML for healthcare solutions, we propose a novel method, U-Fair, which accounts for the gender difference in PHQ-8 distribution and leverages on uncertainty as a MTL task reweighing mechanism to achieve better gender fairness for depression detection. Our key contributions are as follow:

- We conduct the first analysis to investigate how MTL impacts fairness in *depression detection* by using each PHQ-8 subcriterion as a task. We show that a simplistic baseline MTL approach runs the risk of incurring negative transfer and may not improve on the Pareto frontier. A Pareto frontier can be understood as the set of optimal solutions that strike a balance among different objectives such that there is no better solution beyond the frontier.
- We propose a simple yet effective approach that leverages gender-based aleatoric uncertainty which improves the fairness-accuracy trade-off and alleviates the negative transfer phenomena and improves on the Pareto-frontier beyond a unitask method.
- We provide the very first results connecting the empirical results obtained via ML experiments with the *empirical findings* obtained via the *largest study conducted on the PHQ-8*. Interestingly, our results highlight the intrinsic relationship between task difficulty as quantified by aleatoric uncertainty and the discrimination capacity of each item of the PHQ-8 subcriterion.

2. Literature Review

Gender difference in depression manifestation has long been studied and recognised within fields such as

Study	Problem	Multimodal	Approach		Evaluation		
			Uncertainty	NFM Measures	PF	NT	ND
Zanna et al. (2022)	Anxiety	×	✓	2	×	×	1
Li et al. (2023a)	Healthcare prediction	×	×	2	×	×	1
Li et al. (2023b)	Organ transplant	×	×	2	×	×	1
Ban and Ji (2024)	Resource allocation	×	×	2	✓	×	3
Li et al. (2024)	Risk factor prediction	×	×	2	×	×	1
U-Fair (**Ours**)	Depression detection	✓(AVT)	✓	4	✓	×	2

Table 1: Comparative Summary with existing MTL Fairness studies. Abbreviations (sorted): A: Audio. NFM: Number of Fairness Measures. NT: Negative Transfers. ND: Number of Datasets. PF: Pareto Frontier. T: Text. V: Visual.

medicine (Barsky et al., 2001) and psychology (Hall et al., 2022). Anecdotal evidence has also often supported this view. Literature indicates that females and males tend to show different behavioural symptoms when depressed (Barsky et al., 2001; Ogrodniczuk and Oliffe, 2011). 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; Jansz et al., 2000).

Recent works have indicated that ML bias is present within mental health analysis (Zanna et al., 2022; Bailey and Plumbley, 2021; Cheong et al., 2024a,b; Cameron et al., 2024; Spitale et al., 2024). Zanna et al. (2022) proposed an uncertainty-based 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. Cheong et al. (2023b, 2024a) demonstrated that bias exists in existing mental health algorithms and datasets and subsequently proposed a causal multimodal method to mitigate the bias present.

MTL is noted to be particularly effective when the tasks are correlated (Zhang and Yang, 2021). Existing works using MTL for depression detection has proven fruitful. Ghosh et al. (2022) adopted a MTL approach by training the network to detect three closely related tasks: depression, sentiment and emotion. Wang et al. (2022) proposed a MTL approach using word vectors and statistical features. Li et al. (2022) implemented a similar strategy by using depression and three other auxiliary tasks: topic, emotion and dialog act. Gupta et al. (2023) adopted a multimodal, multiview and MTL approach where the subtasks are depression, sentiment and emotion.

In concurrence, although MTL has proven to be effective at improving *fairness* for other tasks such as healthcare predictive modelling (Li et al., 2023a), organ transplantation (Li et al., 2023b) and resource allocation (Ban and Ji, 2024), this approach has been underexplored for the task of depression detection.

Comparative Summary: Our work differs from the above in the following ways (see Table 1). First, our work is the first to leverage an MTL approach to improve gender fairness in *depression detection*. Second, we utilise an MTL approach where each task corresponds to each of the PHQ-8 subtasks (Kroenke et al., 2009) in order to exploit gender-specific differences in PHQ-8 distribution to achieve greater fairness. Third, we propose a novel gender-based uncertainty MTL loss reweighing to achieve fairer performance across gender for

3. Methodology: U-Fair

In this section, we introduce U-Fair, which uses aleatoric-uncertainties for demographic groups to reweight their losses.

3.1. PHQ-8 Details

One of the standardised and most commonly used depression evaluation method is the PHQ-8 developed by Kroenke et al. (2009). In order to arrive at the final classification (depressed vs non-depressed), the protocol is to first obtain the subscores of each of the PHQ-8 subitem as follows:

- PHQ-1: Little interest or pleasure in doing things,
- PHQ-2: Feeling down, depressed, or hopeless,
- PHQ-3: Trouble falling or staying asleep, or sleeping too much,
- PHQ-4: Feeling tired or having little energy,
- PHQ-5: Poor appetite or overeating,

- PHQ-6: Feeling that you are a failure,
- PHQ-7: Trouble concentrating on things,
- PHQ-8: Moving or speaking so slowly that other people could have noticed.

Each PHQ-8 subcategory is scored between 0 to 3, with the final PHQ-8 total score (TS) ranging between 0 to 24. The PHQ-8 binary outcome is obtained via thresholding. A PHQ-8 TS of ≥ 10 belongs to the depressed class ($Y = 1$) whereas TS ≤ 10 belongs to the non-depressed class ($Y = 0$).

Most existing works focused on predicting the final binary class (Y) (Zheng et al., 2023; Bailey and Plumbley, 2021). Some focused on predicting the PHQ-8 total score and further obtained the binary classification via thresholding according to the formal definition (Williamson et al., 2016; Gong and Poellabauer, 2017). Others adopted a bimodal setup with 2 different output heads to predict the PHQ-8 total score as well as the PHQ-8 binary outcome (Valstar et al., 2016; Al Hanai et al., 2018).

3.2. Problem Formulation

In our work, *in alignment with how the PHQ-8 works*, we adopt the approach where each PHQ-8 subcategory is treated as a task t. The architecture is adapted from Wei et al. (2022). For each individual $i \in I$, we have 8 different prediction heads for each of the tasks, $[t_1, ..., t_8] \in T$, to predict the score $y_t^i \in \{0, 1, 2, 3\}$ for each task or sub PHQ-8 category. The ground-truth labels for each task t is transformed into a Gaussian-based soft-distribution $p_t(x)$, as soft labels provide more information for the model to learn from (Yuan et al., 2024). x is the input feature provided to the model. Each of the classification heads are trained to predict the probability $q_t(x)$ of the 4 different score classes $y_t^i \in \{0, 1, 2, 3\}$. During inference, the final $y_t^i \in \{0, 1, 2, 3\}$ is obtained by selecting the score with the maximum probability. The PHQ-8 Total Score TS and final PHQ-8 binary classification $\hat{Y}$ for each individual $i \in I$ are derived from each subtask via:

$$TS = \sum_{t=1}^{8} y_t, \tag{1}$$

and

$$\hat{Y} = 1 \text{ if } TS \geq 10, \text{ else } \hat{Y} = 0. \tag{2}$$

$\hat{Y}$ thus denotes the final predicted class calculated based on the summation of y_t. We study the problem of fairness in depression detection, where the goal is to predict a correct outcome $y^i \in Y$ from input $\mathbf{x}^i \in X$ based on the available dataset D for individual $i \in I$. In our setup, $Y = 1$ denotes the PHQ-8 binary outcome corresponding to "depressed" and $Y = 0$ denotes otherwise. Only gender was provided as a sensitive attribute S.

3.3. Unitask Approach

For our single task approach, we use a Kullback-Leibler (KL) Divergence loss as follows:

$$\mathcal{L}_{STL} = \sum_{t \in T} p_t(x) \log \left(\frac{p_t(x)}{q_t(x)} \right). \tag{3}$$

$p_t(x)$ is the soft ground-truth label for each task t and $q_t(x)$ is the probability of the 4 different score classes $y_t \in \{0, 1, 2, 3\}$ as explained in Section 3.1.

3.4. Multitask Approach

For our baseline multitask approach, we extend the loss function in Equation 3 to arrive at the following generalisation:

$$\mathcal{L}_{MTL} = \sum_{t \in T} w_t \mathcal{L}_t. \tag{4}$$

$\mathcal{L}_t$ is the single task loss $\mathcal{L}_{STL}$ for each t as defined in Equation 3. We set $w_t = 1$ in our experiments.

3.5. Baseline Approach

To compare between the generic multitask approach in Equation 4 and an *uncertainty-based* loss reweighting approach, we use the commonly used multitask learning method by Kendall et al. (2018) as the baseline uncertainty weighting (UW) appraoch. The uncertainty MTL loss across tasks is thus defined by:

$$\mathcal{L}_{UW} = \sum_{t \in T} \left(\frac{1}{\sigma_t^2} \mathcal{L}_t + \log \sigma_t \right), \tag{5}$$

where $\mathcal{L}_t$ is the single task loss as defined in Equation 3. σ_t is the learned weight of loss for each task t and can be interpreted as the aleatoric uncertainty of the task. A task with a higher aleatoric uncertainty will thus lead to a larger single task loss $\mathcal{L}_t$ thus preventing the trained model to optimise on that task. The higher σ_t, the more difficult the task t. $\log \sigma_t$ penalizes the model from arbitrarily increasing σ_t to reduce the overall loss (Kendall et al., 2018).

3.6. Proposed Loss: U-Fair

To achieve fairness across the different PHQ-8 tasks, we propose the idea of task prioritisation based on the model's task-specific uncertainty weightings. Motivated by literature highlighting the existence of gender difference in depression manifestation (Barsky et al., 2001), we propose a novel gender based uncertainty reweighting approach and introduce U-Fair Loss which is defined as follows:

$$\mathcal{L}_{U-Fair} = \frac{1}{|S|} \sum_{s \in S} \sum_{t \in T} \left(\frac{1}{(\sigma_t^s)^2} \mathcal{L}_t^s + \log \sigma_t^s \right). \quad (6)$$

For our setting, s can either be male s_1 or female s_0 and $|S| = 2$. Thus, we have the uncertainty weighted task loss for each gender, and sum them up to arrive at our proposed loss function $\mathcal{L}_{MMFair}$.

This methodology has two key benefits. First, fairness is optimised implicitly as we train the model to optimise for task-wise prediction accuracy. As a result, by not constraining the loss function to blindly optimise for fairness at the cost of utility or accuracy, we hope to reduce the negative impact on fairness and improve the Pareto frontier with a constraint-based fairness optimisation approach (Wang et al., 2021b). Second, as highlighted by literature in psychiatry (Leung et al., 2020; Thibodeau and Asmundson, 2014), each task has different levels of uncertainty in relation to each gender. By adopting a gender based uncertainty loss-reweighting approach, we account for such uncertainty in a principled manner, thus encouraging the network to learn a better *joint-representation* due to the MTL and the gender-base aleatoric uncertainty loss reweighing approach.

4. Experimental Setup

We outline the implementation details and evaluation measures here. We use DAIC-WOZ (Valstar et al., 2016) and E-DAIC (Ringeval et al., 2019) for our experiments. Further details about the datasets can be found within the Appendix.

4.1. Implementation Details

We adopt an attention-based multimodal architecture adapted from Wei et al. (2022) featuring late fusion of extracted representations from the three different modalities (audio, visual, textual) as illustrated in Figure 1. The extracted features from each modality are concatenated in parallel to form a feature map as input to the subsequent fusion layer. We have 8 different attention fusion layers connected to the 8 output heads which corresponds to the t_1 to t_8 tasks. For all loss functions, we train the models with the Adam optimizer (Kingma and Ba, 2014) at a learning rate of 0.0002 and a batch size of 32. We train the network for a maximum of 150 epochs and apply early stopping.

4.2. Evaluation Measures

To evaluate performance, we use F1, recall, precision, accuracy and unweighted average recall (UAR) in accordance with existing work (Cheong et al., 2023c). To evaluate group fairness, we use the most commonly-used definitions according to (Hort et al., 2022). s_1 denotes the male majority group and s_0 denotes the female minority group for both datasets.

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

$$\mathcal{M}_{SP} = \frac{P(\hat{Y} = 1|s_0)}{P(\hat{Y} = 1|s_1)}. \quad (7)$$

  According to $\mathcal{M}_{SP}$, 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).

$$\mathcal{M}_{EOpp} = \frac{P(\hat{Y} = 1|Y = 1, s_0)}{P(\hat{Y} = 1|Y = 1, s_1)}. \quad (8)$$

  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.:

$$\mathcal{M}_{EOdd} = \frac{P(\hat{Y} = 1|Y = i, s_0)}{P(\hat{Y} = 1|Y = i, s_1)}. \quad (9)$$

  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.

$$\mathcal{M}_{EAcc} = \frac{\mathcal{M}_{ACC,s_0}}{\mathcal{M}_{ACC,s_1}}. \tag{10}$$

For all fairness measures, the ideal score of 1 thus indicates that both measures are equal for s_0 and s_1 and is thus considered "perfectly fair". We adopt the approach of existing work which considers 0.80 and 1.20 as the lower and upper fairness bounds respectively (Zanna et al., 2022). Values closer to 1 are fairer, values further form 1 are less fair. For all binary classification, the "default" threshold of 0.5 is used in alignment with existing works (Wei et al., 2022; Zheng et al., 2023).

5. Results

For both datasets, we normalise the fairness results to facilitate visualisation in Figures 2 and 3.

5.1. Uni vs Multitask

For DAIC-WOZ (DW), we see from Table 2, we find that a multitask approach generally improves results compared to a unitask approach (Section 3.3). The baseline loss re-weighting approach from Equation 5 managed to further improve *performance*. For example, we see from Table 2 that the overall classification accuracy improved from 0.70 within a vanilla MTL approach to 0.82 using the baseline uncertainty-based task reweighing approach.

However, this observation is not consistent for E-DAIC (ED). With reference to Table 3, a unitask approach seems to perform better. We see evidence of *negative transfer*, i.e. the phenomena where learning multiple tasks concurrently result in lower performance than a unitask approach. We hypothesise that this is because ED is a more challenging dataset. When adopting a multitask approach, the model completely relies on the easier tasks thus negatively impacting the learning of the other tasks.

Moreover, performance improvement seems to come at a cost. This may be due to the fairness-accuracy trade-off (Wang et al., 2021b). For instance in DW, we see that the fairness scores $\mathcal{M}_{SP}$, $\mathcal{M}_{EOpp}$, $\mathcal{M}_{Odd}$ and $\mathcal{M}_{Acc}$ reduced from 0.86, 0.78, 0.94 and 0.76 to 1.23, 1.70, 1.31 and 1.25 respectively. This is consistent with the analysis across the Pareto frontier depicted in Figures 2 and 3.

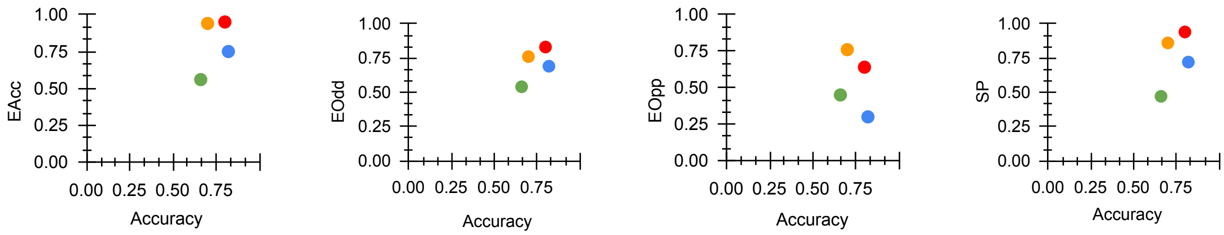
	Measure	Approach	Binary Outcome
Performance Measures	Acc	Unitask	0.66
		Multitask	0.70
		Baseline UW	**0.82**
		U-Fair **(Ours)**	0.80
	F1	Unitask	0.47
		Multitask	0.53
		Baseline UW	0.29
		U-Fair **(Ours)**	**0.54**
	Precision	Unitask	0.44
		Multitask	0.50
		Baseline UW	0.22
		U-Fair **(Ours)**	**0.56**
	Recall	Unitask	0.50
		Multitask	0.57
		Baseline UW	0.43
		U-Fair **(Ours)**	**0.60**
	UAR	Unitask	0.60
		Multitask	**0.65**
		Baseline UW	0.64
		U-Fair **(Ours)**	0.63
Fairness Measures	$\mathcal{M}_{SP}$	Unitask	0.47
		Multitask	0.86
		Baseline UW	1.23
		U-Fair **(Ours)**	**1.06**
	$\mathcal{M}_{EOpp}$	Unitask	0.45
		Multitask	**0.78**
		Baseline UW	1.70
		U-Fair **(Ours)**	1.46
	$\mathcal{M}_{EOdd}$	Unitask	0.54
		Multitask	0.76
		Baseline UW	1.31
		U-Fair **(Ours)**	**1.17**
	$\mathcal{M}_{EAcc}$	Unitask	1.44
		Multitask	0.94
		Baseline UW	1.25
		U-Fair **(Ours)**	**0.95**

Table 2: Results for **DAIC-WOZ**. Full table results for DW, Table 6, is available within the Appendix. Best values are highlighted in **bold**.

5.2. Uncertainty & the Pareto Frontier

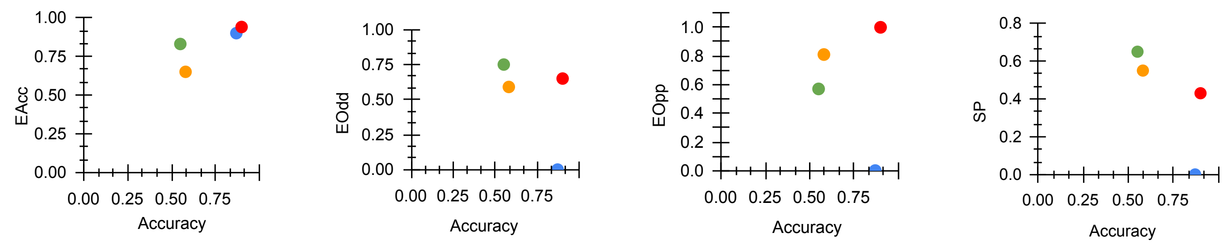
Our proposed loss reweighting approach seems to address the negative transfer and Pareto frontier challenges. Although accuracy dropped slightly from 0.82 to 0.80, fairness largely improved compared to the baseline UW approach (Equation 5). We see from Table 2 that fairness improved across $\mathcal{M}_{SP}$, $\mathcal{M}_{EOpp}$, $\mathcal{M}_{EOdd}$ and $\mathcal{M}_{Acc}$ from 1.23, 1.70, 1.31, 1.25 to 1.06, 1.46, 1.17 and 0.95 for DW.

For ED, the baseline UW which adopts a task based difficulty reweighting mechanism seems to somewhat mitigate the task-based negative transfer which



(a) $\mathcal{M}_{EAcc}$ vs Acc (b) $\mathcal{M}_{EOdd}$ vs Acc (c) $\mathcal{M}_{EOpp}$ vs Acc (d) $\mathcal{M}_{SP}$ vs Acc

Figure 2: Fairness-Accuracy Pareto Frontier across the **DAIC-WOZ** results. Upper right indicates better Pareto optimality, i.e. better fairness-accuracy trade-off. **Orange**: Unitask. **Green**: Multitask. **Blue**: Multitask UW. **Red**: U-Fair. Abbreviations: Acc: accuracy.



(a) $\mathcal{M}_{EAcc}$ vs Acc (b) $\mathcal{M}_{EOdd}$ vs Acc (c) $\mathcal{M}_{EOpp}$ vs Acc (d) $\mathcal{M}_{SP}$ vs Acc

Figure 3: Fairness-Accuracy Pareto Frontier across the **E-DAIC** results. Upper right indicates better Pareto optimality, i.e. better fairness-accuracy trade-off. **Orange**: Unitask. **Green**: Multitask. **Blue**: Multitask UW. **Red**: U-Fair. Abbreviations: Acc: accuracy.

improves the unitask performance but not overall performance nor fairness measures. Our proposed method which takes into account the gender difference may have somewhat addressed this task-based negative transfer. In concurrence, U-Fair also addressed the initial bias present. We see from Table 3 that fairness improved across all fairness measures. The scores improved from 3.86, 2.31, 8.21, 0.92 to 1.67, 1.00, 5.00 and 0.94 across $\mathcal{M}_{SP}$, $\mathcal{M}_{EOpp}$, $\mathcal{M}_{EOdd}$ and $\mathcal{M}_{Acc}$.

The Pareto frontier across all four measures illustrated in Figures 2 and 3 demonstrated that our proposed method generally provides better accuracy-fairness trade-off across most fairness measures for both datasets. With reference to Figure 2, we see that U-Fair, generally provides a slightly better Pareto optimality compared to other methods. This improvement in the Pareto frontier is especially pronounced for Figure 3(c). The difference in the Pareto frontier between our proposed method and other compared methods is greater in ED (Fig 3), the more challenging dataset, compared to that in DW (Fig 2).

For DW, with reference to Figures $4(a)$ and $4(b)$, we see that there is a difference in task difficulty. Task 4 and 6 is easier for females whereas task 7 is easier for males. For ED, with reference to Figures $4(c)$, $4(d)$ and Table 5, Task 4 seems to be easier for females whereas task 7 seems easier for males. Thus, adopting a gender-based uncertainty reweighting approach might have ensured that the tasks are more appropriately weighed leading towards better performance for both genders whilst mitigating the negative transfer and Pareto frontier challenges.

5.3. Task Difficulty & Discrimination Capacity

A particularly relevant and exciting finding is that each PHQ-8 subitem's task difficulty agree with its *discrimination capacity* as evidenced by the rigorous study conducted by de la Torre et al. (2023). This largest study to date assessed the internal structure, reliability and cross-country validity of the PHQ-8 for the assessment of depressive symptoms. *Discrimination capacity* is defined as the ability of item to distinguish whether a person is depressed or not.

With reference to Table 5, it is noteworthy that the task difficulty captured by $\frac{1}{\sigma^2}$ in our experiments

	Measure	Approach	Binary Outcome
Performance Measures	Acc	Unitask	0.55
		Multitask	0.58
		Baseline UW	0.87
		U-Fair **(Ours)**	**0.90**
	F1	Unitask	**0.51**
		Multitask	0.45
		Baseline UW	0.27
		U-Fair **(Ours)**	0.45
	Precision	Unitask	0.36
		Multitask	0.32
		Baseline UW	0.28
		U-Fair **(Ours)**	**0.46**
	Recall	Unitask	**0.87**
		Multitask	0.80
		Baseline UW	0.26
		U-Fair **(Ours)**	0.45
	UAR	Unitask	0.63
		Multitask	0.67
		Baseline UW	0.60
		U-Fair **(Ours)**	**0.70**
Fairness Measures	$\mathcal{M}_{SP}$	Unitask	0.65
		Multitask	**1.25**
		Baseline UW	3.86
		U-Fair **(Ours)**	1.67
	$\mathcal{M}_{EOpp}$	Unitask	0.57
		Multitask	0.81
		Baseline UW	2.31
		U-Fair **(Ours)**	**1.00**
	$\mathcal{M}_{EOdd}$	Unitask	**0.75**
		Multitask	1.41
		Baseline UW	8.21
		U-Fair **(Ours)**	5.00
	$\mathcal{M}_{EAcc}$	Unitask	0.83
		Multitask	0.65
		Baseline UW	0.92
		U-Fair **(Ours)**	**0.94**

Table 3: Results for **E-DAIC**. Full table results for ED, Table 7, is available within the Appendix. Best values are highlighted in **bold**.

Method	Prec.	Rec.	F1
Ma et al. (2016)	0.35	**1.00**	0.52
Song et al. (2018)	0.32	0.86	0.46
Williamson et al. (2016)	-	-	0.53
Song et al. (2018)	0.60	0.43	0.50
U-Fair (Ours)	0.52	0.60	**0.57**

Table 4: Comparison with other models which used extracted features for DAIC-WOZ. Best results highlighted in **bold**.

corresponds to the discrimination capacity (DC) of each task. The higher σ_t, the more difficult the task t. In other words, the lower the value of $\frac{1}{\sigma^2}$, the more difficult the task. For instance, in their study, PHQ-1, 2 and 6 were the items that has the greatest ability to discriminate whether a person is depressed. This is in alignment with our results where PHQ-1,2 and 8 are easier across both datasets. PHQ-3 and PHQ-5 are the least discriminatory or more difficult tasks as evidenced by the values highlighted in red.

		$\frac{1}{\sigma^2}$			
	DC	DW-F	DW-M	ED-F	ED-M
PHQ-1	**3.06**	**1.50**	**1.41**	**1.69**	**1.69**
PHQ-2	**3.42**	**1.41**	**1.47**	**1.38**	**1.41**
PHQ-3	**1.91**	**0.62**	**0.64**	**0.51**	**0.58**
PHQ-4	2.67	0.82	0.68	0.91	0.60
PHQ-5	**2.22**	**0.61**	0.69	**0.51**	**0.58**
PHQ-6	**2.86**	0.73	**0.59**	0.63	0.60
PHQ-7	2.55	0.75	0.80	0.61	0.89
PHQ-8	2.43	**1.58**	**1.72**	**1.69**	**1.70**

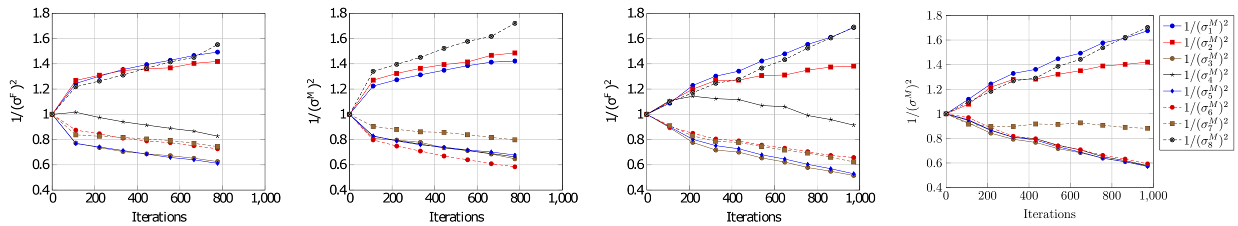
Table 5: Discrimination capacity (DC) vs $\frac{1}{\sigma^2}$. Lower $\frac{1}{\sigma^2}$ values implies higher task difficulty. **Green**: top 3 highest scores. **Red**: bottom 2 lowest scores. Our results are in harmony with the largest and most comprehensive study on the PHQ-8 conducted by de la Torre et al. (2023). DW: DAIC-WOZ. ED: E-DAIC. F: Female. M: Male.

6. Discussion and Conclusion

Our experiments unearthed several interesting insights. First, overall, there are certain gender-based differences across the different PHQ-8 distribution labels as evidenced in Figure 4. In addition, each task have slightly different degree of task uncertainty across gender. This may be due to a gender difference in PHQ-8 questionnaire profiling or inadequate data curation. Thus, employing a gender-aware approach may be a viable method to improve fairness and accuracy for depression detection.

Second, though a multitask approach generally performs better than a unitask approach, this comes with several caveats. We see from Table 5 that each task has a different level of difficulty. Naively using all tasks may worsen performance and fairness compared to a unitask approach if we do not account for task-based uncertainty. This is in agreement with existing literature which indicates that there can be a mix of positive and negative transfers across tasks (Li et al., 2023c) and tasks have to be related for performance to improve (Wang et al., 2021a).

Third, understanding, analysing and improving upon the fairness-accuracy Pareto frontier within



(*a*) DAIC-WOZ:**Female** (*b*) DAIC-WOZ: **Male** (*c*) E-DAIC: **Female** (*d*) E-DAIC: **Male**

Figure 4: Task-based weightings for both gender and datasets.

the task of depression requires a nuanced and careful use of measures and datasets in order to avoid the fairness-accuracy trade-off. Moreover, there is a growing amount of research indicating that if using appropriate methodology and metrics, these trade-offs are not always present (Dutta et al., 2020; Black et al., 2022; Cooper et al., 2021) and can be mitigated with careful selection of models (Black et al., 2022) and evaluation methods (Wick et al., 2019). Our results are in agreement with existing works indicating that state-of-the-art bias mitigation methods are typically only effective at removing epistemic discrimination (Wang et al., 2023), i.e. the discrimination made during model development, but not aleatoric discrimination. In order to address aleatoric discrimination, i.e. the bias inherent within the data distribution, and to improve the Pareto frontier, better data curation is required (Dutta et al., 2020). Though our results are unable to provide a significant improvement on the Pareto frontier, we believe that this work presents the first step in this direction and would encourage future work to look into this.

In sum, we present a novel gender-based uncertainty multitask loss reweighting mechanism. We showed that our proposed multitask loss reweighting is able to improve fairness with lesser fairness-accuracy trade-off. Our findings also revealed the importance of accounting for negative transfers and for more effort to be channelled towards improving the Pareto frontier in depression detection research.

ML for Healthcare Implication: Producing a thorough review of strategies to improve fairness is not within the scope of this work. Instead, the key goal is to advance ML for healthcare solutions that are grounded in the framework used by clinicians. In our settings, this corresponds to using each PHQ-8 subcriterion as individual subtask within our MTL-based approach and using a a gender-based uncertainty reweighting mechanism to account for the gender difference in PHQ-8 label distribution. By replicating the inferential process used by clinicians, this work attempts to bridge ML methods with the symptom-based profiling system used by clinicians. Future work can also make use of this property during inference in order to improve the trustworthiness of the machine learning or decision-making model (Huang and Ma, 2022).

In the process of doing so, our proposed method also provide the elusive first evidence that each PHQ-8 subitem's *task difficulty* aligns with its *discrimination capacity* as evidenced from data collected from the largest PHQ-8 population-based study to date (de la Torre et al., 2023). We hope this piece of work will encourage other ML and healthcare researchers to further investigate methods that could bridge ML experimental results with empirical real world healthcare findings to ensure its reliability and validity.

Limitations: We only investigated gender fairness due to the limited availability of other sensitive attributes in both datasets. Future work can consider investigating this approach across different sensitive attributes such as race and age, the intersectionality of sensitive attributes and other healthcare challenges such as cognitive impairment or cancer diagnosis. Moreover, we have adopted our existing experimental approach in alignment with the train-validation-test split provided by the dataset owners as well as other existing works. Future works can consider adopting a cross-validation approach. Other interesting directions include investigating this challenge as an ordinal regression problem (Diaz and Marathe, 2019). Future work can also consider repeating the experiments using datasets collected from other countries and dive deeper into the cultural intricacies of the different PHQ-8 subitems, investigate the effects of the different modalities and its relation to a multitask approach, as well as investigate other important topics such as interpretability and explainability to advance responsible (Wiens et al., 2019) and ethical machine learning for healthcare (Chen et al., 2021).

Acknowledgments

Funding: J. Cheong is supported by the Alan Turing Institute doctoral studentship, the Leverhulme Trust and further acknowledges resource support from METU. A. Bangar contributed to this while undertaking a remote visiting studentship at the Department of Computer Science and Technology, University of Cambridge. H. Gunes' work is supported by the EPSRC/UKRI project ARoEq under grant ref. EP/R030782/1. **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.

References

Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720, 2018.

Andrew Bailey and Mark D Plumbley. Gender bias in depression detection using audio features. *EUSIPCO 2021*, 2021.

Zeynep Sonat Baltaci, Kemal Oksuz, Selim Kuzucu, Kivanc Tezoren, Berkin Kerim Konar, Alpay Ozkan, Emre Akbas, and Sinan Kalkan. Class uncertainty: A measure to mitigate class imbalance. *arXiv preprint arXiv:2311.14090*, 2023.

Hao Ban and Kaiyi Ji. Fair resource allocation in multi-task learning. *arXiv preprint arXiv:2402.15638*, 2024.

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

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.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863, 2022.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*, pages 77–91. PMLR, 2018.

Joseph Cameron, Jiaee Cheong, Micol Spitale, and Hatice Gunes. Multimodal gender fairness in depression prediction: Insights on data from the usa & china. *arXiv preprint arXiv:2408.04026*, 2024.

Bedrettin Cetinkaya, Sinan Kalkan, and Emre Akbas. Ranked: Addressing imbalance and uncertainty in edge detection using ranking-based losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3239–3249, 2024.

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4(1):123–144, 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), 2021.

Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Counterfactual fairness for facial expression recognition. In *European Conference on Computer Vision*, pages 245–261. Springer, 2022.

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, 2023a.

Jiaee Cheong, Selim Kuzucu, Sinan Kalkan, and Hatice Gunes. Towards gender fairness for mental health prediction. In *IJCAI 2023*, pages 5932–5940, US, 2023b. IJCAI.

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*, pages 1–8, USA, sep 2023c.

Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Fairrefuse: Referee-guided fusion for multi-modal causal fairness in depression detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7224–7232, 8 2024a. AI for Good.

Jiaee Cheong, Micol Spitale, and Hatice Gunes. Small but fair! fairness for multimodal human-human and robot-human mental wellbeing coaching, 2024b.

Michelle Chua, Doyun Kim, Jongmun Choi, Nahyoung G Lee, Vikram Deshpande, Joseph Schwab, Michael H Lev, Ramon G Gonzalez, Michael S Gee, and Synho Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, 7(6):711–718, 2023.

A Feder Cooper, Ellen Abrams, and Na Na. Emergent unfairness in algorithmic fairness-accuracy trade-off research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 46–54, 2021.

Jorge Arias de la Torre, Gemma Vilagut, Amy Ronaldson, Jose M Valderas, Ioannis Bakolis, Alex Dregan, Antonio J Molina, Fernando Navarro-Mateu, Katherine Pérez, Xavier Bartoll-Roca, et al. Reliability and cross-country equivalence of the 8-item version of the patient health questionnaire (phq-8) for the assessment of depression: results from 27 countries in europe. *The Lancet Regional Health–Europe*, 31, 2023.

Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4738–4747, 2019.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*, pages 2803–2813. PMLR, 2020.

Yarin Gal. Uncertainty in deep learning. 2016.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognitive Computation*, 14(1), 2022.

Yuan Gong and Christian Poellabauer. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*, pages 69–76, 2017.

Thomas Grote and Geoff Keeling. Enabling fairness in healthcare through machine learning. *Ethics and Information Technology*, 24(3):39, 2022.

Shelley Gupta, Archana Singh, and Jayanthi Ranjan. Multimodal, multiview and multitasking depression detection framework endorsed with auxiliary sentiment polarity and emotion detection. *International Journal of System Assurance Engineering and Management*, 14(Suppl 1), 2023.

Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

Mengjie Han, Ilkim Canli, Juveria Shah, Xingxing Zhang, Ipek Gursel Dino, and Sinan Kalkan. Perspectives of machine learning and natural language processing on characterizing positive energy districts. *Buildings*, 14(2):371, 2024.

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.

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.

Guanjie Huang and Fenglong Ma. Trustsleepnet: A trustable deep multimodal network for sleep stage classification. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 01–04. IEEE, 2022.

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

Patrick Kaiser, Christoph Kern, and David Rügamer. Uncertainty-aware predictive modeling for fair data-driven decisions, 2022.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2014.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H

Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173, 2009.

Selim Kuzucu, Jiaee Cheong, Hatice Gunes, and Sinan Kalkan. Uncertainty as a fairness measure. *Journal of Artificial Intelligence Research*, 81:307–335, 2024.

Doris YP Leung, Yim Wah Mak, Sau Fong Leung, Vico CL Chiang, and Alice Yuen Loke. Measurement invariances of the phq-9 across gender and age groups in chinese adolescents. *Asia-Pacific Psychiatry*, 12(3):e12381, 2020.

Can Li, Sirui Ding, Na Zou, Xia Hu, Xiaoqian Jiang, and Kai Zhang. Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modeling. *Journal of Biomedical Informatics*, 143:104399, 2023a.

Can Li, Dejian Lai, Xiaoqian Jiang, and Kai Zhang. Feri: A multitask-based fairness achieving algorithm with applications to fair organ transplantation. *arXiv preprint arXiv:2310.13820*, 2023b.

Can Li, Xiaoqian Jiang, and Kai Zhang. A transformer-based deep learning approach for fairly predicting post-liver transplant risk factors. *Journal of Biomedical Informatics*, 149:104545, 2024.

Chuyuan Li, Chloé Braud, and Maxime Amblard. Multi-task learning for depression detection in dialogs. *arXiv preprint arXiv:2208.10250*, 2022.

Dongyue Li, Huy Nguyen, and Hongyang Ryan Zhang. Identification of negative transfers in multitask learning using surrogate models. *Transactions on Machine Learning Research*, 2023c.

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.

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.

Raghav Mehta, Changjian Shui, and Tal Arbel. Evaluating the fairness of deep learning uncertainty estimates in medical image analysis, 2023.

Lakshadeep Naik, Sinan Kalkan, and Norbert Krüger. Pre-grasp approaching on mobile robots: a pre-active layered approach. *IEEE Robotics and Automation Letters*, 2024.

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

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *NeurIPS*, 30, 2017.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, and Maja Pantic. Avec'19: Audio/visual emotion challenge and workshop. In *ICMI*, pages 2718–2719, 2019.

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*, pages 99–109, 2020.

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.

Micol Spitale, Jiaee Cheong, and Hatice Gunes. Underneath the numbers: Quantitative and qualitative gender fairness in llms for depression prediction. *arXiv preprint arXiv:2406.08183*, 2024.

Anique Tahir, Lu Cheng, and Huan Liu. Fairness through aleatoric uncertainty. In *CIKM*, 2023.

Michel A Thibodeau and Gordon JG Asmundson. The phq-9 assesses depression similarly in men and women from the general population. *Personality and Individual Differences*, 56:149–153, 2014.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. pages 3–10, 2016.

Marion L Vetter, Thomas A Wadden, Christopher Vinnard, Reneé H Moore, Zahra Khan, Sheri

Volger, David B Sarwer, and Lucy F Faulconbridge. Gender differences in the relationship between symptoms of depression and high-sensitivity crp. *International journal of obesity*, 37(1):S38–S43, 2013.

Hao Wang, Luxi He, Rui Gao, and Flavio Calmon. Aleatoric and epistemic discrimination: Fundamental limits of fairness interventions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 526–536, 2021a.

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.

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.

Yiding Wang, Zhenyi Wang, Chenghao Li, Yilin Zhang, and Haizhou Wang. Online social network individual depression detection using a multitask heterogenous modality fusion approach. *Information Sciences*, 609, 2022.

Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H Chi. Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1748–1757, 2021b.

Ping-Cheng Wei, Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. Multi-modal depression estimation based on sub-attentional fusion. In *European Conference on Computer Vision*, pages 623–639. Springer, 2022.

Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

James R Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F Quatieri. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 11–18, 2016.

Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 506–523. Springer, 2020.

Hua Yuan, Yu Shi, Ning Xu, Xu Yang, Xin Geng, and Yong Rui. Learning from biased soft labels. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

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.

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.

Appendix A. Experimental Setup

A.1. Datasets

For both DAIC-WOZ and E-DAIC, we work with the extracted features and followed the train-validate-test split provided. The dataset owners provided the ground-truths for each of the PHQ-8 sub-criterion and the final binary classification for both datasets.

DAIC-WOZ (Valstar et al., 2016) contains audio recordings, extracted visual features and transcripts collected in a lab-based setting of 100 males and 85 females. The dataset owners provided a standard train-validate-test split which we followed. The dataset owners also provided the ground-truths for each of the PHQ-8 questionnaire sub-criterion as well as the final binary classification.

E-DAIC (Ringeval et al., 2019) contains acoustic recordings and extracted visual features of 168 males and 103 females. The dataset owners provided a standard train-validate-test split which we followed.

Measure	Approach	PHQ-1	PHQ-2	PHQ-3	PHQ-4	PHQ-5	PHQ-6	PHQ-7	PHQ-8	**Binary Outcome**
Acc	Unitask	**0.87**	0.51	0.62	0.57	0.57	0.51	**0.79**	**0.94**	0.66
	Multitask	0.72	0.68	0.57	**0.62**	0.64	**0.68**	0.74	0.89	0.70
	Baseline UW	0.81	**0.70**	**0.64**	0.60	**0.66**	0.62	0.72	0.87	**0.82**
	U-Fair **(Ours)**	0.68	0.66	0.47	0.43	0.43	0.49	0.60	0.74	0.80
F1	Unitask	0.25	0.41	0.44	0.33	**0.33**	**0.53**	0.44	**0.40**	0.47
	Multitask	0.32	0.29	0.50	**0.44**	0.32	0.48	**0.45**	0.29	0.53
	Baseline UW	**0.40**	0.30	**0.51**	0.42	**0.33**	0.31	0.43	0.25	0.29
	U-Fair **(Ours)**	0.29	**0.33**	0.44	0.43	0.27	0.33	0.39	0.00	**0.54**
Precision	Unitask	**1.00**	0.27	0.47	0.31	0.26	0.37	0.67	**0.50**	0.44
	Multitask	0.25	0.25	0.43	**0.39**	0.29	**0.47**	**0.50**	0.25	0.50
	Baseline UW	0.38	0.27	**0.50**	0.37	**0.31**	0.33	0.45	0.20	0.22
	U-Fair **(Ours)**	0.21	**0.27**	0.36	0.30	0.19	0.27	0.32	0.00	**0.56**
Recall	Unitask	0.14	**0.89**	0.41	0.36	0.45	**0.93**	0.33	**0.33**	0.50
	Multitask	0.43	0.33	0.59	0.50	0.36	0.50	0.42	0.33	0.57
	Baseline UW	0.43	0.33	0.53	0.50	0.36	0.29	0.42	0.33	0.43
	U-Fair **(Ours)**	**0.43**	0.44	**0.59**	**0.71**	**0.45**	0.43	**0.50**	0.00	**0.60**
UAR	Unitask	**0.93**	**0.60**	0.58	0.51	0.52	**0.64**	**0.74**	**0.73**	0.60
	Multitask	0.57	0.54	0.57	0.57	0.54	0.62	0.66	0.60	**0.65**
	Baseline UW	0.65	0.56	**0.61**	**0.57**	**0.56**	0.52	0.62	0.62	0.64
	U-Fair **(Ours)**	0.58	0.58	0.49	0.51	0.44	0.47	0.56	0.40	0.63
$\mathcal{M}_{SP}$	Unitask	0.00	1.44	1.92	1.60	0.86	1.44	4.79	**0.96**	0.47
	Multitask	1.92	**0.96**	1.80	1.20	3.51	**1.10**	3.83	2.88	0.86
	Baseline UW	2.88	1.15	1.92	**1.06**	2.16	1.34	1.15	1.44	1.23
	U-Fair **(Ours)**	**0.72**	0.64	**1.28**	1.15	**1.12**	0.66	**0.86**	0.77	**1.06**
$\mathcal{M}_{EOpp}$	Unitask	0.00	1.50	2.00	1.67	0.90	1.50	5.00	1.00	0.45
	Multitask	2.00	**1.00**	1.88	1.25	3.67	**1.14**	4.00	3.00	**0.78**
	Baseline UW	3.00	1.20	2.00	**1.11**	2.25	1.40	1.20	1.50	1.70
	U-Fair **(Ours)**	**0.75**	0.67	**1.33**	1.20	**1.17**	0.69	**0.90**	**0.80**	1.46
$\mathcal{M}_{EOdd}$	Unitask	0.00	1.44	1.90	2.83	1.25	1.53	0.00	0.00	0.54
	Multitask	0.00	1.60	1.83	1.28	9.00	1.88	4.00	0.00	0.76
	Baseline UW	0.00	0.00	2.29	1.49	3.50	2.25	1.50	2.74	1.31
	U-Fair **(Ours)**	**0.80**	**0.80**	**1.43**	**1.16**	**1.33**	**0.75**	**1.00**	0.00	**1.17**
$\mathcal{M}_{EAcc}$	Unitask	0.91	0.81	**0.89**	0.56	**1.20**	0.81	**1.01**	**0.96**	1.44
	Multitask	0.96	**1.09**	0.89	0.89	0.55	1.23	1.01	0.87	0.94
	Baseline UW	**0.96**	1.30	0.84	0.72	0.69	**1.03**	1.08	0.91	1.25
	U-Fair **(Ours)**	1.09	1.16	0.80	**0.96**	0.64	1.28	1.11	1.14	**0.95**

Table 6: Full experimental results for **DAIC-WOZ** across the different PHQ-8 subitems. Best values are highlighted in **bold**.

Measure	Approach	PHQ-1	PHQ-2	PHQ-3	PHQ-4	PHQ-5	PHQ-6	PHQ-7	PHQ-8	**Binary Outcome**
Acc	Unitask	**0.80**	**0.66**	0.59	0.66	0.59	0.61	0.63	0.89	0.55
	Multitask	0.68	0.54	0.48	0.43	0.52	0.54	0.48	0.54	0.58
	Baseline UW	0.75	0.63	**0.61**	**0.73**	**0.73**	0.63	0.59	0.89	0.87
	U-Fair **(Ours)**	0.77	0.61	0.61	0.54	0.71	**0.71**	**0.71**	**0.93**	**0.90**
F1	Unitask	**0.27**	0.24	0.49	**0.60**	**0.47**	**0.45**	**0.49**	**0.25**	**0.51**
	Multitask	0.18	0.32	0.47	0.43	0.40	0.38	0.38	0.07	0.45
	Baseline UW	0.22	**0.36**	**0.54**	0.48	0.29	0.09	0.08	0.00	0.27
	U-Fair **(Ours)**	0.13	0.21	0.39	0.43	0.33	0.33	0.27	0.00	0.45
Precision	Unitask	**0.29**	0.21	0.38	0.45	0.34	0.33	**0.33**	**0.25**	0.36
	Multitask	0.14	0.22	0.33	0.30	0.29	0.28	0.25	0.04	0.32
	Baseline UW	0.20	**0.27**	**0.41**	**0.54**	**0.43**	0.10	0.07	0.00	0.28
	U-Fair **(Ours)**	0.14	0.18	0.35	0.33	0.40	**0.36**	0.27	0.00	**0.46**
Recall	Unitask	**0.25**	0.27	0.69	**0.88**	**0.71**	**0.69**	**0.91**	**0.25**	**0.87**
	Multitask	0.25	**0.55**	**0.81**	0.75	0.64	0.62	0.82	0.25	0.80
	Baseline UW	0.25	0.55	0.81	0.44	0.21	0.08	0.09	0.00	0.26
	U-Fair **(Ours)**	0.13	0.27	0.44	0.63	0.29	0.31	0.27	0.00	0.45
UAR	Unitask	**0.58**	0.51	0.60	**0.69**	**0.60**	**0.60**	**0.65**	**0.60**	0.63
	Multitask	0.50	0.52	0.58	0.53	0.55	0.55	0.58	0.47	0.67
	Baseline UW	0.54	**0.59**	**0.67**	0.64	0.56	0.43	0.40	0.48	0.60
	U-Fair **(Ours)**	0.50	0.48	0.56	0.56	0.57	0.57	0.55	0.50	**0.70**
$\mathcal{M}_{SP}$	Unitask	0.26	2.78	0.81	**1.12**	0.94	1.44	**1.03**	**0.52**	0.65
	Multitask	5.67	2.63	**1.19**	1.40	**0.98**	1.44	1.24	0.41	**1.25**
	Baseline UW	**1.55**	**1.29**	2.58	2.47	2.06	2.32	5.67	0.00	3.86
	U-Fair **(Ours)**	2.06	2.83	1.26	2.67	3.61	**1.29**	1.29	0.00	1.67
$\mathcal{M}_{EOpp}$	Unitask	0.17	1.80	0.53	0.72	0.61	**0.93**	0.67	0.33	0.57
	Multitask	3.67	1.70	0.77	**0.90**	0.63	0.93	0.80	0.26	0.81
	Baseline UW	**1.00**	**0.83**	1.67	1.60	**1.33**	1.50	3.67	0.00	2.31
	U-Fair **(Ours)**	1.33	1.83	**0.82**	1.73	2.33	0.83	**0.83**	0.00	**1.00**
$\mathcal{M}_{EOdd}$	Unitask	0.35	3.65	1.39	1.38	**1.00**	**1.46**	**1.40**	**0.74**	**0.75**
	Multitask	7.00	3.42	**1.29**	**1.63**	1.03	1.53	1.43	0.41	1.41
	Baseline UW	3.00	**1.76**	4.20	6.11	2.00	0.00	0.00	0.00	8.21
	U-Fair **(Ours)**	**2.80**	3.42	2.22	3.67	3.60	2.25	1.90	0.00	5.00
$\mathcal{M}_{EAcc}$	Unitask	1.13	**0.74**	1.45	0.84	1.14	**0.96**	0.71	1.08	0.83
	Multitask	0.63	0.39	0.77	0.41	**0.94**	0.77	0.54	1.77	0.65
	Baseline UW	1.05	0.71	0.48	**0.99**	0.89	0.81	0.88	1.12	0.92
	U-Fair **(Ours)**	**0.96**	0.64	**1.22**	0.47	0.83	0.74	**1.03**	**1.05**	**0.94**

Table 7: Full experimental results for **E-DAIC** across the different PHQ-8 subitems. Best values are highlighted in **bold**.