“It’s not Fair!” – Fairness for a Small Dataset of Multi-Modal Dyadic Mental Well-being Coaching

Jiaee Cheong*
University of Cambridge
Cambridge, UK
jc2208@cam.ac.uk

Micol Spitale*
University of Cambridge
Cambridge, UK
ms2871@cam.ac.uk

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

***Abstract*—In recent years, the affective computing research community has put ethics at the centre of its research agenda. However, many of the currently available datasets for affective computing are ‘small’, making bias and debias analysis challenging. This paper presents the first work to explore bias analysis and mitigation of a small temporal multi-modal dataset for mental well-being by adopting different data augmentation techniques. This proof-of-concept work’s contributions include: i) introducing a novel small temporal multi-modal dataset of dyadic interactions during mental well-being coaching; ii) providing multi-modal and feature importance analyses evaluated via modelling performance and fairness metrics across both high and low-level features; and iii) proposing a simple and effective data augmentation strategy (MixFeat) to debias the small dataset presented in this paper. We conduct extensive experiments and analyses to compare our proposed method against other baseline data augmentation method across various uni-modal and multi-modal setups. Our results indicate that, regardless of the dimensionality of the dataset at hand, the inclusion of a bias analysis section in the conference papers is viable. This paper is therefore a call to the community to include a bias analysis section in ACII conference submissions, similar to the ablation studies conducted in papers submitted to major machine learning conferences.**

***Index Terms*—dyadic interaction, mental well-being, small dataset, bias, data augmentation, fairness**

I. Introduction

In recent years, the advancement in machine learning (ML), the availability of large-scale datasets and the enhancement in computing have led to the widespread use of machine-learning prediction systems in our society [1]. However, the problem of bias in machine-learning based tools and systems are becoming an increasing source of concern [2]. Such risks are also present in the field of affective computing as affect recognition tools are increasingly deployed in a wide range of high-stake use-cases ranging from driver drowsiness detection [3] to mental well-being prediction [4]–[6]. A wide range of fairness measures and bias mitigation techniques have been proposed to quantify and mitigate the bias present in machine learning models [7]–[9]. As existing approaches chiefly focus on large datasets, they may not be effective for small datasets. However, most of the datasets currently available for affective application scenarios are small, i.e., containing just a few hundred instances of data [10], [11].

We are cognizant of the ACII community’s attempt to be more ethically oriented as exemplified by the mandatory ethics impact statement to guard against the potential risks and harms that could be perpetuated by affect-related technology [1]. In line with this, we hypothesise that bias exists even for small datasets and we contend that every analysis on small datasets should have a bias analysis section. Bias in small datasets is a challenging problem as opposed to larger datasets, we do not have millions of data to leverage to conduct large-scale debiasing. In addition, the data collection studies are often conducted in person, which is time and effort-intensive; hence, collecting more data is often not an option.

This is a non-trivial challenge for the ACII community. Figure 1 considers papers that have been published within the last three editions of ACII Conference (i.e., 2019, 2021, 2022) and illustrates that papers focusing on small datasets typically represent 30% to 40% of the total papers accepted for presentation at the main conference track. Based on this, we consider any dataset that has less than 40 (median) subjects or 500 (median) samples ‘small’. We excluded papers that used large benchmark datasets such as AffectNet. This work presents the first attempt to address the challenge of bias in small datasets and calls the community to include a bias analysis section in ACII conference submission regardless of the dimensionality of the dataset at hand.

It does so by introducing the first work which explores the problem of bias in a small dataset and investigating different data augmentation approaches to debias a small temporal multi-modal mental well-being dataset. Since little is known about these well-being dyadic interactions, this work investigated further the contribution of each individual modality (i.e., face, audio, verbal) and the importance of high and low-level features for data-driven applications. Hence, the main contributions of this paper are as follow. First, we introduce a novel small temporal multi-modal dataset of dyadic interactions between a human coach and 11 coachees over four weeks to promote mental well-being, which can be used to analyse and understand the relationships between face, audio, verbal data and well-being. Second, we provide a thorough multi-modal analysis and a feature importance analysis evaluated using both performance and fairness metrics. Third, we propose a simple

*equal contribution, alphabetical order

[1] https://acii-conf.net/2022/authors/submission-guidelines/

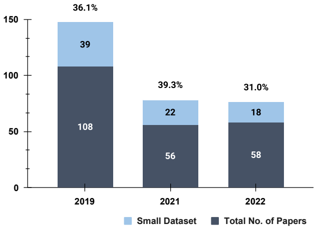

Fig. 1. Proportion of small dataset papers accepted at ACII'19-'22.

and effective data augmentation strategy to reduce the bias in small dataset experimental settings. We compare the proposed method against a baseline data augmentation approach across both single and multiple modalities.

II. Literature Review

A. Dyadic Mental Well-being Coaching

The goal of coaching for mental well-being is to assist the coachee in thriving in their life [12]. Specifically, coaching aims at increasing the coachee's optimism, goal-striving, and general well-being [13]. There exist different coaching strategies to this aim. For instance, *cognitive behavioural* coaching emphasises the connection between thoughts, feelings, and behaviours [13], while *positive psychology-based* coaching encourages the coachee to focus more on the positive aspects of their life rather than the negative ones [14]. Also, *brief-solution focused therapy* (BSFT) has been widely used to encourage coachees to pay attention to the solutions rather than examining problems [15]. Recently, the interest in mental well-being coaching in the fields of affective computing [10], human-agent interactions [16], and machine learning [17] has increased significantly, also due to the pandemic that has exacerbated the need for mental health care [18]. Since such tools are mainly data driven, there is a need for the acquisition of datasets that include data related to mental well-being coaching. Despite this, to date a dataset which contains dyadic interactions during mental well-being coaching has not been introduced.

B. Fairness in Mental Well-being

Though recent attempts at applying ML for the investigation and understanding of mental health has been promising [19], [20], there is only a handful of studies which have looked into bias in mental well-being prediction [21]–[26]. Park et al. [24] conducted their experiments on data collected in a clinical setting with a specific focus on post-partum depression. Zanna et al. [25] conducted their experiments on data collected in the wild with a specific focus on anxiety prediction. Ryan et al. [21] proposed three categories of fairness definitions they deem relevant to mental health. Park et al. [23] analysed bias across gender in mobile mental health assessment and proposed an algorithmic impact remover to mitigate unwanted bias. Bailey and Plumbley [22] attempted to mitigate the gender bias present in the DAIC-WOZ dataset using data re-distribution. [26] examined whether bias exists in existing mental health datasets and algorithms and provided practical suggestions to avoid hampering bias mitigation efforts in ML for mental health. However, all of the existing works consist of relatively large datasets (more than 500 samples or more than 40 subjects) which differ from our small dataset setup. In addition, no investigation has specifically looked into the problem of bias in the context of a human-human dyadic mental well-being coaching setup.

C. Data Augmentation for Bias Mitigation

Bias can be mitigated at the pre-processing, in-processing or post-processing stage [9]. The proposed method falls under the pre-processing data augmentation category which has proven to be effective in mitigating bias [27]. There is minimal work that focus on mitigating bias for a small dataset setup [28]. For a small dataset problem, [28] leverages on a small annotated dataset to debias a larger dataset. This is distinct from our work as it focuses specifically on an item recommendation system. Existing research has indicated that re-sampling outperforms reweighting for correcting sampling bias [29]. Given the above, we propose a simple re-sampling or data augmentation method based on the mixup method proposed in [30]. *Mixup* has proven to be a simple yet highly effective method to address challenges ranging from robustness [31], fairness [32] and regularisation [33]. As a result, *Mixup* has been frequently used as a benchmark for new data augementation techniques and there are recent works proposing new variations of the original method [32], [34], [35].

III. Problem Formulation

We study the problem of model fairness using a machine learning approach, 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, $y_i \in Y$ is thus the outcome where $Y = 1$ denotes "high-PA" (i.e., high positive affect, indicative of higher levels of mental well-being) whereas $Y = 0$ denotes "low-PA" (i.e., low positive affect, indicative of lower levels mental well-being). The fairness measure of a model M is then evaluated according to the sensitive groups of individuals defined by their sensitive attributes A (e.g., gender and race). In our experiments, both the sensitive attributes analysed are binary. They belong to the majority group, e.g.: $A_{race} = 1$ if they are White or $A_{race} = 0$ if otherwise. $\hat{Y}$ denotes the predicted class.

A. Fairness Measures

The fairness measures are similar to that in [36] and [25].

- **Equal Accuracy** (EA), a group-based metric, is used to compare the group fairness between the models. This can be understood as the accuracy gap between the majority and the minority group:

$$EA = |MAE(\hat{Y}|A = 1) - MAE(\hat{Y}|A = 0)|, \quad (1)$$

where MAE represents the Mean Absolute Error (MAE) of the classification task of each sensitive group.

- **Disparate Impact** (DI), measures the ratio of positive outcome ($\hat{Y} = 1$) for both the majority and minority group as represented by the following equation:

$$DI = \frac{Pr(\hat{Y} = 1|A = 0)}{Pr(\hat{Y} = 1|A = 1)} \tag{2}$$

The two measures above represent different aspects of bias. EA evaluates fairness based on the model's predictive performance measured in terms of accuracy. whereas DI evaluates fairness based purely on the predicted outcomes $\hat{Y}$.

B. Proposed Method: MixFeat

Our proposed methodology (MixFeat) is based on the data augmentation technique proposed by [30]. Given a dataset of size N where A represents the audio cue, F represents the facial cue and V represents the verbal cue, the new training sample (A_k, F_k, V_k) is therefore generated as follow:

$$\begin{aligned} A_k &= \lambda_A \cdot A_i + (1 - \lambda_A) \cdot A_j \\ F_k &= \lambda_F \cdot F_i + (1 - \lambda_F) \cdot F_j \\ V_k &= \lambda_V \cdot V_i + (1 - \lambda_V) \cdot V_j \end{aligned} \tag{3}$$

where $i, j \in \{1, ...N\}$, $i \neq j$ and $\lambda_A, \lambda_F, \lambda_V \sim$ Beta(0,1). We use the above method to generate synthetic samples for the minority group to obtain balanced samples across the sensitive attributes of race and gender. The intuition behind this method is that if we generate new samples by mixing up features from other samples with the same sensitive attribute, the new samples will inherit the sensitive-attribute specific features. Thus, this method preserves the relation between the synthetic samples and supervision signal which gives the algorithm more samples to learn from without imposing strong assumptions [30]. Figure 2 outlines the experimental setup and how the method is integrated into the overall classification pipeline.

IV. Dataset and Methods

This section reports the dataset definition and methodology for detecting well-being in human-human dyadic interactions.

A. The AFAR-BSFT Dataset

TABLE I
Ground-truth distribution across different groups.

	Gender		
	Female	Male	p
Low-PA	23.1 ± 5.0	30.4 ± 0.9	**0.01**
High-PA	40.8 ± 5.1	37.3 ± 2.0	**0.02**
	Race		
	White	Non-Cauc	p
Low-PA	26.4 ± 6.3	24.4 ± 4.7	0.57
High-PA	39.7 ± 4.6	38.8 ± 1.3	0.69

We collected a dataset of human-human dyadic interactions between a human well-being coach and 11 participants over four weeks. The human well-being coach was instructed to deliver a Brief-Solution Focused Therapy (BSFT) style coaching, asking participants to focus on solutions rather than analysing the problem [15] for about 20 minutes. After each session, we asked participants to complete the Positive And Negative Affect Scale (PANAS) [37] to evaluate their positive and negative affect. The dataset was collected at the Affective Intelligence and Robotics (AFAR) Lab, and we refer to it as the AFAR-BSFT DB henceforth.

1) Data Collection: 11 participants were recruited via email advertising of the University of Cambridge. We conducted the study in a dedicated room (see Figure 2) where a human well-being coach and one participant were seated in front of each other. Video recordings were done using two external cameras, one facing the participant and the other facing the human coach that can be used for further analysis (beyond the scope of this paper) on dyadic interactions during the coaching practice. We collected 44 videos (11 participants × 4 weeks, 20 mins per session) of dyadic well-being coaching interactions. 3 out of 44 sessions were excluded due to technical issues (e.g., corrupted video or audio recordings).

2) Data Annotation: Two human annotators labelled the gender and race of the participants (with a 100% agreement). This resulted in 7 participants being labelled as males and 4 as females, and 8 participants being labelled as Whites, and 3 as non-Whites. For each sensitive group (gender and race), we report the mean and standard deviation of the target construct. We used *t-test* and *one-way ANOVA* to examine the differences between the means across the different groups as reported in Table I, where a statistical significant difference between gender labels is reported. We evaluated the participants' positive affect using the self-report results of the PANAS questionnaire [37], which has been widely used by practitioners to identify strengths and concerns in mental well-being. We computed the positive affect (PA) and negative affect (NA) sub-scales according to the manual in [37]. We set the threshold value to 33.3, corresponding to the mean value for the American population [37], and we then classified the videos collected into "high-PA" and "low-PA". This resulted in 17 videos for the "low-PA" and 26 videos for the "high-PA" class. Given the small size of the dataset, we decided to limit our problem to a binary classification problem. The AFAR-BSFP DB will be made available for research purposes [2] in the form of feature sets accompanied by the corresponding labels upon the publication of this work.

B. Self-report Affect Detection

1) Dataset Pre-processing: Before extracting the features, we split the audio and video recordings, and we asked a human annotator to transcribe the dyadic interactions between the human coach and the participants manually. The annotator also took note of the timestamp of the speech so that we were able to diarize the audio files.

2) Multi-modal Feature Extraction: Given the audio-visual recordings, we extrapolated multi-modal features as follows. We extracted the facial features using OpenFace 2.0 [38] –

[2] AFAR GitHub: https://github.com/Cambridge-AFAR/AFAR-BSFT-DB

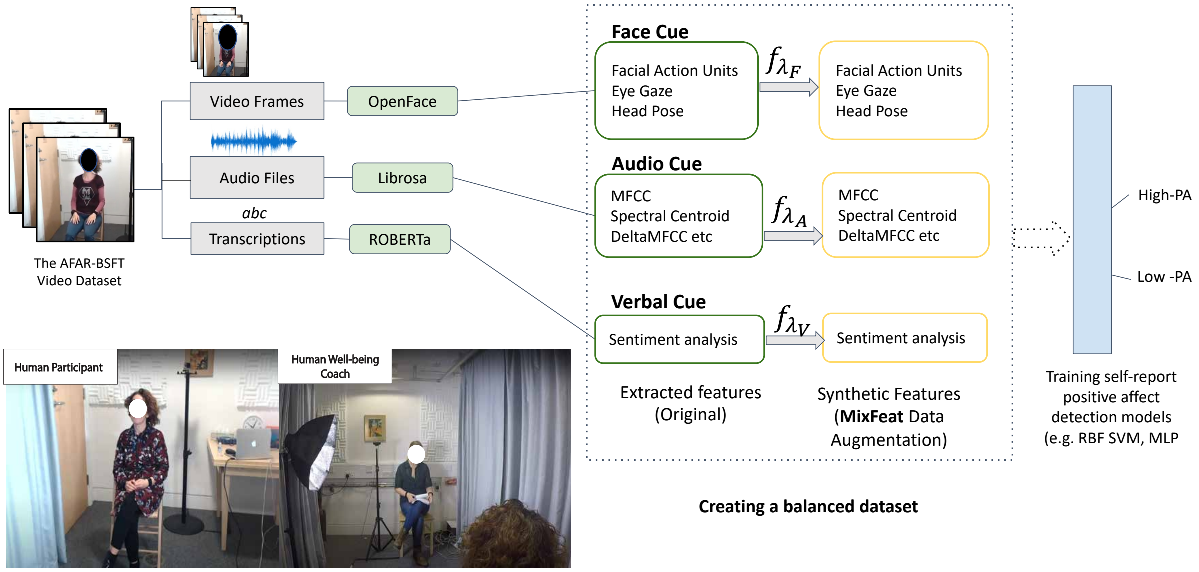

Fig. 2. The model pipeline with our proposed data augmentation technique: **MixFeat**. After extracting the high-level features from the dataset, we generate synthetic sample features using Equation 3. Each modality's feature generation process is chiefly governed by their respective $\lambda \sim$ Beta(0,1) parameters. For the baseline data augmentation method, we conduct random upsampling in place of the synthetic feature generation to obtain a balanced dataset. Setting of the data collection is shown below. Interaction between the participants (on the left side), and the human well-being coach (on the right side).

which represents one of the state-of-the-art tools for extracting facial features within the ACII community, e.g., in [11] – resulting in the following: eye gaze directions, the intensity and presence of 17 facial action units (FAUs), facial landmarks, head pose coordinates, and point-distribution model (PDM) parameters for facial location, scale, rotation and deformation, resulting in 709 facial features. We used librosa[3] to extract the audio features, namely pitch, speech duration, 128 Mel spectrograms, 20 MFCC, 20 delta MFCC, spectral centroid, and RMS, which results in 172 audio features, as in previous works, e.g., [39]. We used ROBERTa[4] to extract the predicted sentiment from the participants' transcriptions resulting in 2 verbal features (label and probability), as in [10].

3) Pre-processing: We first removed constant and null features to prepare the multi-modal features for the machine learning models. Then, we decided to condense the temporal information of each video clip into statistical descriptors as in [10], [11], computing a fixed-length vector for each multi-modal feature of each clip that consists of mean, median, standard deviation, minimum, maximum, and auto-correlation with 1-second lag, resulting in a facial feature vector with size $41 \times 709 \times 6$, in an audio feature vector with size $41 \times 172 \times 6$, and in a verbal feature vector with size $41 \times 2 \times 6$.

4) Feature Selection: We defined the high-level and low-level features as interpretable (e.g., facial action unit, pitch) and not-interpretable (e.g., spectral features) to select the most informative ones for the positive affect detection model [40]. Specifically, the low-level features were 1) facial: facial landmarks, head pose coordinates, and point-distribution model (PDM) parameters, and 2) audio: 128 Mel spectrograms, 20 MFCC, 20 delta MFCC, spectral centroid, and RMS; while the high-level features were 1) facial: facial action units and gaze, 2) audio: pitch and speech duration, and 3) verbal: the sentiment of the speech. Given the differences in dimensionality between low-level and high-level features, we conducted a principal component analysis (PCA) to reduce the size of the features while keeping 80% of the information. The PCA analysis resulted in i) 5 principal components (PCs) for high-level features for face, 10 PCs for low-level features for face, and ii) 2 features for high-level features for audio (no PCA conducted because the number of high-level audio features was already small, i.e., equal to 2), and 3 PCs for low-level features for audio.

5) Data Fusion Strategies: We explored different state-of-the-art data fusion strategies [41], [42]. First, we experimented with early fusion, which consisted of concatenating features from different modalities that resulted in a single vector of features. Then, we experimented with different late fusion strategies, namely majority voting (soft and hard) and stacking (soft and hard). In majority voting, the final decision is made according to the most frequent class label predicted across the different uni-modal models (hard) or the classifier whose predicted class probability is the highest across the different uni-modal models (soft). In stacking, the final decision is made by another classifier (e.g., logistic regression model) fed by either the predicted class label (hard) or the predicted class probabilities (soft) of each uni-modal model.

V. Modeling and Bias Analysis Results

A. Modeling and Feature Selection

We first conducted experiments using various machine learning techniques as in [10], [11] – namely logistic regres-

[3] https://librosa.org/doc/latest/index.html

[4] https://huggingface.co/docs/transformers/model_doc/roberta

sion, linear support vector machine (SVM), random forest tree, bagging, XGBoost, AdaBoost, decision tree, radial basis function support vector machine (RBF-SVM), multi-layer perceptron (MLP), and long-short term memory (LSTM) neural network – and validating them with three different cross-validation approaches (i.e., 5-fold CV, leave-one-subject-out (LOSO), leave-one-week-out). Our results showed that the outperforming models were RBF-SVM and MLP among the machine learning techniques we experimented with. Due to space constraints, we only report the outperforming model results and analyses in the following sections.

TABLE II
UNI-MODAL HIGH VS LOW-LEVEL FEATURE MODELING RESULTS. ABBREVIATIONS. R: RBF SVM. M: MLP. VALUES IN BOLD DENOTE THE HIGHEST ACCURACY OR THE HIGHEST F1 FOR THE SPECIFIC MODALITY ACROSS THE THREE SETS OF EXPERIMENTS.

	Uni-modal					
	Face		Audio		verbal	
	Low	High	Low	High	Low	High
R-Acc	0.27	**0.45**	0.27	**0.65**	N/A	0.50
R- F1	0.33	**0.44**	0.33	**0.68**	N/A	0.43
M-Acc	0.35	**0.43**	0.35	**0.47**	N/A	0.49
M- F1	0.37	**0.37**	0.37	**0.37**	N/A	0.78

TABLE III
MULTI MODAL (BI-MODAL) HIGH VS LOW-LEVEL FEATURE MODELING RESULTS. ABBREVIATIONS. R: RBF SVM. M: MLP. VALUES IN BOLD DENOTE THE HIGHEST ACCURACY OR F1 FOR THE SPECIFIC MODALITY ACROSS THE THREE SETS OF EXPERIMENTS.

	Face and Audio					
	Early		Soft Voting		Stacking	
	Low	High	Low	High	Low	High
R-Acc	**0.46**	0.38	0.45	**0.59**	**0.76**	0.71
R-F1	**0.53**	0.48	0.46	**0.62**	**0.80**	0.76
M-Acc	**0.43**	0.33	0.40	**0.56**	0.71	**0.75**
M-F1	**0.44**	0.40	0.52	**0.65**	0.76	**0.78**

B. *Low vs High Level Feature Analysis*

We trained different experimental models with either the high or low-level features, and compared their performances. Table II reports the results of the uni-modal models, while Table III reports the results of the multi-modal (i.e., face and audio, audio and verbal, face and verbal) models. We have not reported the tri-modal (i.e., face, audio, and verbal) analysis because the verbal feature vector contains only high-level information, making comparison impossible. Our results showed that the models trained with high-level features performed better in terms of accuracy and F1 in all uni-modal and most of the multi-modal setups (see Tables II and III). Therefore in the rest of our work, we only considered high-level features to train the models and conduct the bias analysis.

C. *Uni-modal vs Multi-modal Analysis*

We conducted multiple experiments to compare uni-modal and multi-modal (with either early or late fusion) approaches. The results are collected in the Original column of Tables VI and VII. We found that overall the multi-modal modeling outperformed the uni-modal models. Specifically, the average accuracy score for early, soft voting, and stacking techniques of the models trained with multi-modal data (i.e., face, audio, and verbal) is always higher than the average accuracy score of models trained with uni-modal data (e.g., accuracy score for RBF SVM model trained with only face data is equal to 41%, while the accuracy score for the same model trained with face, audio, and verbal data using an early fusion technique is equal to 54%). Interestingly, the audio modality consistently gives the best accuracy and fairness scores across all three sets of experiments. This could be due to the fact that BSFT coaching is dialogue oriented. Across the fairness metrics, not all multi-modal approaches led to a reduction in bias. For example, the MLP-based soft major voting approach seems to reduce gender and race biases more with respect to audio or face uni-modal approaches, however, the early fusion techniques for both MLP and RBF SVM-based approaches increase both gender and race biases.

D. *Week-based Analysis*

We conducted a longitudinal analysis to understand the effect of time on performances and bias by comparing these metrics over 4 weeks. Our results show that the accuracy of the models increases over the week for both uni-modal and multi-modal approaches, e.g., the overall accuracy of the face and verbal models is outperforming in Week 4 (60%) with respect to the previous weeks, and analogously the overall accuracy for the late fusion strategies is better in Week 3 and 4 with respect to the previous weeks. Across bias, we observe different results for the uni-modal and multi-modal approaches. The gender and race bias in the uni-modal approaches is reduced spreadly throughout the weeks (i.e., the data does not show any patterns); on the other hand, the gender and race bias in the multi-modal approaches seems to be more reduced in the early weeks (i.e., Week 1 and 2) with respect to the last weeks.

VI. DEBIASING APPROACH AND RESULTS

A. *Baseline Method: Data Balancing*

To provide a comparison to our proposed method, we use a baseline data balancing method to mitigate the bias present. As the dataset is highly imbalanced, we employ a similar data balancing method as [36]. We re-sample the minority group by randomly oversampling datapoints to obtain an augmented dataset with samples balanced across both sensitive attributes. After data balancing, we retrain the models and capture the results in Table VI and VII.

B. *Proposed Method: MixFeat Augmentation*

The implementation of our proposed method is similar to that of the baseline method. The key difference is that instead of randomly oversampling data points, we generate synthetic samples according to the method outline in Equation 3.

TABLE IV

Uni-modal Week-based accuracy and bias analysis. Abbreviations. R: RBF SVM. M:MLP. UAR: Unweighted Average Recall.

	Week 1						Week 2						Week 3						Week 4					
	Face		Audio		verbal		Face		Audio		verbal		Face		Audio		verbal		Face		Audio		verbal	
	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M
Female	0.57	0.50	0.50	0.33	0.33	0.50	0.14	0.57	0.43	0.57	0.43	0.14	0.43	0.57	0.43	0.43	0.43	0.29	0.50	0.83	0.33	0.00	0.33	0.50
Male	0.43	0.00	0.67	0.33	0.00	0.67	0.50	0.25	0.75	0.50	0.50	0.75	0.33	0.00	0.67	0.00	0.33	0.67	0.50	0.25	0.75	0.50	0.25	0.75
White	0.50	0.50	0.50	0.33	0.33	0.50	0.38	0.50	0.63	0.63	0.50	0.38	0.43	0.29	0.43	0.14	0.57	0.43	0.71	0.57	0.57	0.14	0.43	0.71
Non-White	0.67	0.00	0.67	0.33	0.00	0.67	0.00	0.33	0.33	0.33	0.33	0.33	0.33	0.67	0.67	0.67	0.00	0.33	0.00	0.67	0.33	0.33	0.00	0.33
Overall Acc	0.50	0.33	**0.56**	0.33	0.22	0.56	0.27	0.45	0.55	0.55	0.45	0.36	0.40	0.40	0.50	0.30	0.40	0.40	0.50	**0.60**	0.50	0.20	0.30	**0.60**
Overall F1	0.67	0.40	0.71	0.40	0.36	0.71	0.43	0.50	0.67	0.62	0.50	0.53	0.50	0.40	0.62	NaN	0.50	0.57	0.67	0.67	0.67	0.33	0.46	0.75
Overall UAR	0.52	0.25	0.58	0.33	0.17	0.58	0.25	0.41	0.53	0.51	0.44	0.40	0.38	0.38	0.55	0.31	0.33	0.43	0.43	0.58	0.50	0.24	0.25	0.57
EA_{Gender}	0.33	0.50	0.17	**0.00**	0.33	0.17	0.36	0.32	0.32	0.07	0.07	0.61	0.10	0.57	0.24	0.43	0.10	0.38	**0.00**	0.58	0.42	0.50	**0.08**	0.25
EA_{Race}	0.17	0.50	0.17	**0.00**	0.33	0.17	0.38	0.17	0.29	0.29	0.17	**0.04**	**0.10**	0.38	0.24	0.52	0.57	0.10	0.71	0.10	0.24	0.19	0.43	0.38
DI_{Gender}	0.80	1.00	1.50	4.00	0.67	1.50	0.88	**0.70**	1.17	**1.05**	1.31	1.17	0.93	0.58	1.40	2.33	0.93	1.17	0.75	0.75	1.20	1.50	**0.60**	1.00
DI_{Race}	0.80	0.00	0.80	1.00	**0.00**	0.80	0.76	0.44	1.14	1.60	0.44	1.14	0.39	**0.00**	0.78	**0.00**	0.93	1.17	0.67	**0.00**	1.17	2.33	0.93	1.00

TABLE V

Multi-modal Week-based accuracy and bias analysis. Abbreviations. R: RBF SVM. M:MLP. UAR: Unweighted Average Recall.

	Week 1						Week 2						Week 3						Week 4					
	Early		Soft Voting		Stacking		Early		Soft Voting		Stacking		Early		Soft Voting		Stacking		Early		Soft Voting		Stacking	
	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M
Female	0.67	0.50	0.67	0.50	0.50	0.67	0.29	0.57	0.57	0.57	0.57	0.29	0.71	0.71	0.71	0.29	0.57	0.43	0.43	1.00	0.57	0.43	0.57	0.14
Male	0.33	0.00	0.67	0.33	0.33	0.67	0.75	0.25	0.75	0.50	0.75	0.75	0.33	0.00	0.67	0.33	0.67	1.00	0.75	0.25	0.75	0.50	0.25	0.75
White	0.50	0.50	0.67	0.50	0.50	0.67	0.50	0.50	0.75	0.63	0.63	0.50	0.71	0.43	0.71	0.14	0.71	0.71	0.63	0.75	0.75	0.50	0.63	0.38
Non-White	0.67	0.00	0.67	0.33	0.33	0.67	0.33	0.33	0.33	0.33	0.67	0.33	0.33	0.67	0.67	0.67	0.33	0.33	0.33	0.67	0.33	0.33	0.00	0.33
Overall Acc	0.56	0.33	0.67	0.44	0.44	**0.67**	0.45	0.45	0.64	0.55	0.64	0.45	0.60	0.50	**0.70**	0.30	0.60	0.60	0.55	**0.73**	0.64	0.45	0.45	0.36
Overall F1	0.71	0.40	0.77	0.55	0.62	0.77	0.57	0.50	0.71	0.62	0.67	0.57	0.67	0.44	0.73	0.22	0.50	0.67	0.71	0.77	0.75	0.57	0.57	0.53
Overall UAR	0.54	0.25	0.67	0.42	0.42	0.67	0.47	0.41	0.60	0.51	0.65	0.47	0.52	0.45	0.69	0.36	0.57	0.62	0.53	0.67	0.60	0.44	0.36	0.40
EA_{Gender}	0.33	0.50	**0.00**	0.17	0.17	**0.00**	0.46	**0.32**	0.18	0.07	0.18	0.46	0.38	0.71	0.05	0.05	0.10	0.57	0.32	0.75	0.18	0.07	0.32	0.61
EA_{Race}	0.17	0.50	**0.00**	0.17	0.17	**0.00**	0.17	0.17	0.42	0.29	0.04	0.17	0.38	0.24	**0.05**	0.52	0.38	0.38	0.29	0.08	0.42	0.17	0.63	0.04
DI_{Gender}	0.80	1.00	2.00	2.00	1.00	2.00	1.40	**0.70**	1.40	**1.05**	2.33	1.40	0.93	0.78	2.33	2.33	1.17	0.93	1.17	0.88	1.40	1.31	**0.70**	1.75
DI_{Race}	0.80	**0.00**	1.00	**0.67**	**0.40**	1.00	1.33	0.44	1.33	1.60	1.07	1.33	0.39	**0.00**	1.17	2.33	1.17	1.75	1.14	**0.00**	1.33	2.00	1.07	1.60

C. *Overall Comparison*

Although both methods are effective in reducing bias, our proposed method seems to produce an outcome that is less variable compared to the baseline method. With reference to Table VI, we see that across the uni-modal experiments, our method consistently produces a more accurate and fairer outcome across most metrics for both sensitive attributes compared to the baseline. The only modality where the baseline performance is better is the "Verbal" Modality using the MLP predictor. Within the multi-modal approach depicted in Table VII, we see that this gap in predictive and fairness performance is diminished. For instance, for soft-voting, the baseline method produces a better outcome across both accuracy and fairness compared to our proposed method whereas our proposed method performs better across early fusion. On the other hand, for stacking, our method performs better in terms of fairness whereas the baseline method performs better in terms of accuracy.

VII. Discussion and Conclusion

Our results indicate the following. First, a multi-modal approach consistently outperforms uni-modal approaches across performance metrics of accuracy and F1 score. However, they may introduce additional bias which is consistent with the findings in [36]. Second, we find that models trained with high-level features performed better in terms of accuracy and F1 in both the uni-modal and multi-modal setups. Third, our results showed that the proposed data augmentation method more consistently improves fairness across both the uni and multi-modal experiments compared to the baseline method. Our results suggest that using high-level features, a multi-modal setup and an interpolation-based data augmentation strategy may produce the best outcome in terms of model performance and fairness measures.

An important takeaway is that first, a multi-modal approach provides more information for a machine learning algorithm to learn from. As a result, this will lead to better performance in terms of accuracy and fairness as supported by previous literature [42]. Second, a multi-modal approach seems to

TABLE VI

Unimodal Debiasing Results. Abbreviations. R: RBF SVM. M:MLP. UAR: Unweighted Average Recall. Values in bold denotes highest accuracy or the fairest outcome for the specific modality across the three sets of experiments.

	Original						Baseline Comparison						Proposed Method					
	Face		Audio		Verbal		Face		Audio		Verbal		Face		Audio		Verbal	
	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M
Female	0.41	0.63	0.44	0.33	0.37	0.33	0.41	0.62	0.45	0.31	0.34	0.34	0.55	0.62	0.66	0.62	0.41	0.52
Male	0.43	0.14	0.71	0.36	0.29	0.71	0.45	0.17	0.83	0.52	0.24	0.83	0.62	0.24	0.97	0.48	0.48	0.72
White	0.48	0.48	0.55	0.31	0.45	0.48	0.48	0.48	0.55	0.31	0.45	0.48	0.62	0.48	0.86	0.55	0.48	0.55
Non-White	0.25	0.42	0.50	0.42	0.08	0.42	0.38	0.31	0.72	0.52	0.38	0.69	0.55	0.38	0.76	0.56	0.66	0.69
Overall Acc	0.41	0.46	0.54	0.34	0.34	0.46	0.43	0.40	0.64	0.41	0.29	0.59	**0.59**	0.43	**0.81**	0.55	0.45	**0.62**
Overall F1	0.57	0.52	0.68	0.37	0.43	0.78	0.51	0.43	0.77	0.56	0.29	0.73	0.72	0.50	0.83	0.68	0.52	0.70
Overall UAR	0.42	0.39	0.58	0.35	0.33	0.52	0.43	0.40	0.64	0.41	0.29	0.59	0.59	0.43	0.81	0.55	0.45	0.62
EA_{Gender}	**0.02**	0.49	0.27	**0.02**	0.08	0.38	0.03	0.45	0.38	0.21	0.10	0.48	0.07	0.38	0.31	0.14	0.07	0.21
EA_{Race}	0.23	**0.07**	0.05	0.11	0.36	0.07	0.10	0.17	0.17	0.21	**0.07**	0.21	**0.07**	0.10	0.10	**0.01**	0.17	0.14
DI_{Gender}	**0.88**	0.72	1.29	1.74	0.91	1.23	0.75	0.63	1.26	1.82	0.67	1.21	0.82	0.75	1.21	**1.19**	**0.94**	1.37
DI_{Race}	0.68	0.12	**0.97**	1.41	0.60	1.06	0.68	0.24	1.08	1.58	0.50	1.12	0.76	0.33	1.33	1.06	0.84	0.96

TABLE VII

Multi Modal Debiasing Results. Abbreviations. R: RBF SVM. M:MLP. UAR: Unweighted Average Recall. Values in bold denotes highest accuracy or the fairest outcome for the specific data fusion method across the three sets of experiments.

	Original						Baseline Comparison						Proposed Method					
	Early		Soft Voting		Stacking		Early		Soft Voting		Stacking		Early		Soft Voting		Stacking	
	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M
Female	0.52	0.70	0.63	0.44	0.56	0.37	0.62	0.62	0.72	0.59	0.48	0.55	0.76	0.66	0.66	0.62	0.62	0.55
Male	0.57	0.14	0.71	0.43	0.50	0.79	0.76	0.34	0.69	0.45	0.66	0.83	0.69	0.59	0.72	0.48	0.73	0.62
White	0.59	0.55	0.72	0.45	0.62	0.55	0.76	0.55	0.86	0.55	0.55	0.66	0.86	0.52	0.90	0.59	0.69	0.66
Non-White	0.42	0.42	0.50	0.42	0.33	0.42	0.62	0.41	0.55	0.48	0.83	0.72	0.59	0.72	0.48	0.52	0.90	0.52
Overall Acc	0.54	0.51	0.66	0.44	0.54	0.51	0.69	0.48	**0.71**	0.52	**0.69**	0.57	**0.72**	0.62	0.69	0.55	0.67	0.59
Overall F1	0.67	0.55	0.74	0.51	0.60	0.63	0.75	0.48	0.71	0.63	0.68	0.74	0.71	0.72	0.69	0.61	0.64	0.56
Overall UAR	0.52	0.42	0.67	0.44	0.53	0.58	0.69	0.48	0.71	0.52	0.57	0.69	0.72	0.62	0.69	0.55	0.67	0.59
EA_{Gender}	**0.05**	0.56	0.08	**0.02**	**0.06**	0.42	0.14	0.28	0.03	0.14	0.17	0.28	0.07	0.07	0.07	0.14	0.11	0.07
EA_{Race}	0.17	**0.14**	0.22	**0.03**	0.29	0.14	**0.14**	**0.14**	0.31	0.07	0.28	0.07	0.28	0.21	0.41	0.07	0.21	0.14
DI_{Gender}	1.10	0.83	1.69	1.34	1.24	1.47	0.76	0.81	**0.88**	1.20	1.57	1.44	**0.96**	1.06	1.15	0.84	**0.90**	1.29
DI_{Race}	0.91	0.13	1.21	1.38	0.85	1.40	0.68	0.26	0.78	**0.94**	1.40	**0.95**	**1.04**	0.68	1.26	**0.94**	0.81	**0.95**

balance out the bias of each individual modality. For the unimodal approach, we see a greater variation in results between the original, baseline and proposed method. However, this variation is less pronounced for the multi-modal approach. We hypothesise that this is because making use of all three modalities provide the algorithm with more comprehensive information to learn from which balances out any gap, imbalances or bias that is introduced by each singular modality, as supported by the literature [43]. Third, our results indicate high-level features seem to contain more information (or less noise) in accordance with [40]. Thus, for a small dataset, it might be better to use higher-level features for the model to learn from. Moreover, by introducing a data augmentation method which relies on high-level features, we will be better able to preserve the subject's anonymity and privacy. In this study, we have relied on external annotation for gender and race. Future work may consider using the labels obtained from participants to avoid introducing labelling bias from external annotation [44] and explore further temporal information by conducting studies with multiple sessions over time to derive longitudinal insights. Future work may also investigate other sensitive attributes such as age.

Ethical Impact Statement

This study has been approved by the ethics committee of the department. Participants signed consent forms consenting for their data to be used within the context of research. Participants were reimbursed with the minimum wage per hour fee. This research attempts to avoid any bias against certain groups of people that could result in discrimination even in small dataset and it cannot be used to deceive or negatively impact human rights. However, our results are limited to the dataset included in this work. Future work should repeat the same analysis on other small datasets to further validate our hypothesis.

The ACII 2022 Conference introduced an ethical statement as a requirement. In line with this, we suggest that papers submitted to the ACII Conference that conduct research using human data, in the form of a large or a small dataset, should have a bias analysis section similar to the ablation studies provided in conference papers submitted to the major machine learning conferences such as CVPR and NeurIPS. This work presents the first effort in this direction by conducting a bias and debias analysis on a small dataset, as a case study of multi-modal human-human dyadic mental well-being coaching.

Acknowledgments

Funding: M. Spitale and H. Gunes are supported by the EPSRC/UKRI under grant ref. EP/R030782/1 (ARoEQ). J. Cheong is funded by the Alan Turing Institute Doctoral Studentship and the Leverhulme Trust.

Open Access: For open access purposes, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Data access: Raw data related to this publication cannot be openly released due to anonymity and privacy issues, however we have released the feature representation: https://github.com/Cambridge-AFAR/AFAR-BSFT-DB.

References

[1] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN computer science*, vol. 2, no. 3, p. 160, 2021.

[2] S. Barocas, M. Hardt, and A. Narayanan, "Fairness in machine learning," *Nips tutorial*, vol. 1, p. 2, 2017.

[3] M. Ngxande, J. Tapamo, and M. Burke, "Bias remediation in driver drowsiness detection systems using generative adversarial networks," *IEEE Access*, vol. 8, pp. 55 592–55 601, 2020.

[4] C. Su, Z. Xu, J. Pathak, and F. Wang, "Deep learning in mental health outcome research: a scoping review," *Translational Psychiatry*, vol. 10, no. 1, pp. 1–26, 2020.

[5] A. B. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications," *Psychological medicine*, vol. 49, no. 9, pp. 1426–1448, 2019.

[6] M. Srividya, S. Mohanavalli, and N. Bhalaji, "Behavioral modeling for mental health using machine learning algorithms," *Journal of medical systems*, vol. 42, no. 5, pp. 1–12, 2018.

[7] M. Hort, Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "Bias mitigation for machine learning classifiers: A comprehensive survey," *arXiv preprint arXiv:2207.07068*, 2022.

[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM CSUR*, vol. 54, no. 6, pp. 1–35, 2021.

[9] J. Cheong, S. Kalkan, and H. Gunes, "The hitchhiker's guide to bias and fairness in facial affective signal processing: Overview and techniques," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 39–49, 2021.

[10] N. I. Abbasi, M. Spitale, J. Anderson, T. Ford, P. B. Jones, and H. Gunes, "Computational audio modelling for robot-assisted assessment of children's mental wellbeing," in *ICSR 2022*. Springer, 2023, pp. 23–35.

[11] L. Mathur, M. Spitale, H. Xi, J. Li, and M. J. Matarić, "Modeling user empathy elicited by a robot storyteller," in *ACII 2021*. IEEE, 2021, pp. 1–8.

[12] V. Hart, J. Blattner, and S. Leipsic, "Coaching versus therapy: A perspective." *Consulting Psychology Journal: Practice and Research*, vol. 53, no. 4, p. 229, 2001.

[13] L. S. Green, L. G. Oades, and A. M. Grant, "Cognitive-behavioral, solution-focused life coaching: Enhancing goal striving, well-being, and hope," *Journal of Positive Psychology*, vol. 1, no. 3, pp. 142–149, 2006.

[14] M. E. Seligman, "Coaching and positive psychology," *Australian Psychologist*, vol. 42, no. 4, pp. 266–267, 2007.

[15] S. De Shazer and Y. Dolan, "More than miracles: The state of the art of solution-focused brief therapy," 2012.

[16] M. Spitale, M. Axelsson, and H. Gunes, "Robotic mental well-being coaches for the workplace: An in-the-wild study on form," in *ACM/IEEE HRI2023*, ser. HRI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 301–310.

[17] J. Xu, S. Song, K. Kusumam, H. Gunes, and M. Valstar, "Two-stage temporal modelling framework for video-based depression recognition using graph representation," *arXiv preprint arXiv:2111.15266*, 2021.

[18] M. Spitale and H. Gunes, "Affective robotics for wellbeing: A scoping review," in *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2022, pp. 1–8.

[19] L. He, M. Niu, P. Tiwari, P. Marttinen, R. Su, J. Jiang, C. Guo, H. Wang, S. Ding, Z. Wang *et al.*, "Deep learning for depression recognition with audiovisual cues: A review," *Information Fusion*, vol. 80, 2022.

[20] Z. Zhang, W. Lin, M. Liu, and M. Mahmoud, "Multimodal deep learning framework for mental disorder recognition," in *FG 2020*. IEEE, 2020, pp. 344–350.

[21] S. RYAN and G. DOHERTY, "Fairness definitions for digital mental health applications," 2022.

[22] A. Bailey and M. D. Plumbley, "Gender bias in depression detection using audio features," in *EUSIPCO 2021*. IEEE, 2021.

[23] J. Park, R. Arunachalam, V. Silenzio, V. K. Singh *et al.*, "Fairness in mobile phone–based mental health assessment algorithms: Exploratory study," *JMIR formative research*, vol. 6, no. 6, p. e34366, 2022.

[24] Y. Park, J. Hu, M. Singh, I. Sylla, I. Dankwa-Mullan, E. Koski, and A. K. Das, "Comparison of methods to reduce bias from clinical prediction models of postpartum depression," *JAMA network open*, vol. 4, no. 4, pp. e213 909–e213 909, 2021.

[25] K. Zanna, K. Sridhar, H. Yu, and A. Sano, "Bias reducing multitask learning on mental health prediction," in *ACII 2022*. IEEE, pp. 1–8.

[26] J. Cheong, S. Kuzucu, S. Kalkan, and H. Gunes, "Towards gender fairness for mental health prediction," in *IJCAI 2023*, 2023.

[27] J. Cheong, S. Kalkan, and H. Gunes, "Counterfactual fairness for facial expression recognition," in *ECCV 2022 Workshops*. Springer, 2023, pp. 245–261.

[28] T. Schnabel and P. N. Bennett, "Debiasing item-to-item recommendations with small annotated datasets," in *ACM RecSys 2020*, 2020, pp. 73–81.

[29] J. An, L. Ying, and Y. Zhu, "Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients," in *ICLR 2021*, 2021.

[30] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR 2018*, 2018.

[31] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," in *ICLR 2020*, 2020.

[32] C.-Y. Chuang and Y. Mroueh, "Fair mixup: Fairness via interpolation," in *ICLR 2021*, 2021.

[33] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *IEEE/CVF ICCV 2019*, 2019, pp. 6023–6032.

[34] D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt, "Pixmix: Dreamlike pictures comprehensively improve safety measures," in *IEEE/CVF CVPR 2022*, 2022, pp. 16 783–16 792.

[35] X. Hao, Y. Zhu, S. Appalaraju, A. Zhang, W. Zhang, B. Li, and M. Li, "Mixgen: A new multi-modal data augmentation," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 379–389.

[36] S. Yan, D. Huang, and M. Soleymani, "Mitigating biases in multimodal personality assessment," in *ICMI 2020*, 2020, pp. 361–369.

[37] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[38] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *FG 2018*. IEEE, 2018, pp. 59–66.

[39] E. M. Benssassi and J. Ye, "Investigating multisensory integration in emotion recognition through bio-inspired computational models," *IEEE Transactions on Affective Computing*, 2021.

[40] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019.

[41] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, pp. 345–379, 2010.

[42] L. Mathur and M. J. Matarić, "Introducing representations of facial affect in automated multimodal deception detection," in *ICMI 2020*, 2020, pp. 305–314.

[43] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello, "Bias and fairness in multimodal machine learning: A case study of automated video interviews," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 268–277.

[44] J. Cheong, S. Kalkan, and H. Gunes, "Causal structure learning of bias for fair affect recognition," in *WACV 2023*, January 2023, pp. 340–349.