Debiasing Multimodal Models for PTSD Analysis

Merna Bibars
Georgia Institute of Technology
mbibars3@gatech.edu

Ajeet Subramanian
Georgia Institute of Technology
asubramanian91@gatech.edu

Aarushi Wagh
Georgia Institute of Technology
awagh31@gatech.edu

Abstract

The diagnosis of Post-Traumatic Stress Disorder (PTSD) poses significant challenges due to symptom overlap with other mental health conditions and the potential for bias in evaluation, particularly concerning gender. This study investigates methods to measure and mitigate gender bias in PTSD diagnosis using multimodal machine learning models. By employing a combination of resampling and adversarial debiasing techniques, the study evaluates the bias and performance trade-offs in a Random Forest, Perceiver and DeepConvLSTM-based model trained on the extended DAIC-WOZ dataset. Experimental results demonstrate improved fairness as measured by the Statistical Parity Ratio (SPR), with minimal compromises to model accuracy. Our multimodal model acheived a 76 f1-score for PTSD classification beating previous work. The code for our project is in the following github repository: https://github.com/ajeet-sub/VLM-Debiasing-Project

1. Introduction

Post-Traumatic Stress Disorder (PTSD) is a severe mental health condition that develops in response to traumatic experiences. Its accurate diagnosis is essential for effective treatment but remains a challenge due to symptom overlap with other mental health disorders such as anxiety and depression. The complex and subjective nature of PTSD symptoms often leads to variability in diagnosis, even among trained professionals [6].

In recent years, machine learning (ML) models have shown significant promise in improving diagnostic accuracy by analyzing diverse data sources, including clinical transcripts, audio recordings, and facial expressions. Multimodal approaches, which integrate these modalities, are particularly effective in capturing the nuances of PTSD symptoms [3][4]. Despite their potential, these models are not without shortcomings. A critical concern is the potential for these systems to inherit biases present in the training data, which can disproportionately affect specific demographic groups.

Gender bias in healthcare has been widely documented, with diagnostic disparities often observed between male and female patients. For instance, women may be underdiagnosed for PTSD or misdiagnosed with other disorders due to societal perceptions or gendered patterns in symptom presentation. When such biases are embedded in machine learning models, the disparities are not only perpetuated but also automated, posing significant ethical and clinical risks.

Ensuring fairness in ML systems used for healthcare applications is a priority, as biased models can exacerbate inequalities and harm patient outcomes. Fairness in this context requires that model predictions are not influenced by sensitive attributes such as gender, race, or socioeconomic status. However, despite the growing body of research on fairness in machine learning, there is a notable lack of studies addressing bias in PTSD diagnostic models.

This study aims to bridge this gap by focusing on measuring and mitigating gender bias in multimodal machine learning models for PTSD detection. We propose a systematic evaluation of bias using Statistical Parity Ratio (DPR) and explore debiasing techniques such as resampling and adversarial learning.

This work contributes to the field by providing insights into the trade-offs between fairness and model performance, offering a framework that ensures ethical and effective deployment of ML models in mental health diagnostics.

2. Related Work

Existing research on PTSD detection has made significant strides, particularly with the adoption of advanced machine learning architectures such as stochastic transformers and multimodal embeddings. These approaches leverage data from multiple modalities—audio, visual, and tex-

tual—to enhance diagnostic accuracy. However, the issue of bias in such systems, especially gender bias, has largely been overlooked.

For instance, [4] proposed a stochastic transformer-based model for PTSD PCL-C score prediction, incorporating uncertainty quantification to improve reliability. Their approach utilized video data to extract nuanced behavioral cues, highlighting the importance of multimodal data in capturing PTSD symptoms and acheiving a SOTA RMSE of 1.98. In a follow-up study, [3] applied a similar framework to audio data, demonstrating the potential for speech-based models in PTSD PCL-C score prediction reaching a RMSE of 2.92.

Other work focusing on PTSD binary classification approached the problem using only text modality such as [8]. They compared the use of zero-shot text augmentation with GPT embedding to a TF-IDF embedding as input to an SVC model. Their model acheived an f1-score of 71.

However, while these methods achieved impressive diagnostic performance, the studies did not address the fairness of these models across demographic groups, leaving a critical gap in understanding and mitigating bias. In related fields, researchers have explored debiasing methods for machine learning models used in mental health diagnostics. [9] investigated adversarial debiasing in a multimodal personality assessment, focusing on removing sensitive attribute information (e.g., gender or race) from feature embeddings. Their work demonstrated that adversarial learning can significantly reduce bias while maintaining model performance, offering a compelling framework for fairness in multimodal systems.

Moreover, [1] examined gender bias in depression detection models using audio features. They employed statistical parity and sufficiency as fairness metrics, demonstrating that resampling techniques could balance the training data and reduce bias. However, these methods often led to a trade-off with model performance, underscoring the challenge of achieving both fairness and accuracy.

Despite these advancements, a consistent baseline for bias measurement in PTSD-related tasks remains absent. Existing studies in PTSD detection rarely evaluate the fairness of their models, focusing instead on improving diagnostic accuracy. This gap is particularly concerning given the ethical implications of deploying biased systems in clinical settings.

This study builds on these approaches by applying bias measurement and debiasing techniques, specifically resampling, to PTSD diagnostic models. By focusing on the Statistical Parity Ratio (SPR) as a fairness metric, this study aims to provide a comprehensive evaluation of bias and propose solutions that balance fairness with diagnostic performance. To our knowledge, this is the first study to explore demographic bias in PTSD diagnosis.

3. Problem Statement

This research aims to evaluate and mitigate gender bias in PTSD diagnosis using a multimodal deep learning model. The absence of established baselines for bias measurement and debiasing in PTSD-related tasks motivates this study. The primary research question is: How can we effectively measure and mitigate the gender bias present in the evaluation of PTSD patients using a multimodal model?

4. Methods

The first step to answer our research question is to utilise a biased PTSD dataset as an input to a deep learning model. Then, measure the bias using any of the measurements used previously in the literature. After calculating the bias quantitatively, we apply two debiasing techniques to the biased data and deep learning model, then recalculate the bias measurment. A block diagramm showing an overview of the methodology is shown on Figure 1

4.1. Dataset

The extended DAIC-WOZ dataset [5, 2, 7] is a comprehensive resource designed for diagnosing mental health conditions, including PTSD, depression, and anxiety. It comprises semistructured clinical interviews with data available from three modalities:

- Audio: Audio recordings of the interviews are preprocessed to extract meaningful features. These include:
  - Spectrogram embeddings generated using the pre-trained Audio Spectrogram Transformer (AST), which captures frequency-domain features over time.
  - eGeMAPS (extended Geneva Minimalistic Acoustic Parameter Set) features, commonly used for emotion recognition tasks.
  - MFCC (Mel-Frequency Cepstral Coefficients), which represent the short-term power spectrum of sound.
- Text: Transcripts of the interviews are processed using RoBERTa, a state-of-the-art transformer-based language model. RoBERTa encodes semantic and contextual information, producing high-dimensional embeddings that represent the textual content of the conversation.
- Visual: Video recordings are analyzed to capture nonverbal cues such as facial expressions, gaze, and posture. DenseNet, a convolutional neural network known for its efficiency in feature extraction, is used to process video frames and generate visual embeddings.

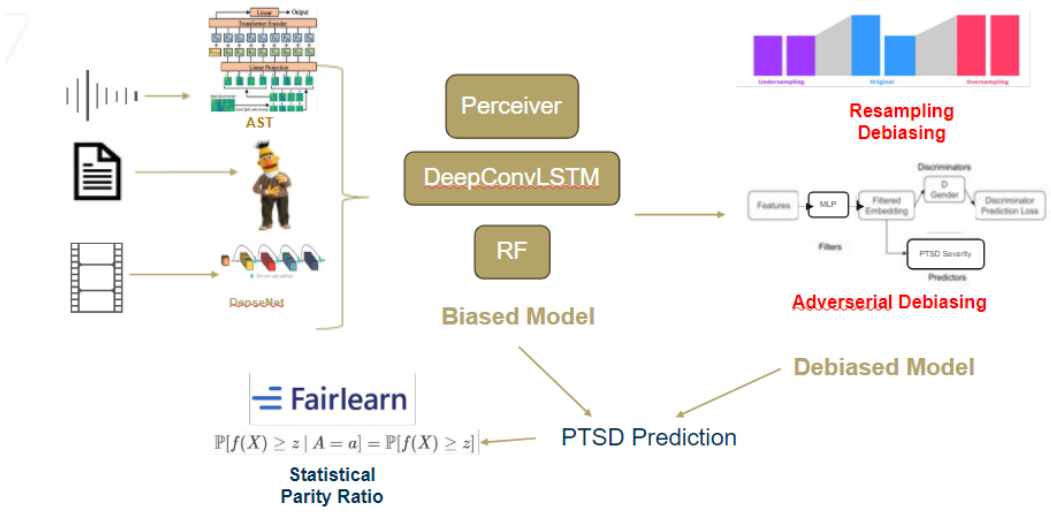

Figure 1. A block diagram showing the entire methodology. The biased data is used as input to any of the 3 models (RF, DeepConvLSTM, Perceiver) and the biased model is use to calculate the SPR. Two debiasing methods are then investigated followed by a re-calculation of the SPR.

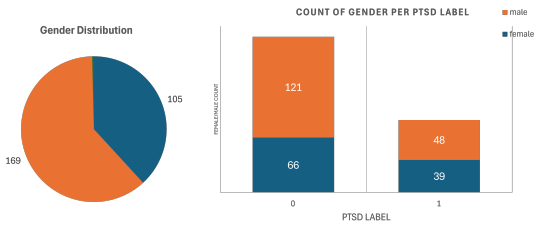

Figure 2. Distribution of gender demographic across PTSD labels in the E-DAIC dataset.

Each participant is labeled with:

- PTSD Severity Scores (PCL-C): Ranging from 0 to 85, where higher scores indicate greater severity of PTSD.
- Binary PTSD Diagnosis: A label indicating the presence (1) or absence (0) of PTSD.

In this study we focused on the binary classification of PTSD. The data shows a bias in the distribution of genders across both PCL-C score and presence or absence of PTSD as shown in Figure 2.

The dataset is split into training (60%), validation (20%), and testing (20%) subsets to evaluate model performance.

4.2. Feature Extraction and Projection

The extracted embeddings from the modalities are treated in two ways:

- The features are flattened and projected to a unified dimension of [768,1] to be the input to the perceiver model
- The features are used with the temporal dimension maintained as input to the DeepConvLSTM

For our experiments, we ended up only using the audio and text feature embeddings, as we didn't have the time to perform experiments on the given

4.3. Models

Several models were employed to predict PTSD and measure the SPR:

- Random Forest (RF): A classic machine learning model was used as a baseline. The RF trained on unimodal audio features provided initial performance benchmarks. However, the multimodal RF model failed to learn meaningful patterns and consistently predicted the same output for all inputs, showing poor generalization and an inability to leverage multimodal data.
- Multimodal Perceiver: This architecture, designed for processing multimodal inputs, encountered similar issues as the RF model. It exhibited no learning from the data and produced uniform predictions, highlighting limitations in its capacity to handle the complexity of the dataset.
- DeepConvLSTM: This model showed significant promise. It is designed to process sequential data effectively. Both a unimodal and multimodal models were built to process audio features and text features separately and multimodal features. The DeepConvLSTM demonstrated the ability to learn meaningful

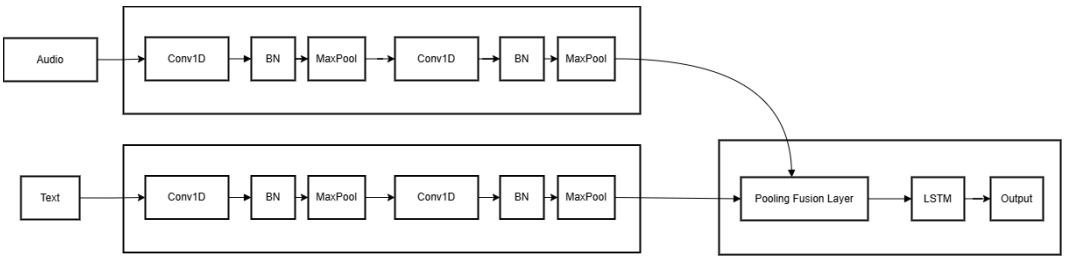

Figure 3. The architecture for the multimodal DeepConvLSTM.

patterns, achieving better performance compared to the other models. The architecture of the multimodal model is shown in Figure 3. The unimodal architecture is similar to the multimodal one, but without the branches for the other modalities. Both batchnorm, and dropout were incorporated throughout the architecture.

4.4. Bias Measurment

Statistical Parity (SP) is a fairness metric used to evaluate whether the predictions of a model are independent of sensitive attributes such as gender, race, or age. In the context of this study, SP assesses whether the likelihood of receiving a positive prediction (e.g., a PTSD diagnosis) is the same across genders. A model satisfies statistical parity if the predictions are distributed equally across all groups defined by a sensitive attribute A. For a sensitive attribute A (e.g., gender), the SP condition can be expressed mathematically for classification as follows:

$$SP = \frac{Pr(f(X) = 1|A = group_1)}{Pr(f(X) = 1|A = group_2)}$$

Where:

- $f(X)$ is the model's prediction (e.g., PTSD diagnosis).
- A is the sensitive attribute (e.g., gender).
- $group_1$ and $group_2$ are the two demographic groups (e.g., male and female).

A value of SP close to 1 indicates fairness in the model's predictions

4.5. Debiasing Methods

4.5.1 Data Resampling

For our baseline approach, we will be implementing a form of data resampling akin to [1], which ensures equal gender representation in the data before training the model. We will be doing this by undersampling the male group, taking a random sample of the data such that there is equal gender representation while also maintaining the same training, validation, and development split.

4.5.2 Adversarial Debiasing

Our main approach, which will be our novel implementation, is an adversarial debiasing approach similar to [9]. In this approach, the model will consist of 2 separate neural networks:

- Our DeepConvLSTM, which is trained on the feature embeddings from the E-Daic data and returns a binary PTSD classification.
- An adversarial network with the goal of predicting the gender from the PTSD class predicted by the DeepConvLSTM.

After each iteration of the training loop, the gradient of the adversarial network is added to the weight update stage of the DeepConvLSTM training, which in turn will reduce the amount of information on the gender variable on the prediction of PTSD. By doing this, we can retain all the information from the training data while removing bias. We hypothesize that this will lead to similar performance levels to unbiased models while providing much better SP values.

5. Experimental Setup

We compare between unimodal and multimodal DeepConvLSTM to see if there is an effect on the bias depending on the modalities. The perceiver was also trained on both unimodal and multimodal features, however the transformer model wasn't able to learn any patterns from the data compared to the DeepConvLSTM.

5.1. Desired Output

The goal is to develop a multimodal model for PTSD classification while reducing gender bias in predictions. The desired outputs for this study are:

- **Statistical Parity Ratio (SP):**
  - Without Debiasing: Models, such as RF and initial DeepConvLSTM implementations, exhibit substantial bias with SP $\ll 1$, indicating that gender significantly affects predictions.

  - Resampling Debiasing: Relatively big improvement in SP (SP $\ll 1$), though predictive performance is more affected due to the reduction of data by undersampling.
  - Adversarial Debiasing: Slight improvement in SP (SP ≤ 1), approaching fairness with a smaller decrease in predictive performance.

- **Model Performance:** Maintain high F1 Scores for binary classification while balancing fairness.

5.2. Loss and Metrics

For the DeepConvLSTM, the loss function we are using is a weighted binary cross entropy with logits where the weights are the ratio of the postive to the negative class. The choice of weighted loss is to avoid the effect of the class imbalance on the results and focus on only the gender imbalance. In our experiment with the adversarial component, the adversarial network will have just a binary cross entropy loss function.

$$Loss = -\beta Y_{true} log(Y_{pred}) - (1-\beta)(1-Y_{true}) log(1-Y_{pred})$$

Where $\beta = 1 - \frac{\sum_y y}{|Y_true|}$

To measure the actual performance of our model, we will be using an f1 score as well as the SP ratio values. The f1 score will help us determine how accurate our model is in predicting PTSD while the SP value determines how biased the predictions are with respect to gender.

$$f1 - score = 2 \times \frac{precision \times recall}{precision + recall}$$

where $precision = \frac{TP}{TP+FP}$ and $recall = \frac{TP}{TP+FN}$

6. Results

Table 1 shows the results of both biased and debiased models across the different modalities. We can see that the best validation f1-score is achieved using the text and audio modality with the raw biased data (76.09). This beats the previous score in the literature of (71) [8] using only text modality. However, that model still shows a little bias with 0.74 SPR. The results also show that even though debiasing using undersampling of the data to have equal proportion of genders improved the SPR to 0.78; the performance of the model decreased to 73.1 f1-score as expected.

The most significant improvement in SPR is for the audio modality which increased from 0.55 to 0.95 by resampling gender only. However that led to a severe decrease in the performance of f1-score from 74 to 53. As expected the adverserial debiasing method for the audio modality improved the SPR by a lower extent (from 0.55 to 0.88) however the decrease in model performance is much less compared to the resampled gender method. The f1-score for the adverserial debiased model is 71.22 compared to 53 for the undersampling debiased model.

7. Ablations

To understand the contribution of different modalities and encoders to the overall performance and fairness of the PTSD diagnostic models, we conducted a series of ablation studies. The goal was to evaluate how visual, audio, and text embeddings, processed by various encoders, influenced the metrics of interest. Table 2 summarizes the results.

7.1. Visual Embeddings

For visual embeddings, we experimented with three encoders: DenseNet, BoVW_OpenFace, and VGG. Among these:

- DenseNet: Performed the best, achieving an F1 score of 49% with a Random Forest (RF) model. However, the fairness metric (SP Ratio) remained low at 0.447, indicating persistent bias in predictions.

- BoVW_OpenFace and VGG: Failed to learn meaningful patterns from the data, resulting in F1 scores of 47% and an SP Ratio of 0.

7.2. Audio Embeddings

For audio features, we tried using two features:

- AST (Audio Spectrogram Transformer): Successfully extracted embeddings that worked well for regression tasks with the Perceiver model, achieving an RMSE of 21. However, the model could not address bias, as reflected by an SP Ratio of 0.

- OpenSMILE_eGeMAPS: Although a widely used feature set for emotion recognition, the extracted features were too large to be processed effectively, and no results were obtained.

7.3. Text Embeddings

Text embeddings were extracted using RoBERTa, and two models were evaluated:

- Random Forest (RF): Achieved the best results among all experiments, with an F1 score of 67% and an SP Ratio of 0.568. This highlights RF's suitability for leveraging text embeddings effectively.

- Perceiver: Struggled to learn from text embeddings, resulting in poor regression performance (RMSE = 21) and an SP Ratio of 0.

Modality	Training F1	Validation F1	Validation SPR	Bias
Text	100	70	0.648	biased
Text	100	56.26	0.95	resampled gender
Text	99.41	62.35	1.23	Adversarial
Audio	99	74	0.55	biased
Audio	97	53	0.95	resampled gender
Audio	94.99	71.22	0.88	Adversarial
Text-Audio	94.46	76.09	0.74	biased
Text-Audio	82.9	73.1	0.78	resampled gender

Table 1. F1 score of training and validation data sets with modality and debiasing method used with DeepConvLSTM

Data	**Encoder**	**Dimensions**	**Model**	**Metric**	**SP Ratio**
Video	DenseNet	(648, 1922)	RF	49% (F1)	**0.447**
Video	BoVW_OpenFace	(6485, 102)	RF	47% (F1)	0
Video	VGG	(648, 4098)	RF	47% (F1)	0
Audio	AST	(1214, 768)	Perceiver	**21 (RMSE)**	0
Audio	OpenSMILE_eGeMAPS	Features too large, could not be extracted	-	-	-
Text	RoBERTa	(1278, 768)	RF	**67% (F1)**	**0.568**
Text	RoBERTa	(1278, 768)	Perceiver	21 (RMSE)	0

Table 2. Ablation Studies: Impact of modality, encoder, and model on performance and fairness.

8. Discussion

Based on the results, it can be seen that using both debiasing methods leads to a SPR closer to 1, with resampling leading to a better value than the adversarial method. The adversarial method, however, had only slight, if not similar results to the biased model in terms of performance (F1-score), mathing our hypothesis and showing its promise as a strong debiasing method for multimodal models.

8.1. Limitations

Our experimentation, while providing promising results, comes with several limitations. For one, the actual model we use for PTSD diagnosis (DeepConvLSTM) is not the state of the art, with its performance paling in comparison to the stochastic transformer model in [3]. Though we are primarily addressing bias, it would be better to test it on the best models in the field.

A key immediate next step to expand on this paper would be to implement the visual features of the e-daic dataset, and train the model on all 3 modalities. Adding onto that, implementing the adversarial method in a multimodal context would also provide more comprehensive results to our study. The lack of these results in our paper is a strong weakness that is present, but can be fixed with more time.

8.2. Future Work Suggestions

In terms of work beyond the paper, a key pivot that could lead to improved results is the implementation of the state of the art stochastic transformer from [3]. Seeing how debiasing methods, especially the adversarial approach, improve on its fairness could lead to stronger results that can be directly applied in the field. Furthermore, pivoting away from the e-daic dataset to train the model is something worth considering. Not having access to the raw data (outside) of audio, heavily bottlenecks any of the models used because of the quality of feature extractions. Finding higher quality data (or raw video data) might provide stronger results.

9. Conclusion

For this project, we implement both a resampling and adversarial debiasing technique to mitigate the gender bias present in the e-daic dataset for PTSD analysis. The resampling method led to the most unbiased results, but with the worst performance, while the adversarial method was still very unbiased with only a slight performance drop. We find that the adversarial method shows strong promise as a debiasing method that doesn't come at the cost of information loss. Given its scalability to larger models, it can become a staple method in the field of PTSD analysis to mitigate biases in automated diagnoses.

References

[1] Andrew Bailey and Mark D. Plumbley. Gender bias in depression detection using audio features. *EUSIPCO*, 2021. 2, 4

[2] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068, 2014. 2

[3] Mamadou Dia, Ghazaleh Khodabandelou, and Alice Othmani. Paying attention to uncertainty: A stochastic multimodal transformers for post-traumatic stress disorder detection using video. *Computer Methods and Programs in Biomedicine*, 257(10843), 2024. 1, 2, 6

[4] Mamadou Dia, Ghazaleh Khodabandelou, and Alice Othmani. A novel stochastic transformer-based approach for post-traumatic stress disorder detection using audio record- ing of clinical interviews. *2023 IEEE 36th Inter- national Symposium on Computer-Based Medical Sys- tems (CBMS)*, pages 700–705, June 2023. 1, 2

[5] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). 2

[6] Ellen C. Meltzer, Tali Averbuch, Jeffrey H. Samet, Richard Saitz, Khelda Jabbar, Christine Lloyd-Travaglini, and Jane M. Liebschutz. Discrepancy in diagnosis and treatment of post-traumatic stress disorder (ptsd): treatment for the wrong reason. *The journal of behavioral health services research*, 39:190–201, 2012. 1

[7] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pages 3–12, 2019. 2

[8] Yuqi Wu, Jie Chen, Kaining Mao, and Yanbo Zhang. Automatic post-traumatic stress disorder diagnosis via clinical transcripts: A novel text augmentation with large language models. In *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–5, 2023. 2, 5

[9] Shen Yan, Di Huang, and Mohammad Soleymani. Mitigating biases in multimodal personality assessment. *2020 International Conference on Multimodal Interaction*, pages 361–369, 2020. 2, 4

Student Name	Contributed Aspects	Details
Merna Bibars	Feature Extraction and Implementation	Extracted audio features using AST. Implemented unimodal and multimodal DeepConvLSTM and trained the on the PTSD dataset.
Ajeet Subramanian	Feature Extraction and Implementation	Extracted text embeddings from transcript data using the Roberta Model. Also implemented the adversarial debiasing method to the DeepConvLstm.
Aarushi Wagh	Feature Extraction and Implementation	Extracted visual embeddings using Densenet. Also implemented the Random Forest model, and the bias measurement script to calculate Statistical Parity Ratio and Difference.

Table 3. Contributions of team members.