Bias Reducing Multitask Learning on Mental Health Prediction

Khadija Zanna, Kusha Sridhar, Han Yu, Akane Sano
Department of Electrical and Computer Engineering
Rice University
Houston, USA
(khzanna, kh82, hy29, Akane.Sano)@rice.edu

***Abstract*—There has been an increase in research in developing machine learning models for mental health detection or prediction in recent years due to increased mental health issues in society. Effective use of mental health prediction or detection models can help mental health practitioners re-define mental illnesses more objectively than currently done, and identify illnesses at an earlier stage when interventions may be more effective. However, there is still a lack of standard in evaluating bias in such machine learning models in the field, which leads to challenges in providing reliable predictions and in addressing disparities. This lack of standards persists due to factors such as technical difficulties, complexities of high dimensional clinical health data, etc., which are especially true for physiological signals. This along with prior evidence of relations between some physiological signals with certain demographic identities restates the importance of exploring bias in mental health prediction models that utilize physiological signals. In this work, we aim to perform a fairness analysis and implement a multi-task learning based bias mitigation method on anxiety prediction models using ECG data. Our method is based on the idea of epistemic uncertainty and its relationship with model weights and feature space representation. Our analysis showed that our anxiety prediction base model introduced some bias with regards to age, income, ethnicity, and whether a participant is born in the U.S. or not, and our bias mitigation method performed better at reducing the bias in the model, when compared to the reweighting mitigation technique. Our analysis on feature importance also helped identify relationships between heart rate variability and multiple demographic groupings.**

***Index Terms*—bias, epistemic uncertainty, fairness metric, Monte-Carlo dropout, protected label, multi-task learning**

I. Introduction and Background

Irrespective of the advancements that machine learning has made possible in several fields such as language technologies, computer vision and medical applications, negative bias is often embedded in the essence of machine learning algorithms. Negative bias is an erroneous assumption made by an algorithm, that is systemically prejudiced against certain groups of people. Negative biases can be encoded in algorithms due to a number of factors, the first being imbalance in the representation of different population categories in the training data. If certain demographics are lacking from the sample data, models trained on this data often do not generalize when applied to new data that contains those missing demographics [12]. The second factor that could introduce negative bias in machine learning algorithms is biased human labeling. This is due to the fact that data that are fed into models, especially supervised or semi-supervised models which are widely used in various jurisdictions, are manually labeled by humans who are inherently biased. These models ultimately reflect people's impressions and sustain or further magnify bias from the labeled data [34]. Training and labeled data aside, there is still risk of introducing bias in the functional form of a model through features and modeling techniques [12].

Due to the expanding popularity of machine learning and the inherent biases that come with it, there has been an increased focus on bias and fairness in the field. There are several works on how to accurately define and measure fairness in systems [18], [21], [29], how to analyze and mitigate bias using various techniques [12], [22], [38], and a few works that assess the trade-offs between fairness and accuracy in these models [26], [33].

Mental health poses a significant challenge for an individual's well-being, and it is estimated that 792 million people lived with a mental health disorder in 2017 [27]. This is slightly more than one in ten people globally (10.7%). Rising statistics like this has led to an increase in research on mental health, including mental health and well-being prediction using physiological signals over the past couple of years. Several authors research on predicting stress levels, and various mental health conditions using data collected both in clinical settings and in the wild [1], [7], [30], [31], [36], [39].

With the rise in popularity of this field of research and its widespread applications in psychiatry and psychology, the need for effective bias mitigation techniques has become apparent, especially with the sensitive nature of physiological data. Previous research that explored emotional responses captured via physiological signals found some relations between blood pressure, electrodermal activity and race, and blood pressure and gender [40]. These findings re-iterate

This work was supported by NSF #1840167 and #2047296.

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

the importance of exploring possible bias in mental health prediction models that utilize physiological signals.

Despite several relatively successful bias mitigation efforts in general machine learning literature, when it comes to the field of mental health prediction and emotion recognition, there is still a lack of standards in methods for reducing bias. This ultimately leads to many challenges in providing reliable predictions and in addressing disparities [24]. This lack of standards persists due to factors such as technical difficulties in regard to data collection, complexities of high dimensional clinical health data, lack of knowledge of underlying causal structures, and challenges to algorithm evaluation [17], [25].

Only a few works to date have explored methods to reduce bias in mental health prediction models, however there has been research on bias analysis and mitigation in emotion recognition models mostly on the facial recognition aspect of computer vision. Majority of these works state non-representative data ratio as a major cause of bias in facial recognition models [5], [13], [15], [23], [37], and gender, age and skin-tone to be some of attributes data is unbalanced by [8], [15], [19]. The methods these papers have introduced are often data and model-driven, making them non-transferable to emotion detection using other types of data [37]. There have also been a few works that consider bias in emotion recognition using speech data [12], [14]. They mention gender as a leading demographic that is a source of bias in models [12]. Other works conducted analyses on emotions in general healthcare procedures, and found that emotional responses in medical practices are heavily culturally mediated with both individual factors like gender, age, occupation, and social factors like food habits, availability, etc. coming into play [2].

Park et al. explored the performance of different methods to reduce bias for clinical prediction algorithms for post-partum depression [24]. They implemented three methods, reweighting, prejudice removal, and removal of the race label to logistic regression, random forest, and extreme gradient boosting models, to mitigate bias based on race. They found that reweighing improves their chosen fairness metrics without compromising accuracy, prejudice remover performed less reliably, and omitting the race label made no significant difference to the fairness metric. Although this study provided promising results, there are some limitations that come with it. First, the authors conducted their experiments based on data collected in a clinical setting, and these results might not hold true on data from the wild. They also implemented these methods on well-known interpretable, and often used models in fairness literature [33], and only analyzed bias based on race without exploring other factors. To counter some of these limitations, we will be testing our method on a Long Short term memory (LSTM)-based anxiety prediction model using physiological data collected in the wild. The reason we are using anxiety to assess fairness and our bias mitigation method is due to the fact that most scientific work addressing anxiety and its disorders has thus far been conducted among European Americans. This has created striking gaps of inequalities in anxiety disorder research and practice [43], and with this work, we aim to aid in closing some of this gap.

One area of interest in the field of fairness in general is the trade-off between accuracy and fairness in models [22], [33]. It is often the case that bias is first introduced to models from the data, and that means using sensitive information in the functional form of the model will improve prediction accuracy [22]. However, it is well known that in some jurisdictions, using different models either explicitly or implicitly for different protected groups is not allowed [22]. In this paper, we develop a technique that enables us to optimize accuracy and fairness while improving interpretability of the model, without explicitly using any of the sensitive information in the functional form of our model. We propose the use of Multitask Learning (MTL), which has been proven to improve interpretablity in models that use multidimensional health data [42], and has been used in previous research to optimize both accuracy and fairness [22], along with Monte Carlo dropout to utilize model uncertainty to improve fairness without sacrificing either computational complexity or accuracy.

Our contribution can be summarized as

(i) analyses of bias based on different demographic information on data collected in the wild for anxiety prediction,

(ii) development of an MTL-based bias mitigation technique to optimize fairness while preserving accuracy and improving interpretability,

(iii) evaluation of the proposed method on LSTM-based models and a large scale of public biobehavioral dataset with various demographics, and

(iv) comparison of our method against a conventional method, reweighting introduced by [24] for both accuracy and fairness measures.

II. Data and Methods

A. Dataset

In our experiments, we used a part of the TILES dataset, physiological data collected by Mundnich et al. [20] from over 200 hospital workers. We used 10 weeks worth of electrocardiogram (ECG) data collected with the OMSignal smart garment (15-second long ECG signal in 250 Hz every 5 minutes). We extracted 25 frequency and time-domain ECG features, such as the statistical characteristics of the peak-to-peak intervals and the energy of the signal in various frequency bands. Table I shows the full list of the extracted tables, and the definition of each features can be referred to [4]. We made use of self-reported anxiety levels which were measured using the State Trait Anxiety Inventory giving a value in the range 20 to 80 for each participant. We further binarized these scores using personalized z-score. The original released data contains anxiety labels that were reported in a 5-point scale by subjects. Following the setting in [10], we calculated the z-score of labels for each participant separately, and marked labels below the personalized average as negative (0) and labels above the mean as positive (1). We used 2 hours of the data (5 minutes x 24 steps) to infer upcoming anxiety labels. Under this scenario, we have in total 920 samples with 506 negative samples and 414 positive samples.

TABLE I
LIST OF EXTRACTED FEATURES

Category	Features
Time Domain	mean_nni, sdnn, sdsd, nni_50, pnni_50, nni_20, pnni_20,rmssd, median_nni, range_nni, cvsd, cvnni, mean_hr, max_hr, min_hr, std_hr
Frequency Domain	lf, hf, lf_hf_ratio, lfnu, hfnu, total_power, vlf
Other	cardiac sympathetic index, cardiac vagal index

We used gender, age, race, income, shift, ethnicity, born in the US, English as a native language, and work hours as the demographic data (referred to as the 9 protected labels in this paper) to test our model fairness on. These data were collected via surveys administered to the participants during the study. We binary-encoded each of the protected labels to assign them as privileged and unprivileged classes to effectively analyze them using our chosen fairness metrics. We chose the class with the higher number of participants as the privileged class (denoted by 1), and the other with less participants as the unprivileged class (denoted by 0). Figure 1 below shows the distribution of the participants in our data based on the different demographic groups. We split the data into 75% training and 25% testing sets to fit into the model.

Fig. 1. Distributions of Demographic Data in the TILES dataset

B. Fairness Terminology

In this section, we introduce and define some of the concepts and terminologies generally used in algorithmic fairness research that are relevant to this paper.

- **Protected (sensitive) label**: An attribute that partitions a population into groups whose outcomes should have parity (such as race, gender, income, etc.).
- **Privileged class**: A protected label value indicating a group that is at an advantage.
- **Unprivileged class**: A protected label value indicating a group that is at a disadvantage.
- **Disparate impact ratio (DIR)**: This is a fairness metric which is the ratio of positive outcomes (anxiety=1) in the unprivileged class, divided by the ratio of positive outcomes in the privileged class, as shown in equation 1. It is the measure of how different outcomes are for different groups, based on the results of a model [9]. In this paper, we assume an acceptable lower bound of 0.8, and a higher bound of 1.2, with 1 being the the ideal score.

$$\text{DIR} = \frac{Pr(Y=1|D=unprivileged)}{Pr(Y=1|D=privileged)} \quad (1)$$

- **Equalized odds**: This fairness metric enforces that the model correctly identify the positive outcome at equal rates across groups, and misclassify the positive outcome at equal rates across groups (creating the same proportion of True Positives and False Positives across groups). In the context of this paper, we choose to compare false negative instead of true positive rates between the privileged and unprivileged classes because a negative outcome (anxiety=0) is more desirable in our study.

C. Multi-task Learning Based Bias Mitigation Method

Our proposed methodology is based on the premise of epistemic uncertainty in Bayesian uncertainty estimation. Epistemic uncertainty refers to uncertainty in the structure and parameters of a model caused by a lack of knowledge, which has also been proven to correlate with model weights [16], [35]. We hypothesize that when the model is most uncertain about the protected label, the weights of that model will lack knowledge at certain regions of the feature space related to the protected label, and therefore using these weights for anxiety prediction will minimize the influence of that protected label on the final anxiety prediction. Our proposed method will in turn minimize any bias that might be introduced to the model through imbalances in the data based on the protected label.

Figure 2 presents a visual diagram of our proposed method. We utilize multi-task learning to predict anxiety and one of the protected labels (e.g gender). We save the model periodically in our experiments, and implement Monte Carlo dropout which allows us to get uncertainty estimations of these saved models. Modeling uncertainty with Monte Carlo dropout works by running multiple forward passes through the model with a different dropout mask every time. Given a trained neural network model with dropout f_{nn}, to derive uncertainty for one sample x, we collect the predictions of T inferences with different dropout masks. Here, $f_{nn}^{d_i}$ represents the model with dropout mask d_i. So we obtain a sample of the possible model outputs for sample x as $f_{nn}^{d_0}(x), ..., f_{nn}^{d_T}(x)$.

By computing the average and variance of this sample, we obtain an ensemble prediction, which is the mean of the models posterior distribution for this sample and an estimate of the uncertainty of the model regarding x.

$$\text{predictive posterior mean: } p = \frac{1}{T}\sum_{i=1}^{T} f_{nn}^{d_i}(x) \quad (2)$$

$$\text{uncertainty: } c = \frac{1}{T}\sum_{i=1}^{T}[f_{nn}^{d_i}(x) - p]^2 \quad (3)$$

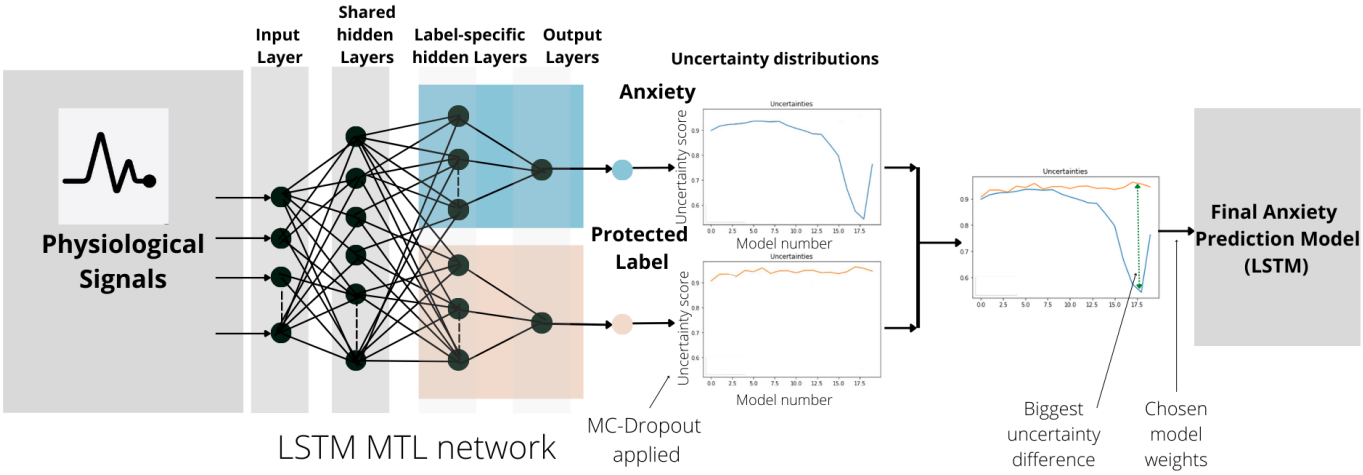

Fig. 2. Process Diagram

More information on this can be found in the original publication by Gal et. al [11].

From the distribution of Monte Carlo predictions obtained, we select the model prediction where the difference in uncertainties between anxiety and the protected label is highest, and then extract the parameters of the model at this desired point, and utilize them in our final anxiety prediction model.

III. Experiments

A. Analysis of Base Model

To start off our experiments, we calculated the disparate impact ratio of the original dataset for a combination of every protected label and anxiety, to understand if the data itself was biased in terms of the anxiety labels. Next, we ran a basic LSTM model for anxiety prediction and analyzed our results for fairness by calculating the disparate impact ratio and equalized odds for our prediction against all 9 protected labels. We conducted these analyses to determine whether and where the model was introducing bias, and ensure that we only apply our bias mitigation method on aspects of the data that are actually biased.

After performing fairness analysis on the base model, we selected the protected labels by which our model showed most bias, and tested our method on.

B. Method Implementation

To implement our method, we ran the MTL model with the structure shown in Figure 3 on a combination of anxiety plus each of the protected labels that showed bias in the base model, and we assigned loss weights of the ratio 4.5:0.5, 4.5 for anxiety and 0.5 for the demographic label. We used this ratio to ensure that anxiety prediction is prioritized by the model, making sure that its uncertainty score is kept lower than the demographic prediction. We arrived at this combination by experimenting with a few different combinations.

We ran these MTL networks for 100 epochs each, and saved the weights at every 5 epochs, making it a total of 20 weight

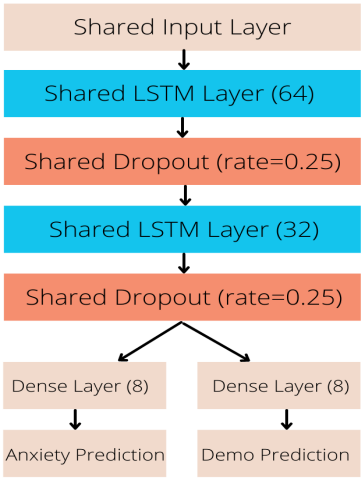

Fig. 3. Structure of Multi-talk Model

combinations, and we determined these parameters using a grid search. With this, we used Monte Carlo dropout, and ran some analysis on the Monte Carlo outputs to calculate the uncertainties of each of the saved weight combinations. We computed the uncertainties by calculating the sample variance of the different forward passes using the equation 3.

We plotted the uncertainties and identified the model that showed the biggest difference in uncertainty scores between anxiety and the demographic prediction.

C. Saliency Maps

To gain a better understanding of how the features influence our anxiety prediction and the different protected labels, and how our proposed method affects the models, we utilized a saliency map technique [28] to visualize the importance of model weights on our input time axis and features. According to the predicted class c, for example, the decision-making process of the model can be represented as $S_c(I) = w_c^T I + b_c$, where $I, S_c(I), w, b$ are the model input, output, weights and

bias, respectively. The essential of this method is to calculate the model weights on the input layer by gradients, e.g., $w = \frac{\partial S_c}{\partial I}|_{I_0}$. The calculated w represents the model saliency regrading to the input layer. In this study, we fetched the saliency maps for all the test samples and calculated the average saliency map to develop the intuition of important features and time steps in general.

IV. Results

A. Initial Analysis

After analyzing the dataset before running any models, the TILES dataset appeared to be relatively balanced in terms of the disparate impact ratio. Figure 4 shows the distribution of these scores and their proximity to the ideal score of 1. For most of the protected labels, the scores for the training data were in the range of 0.8 to 1.2, with the exception of income and shift with scores of 1.96 and 1.3 respectively. As for the testing data, the scores for race, shift, ethnicity, bornUS, and lang were in the range, while scores for gender, age income are below, and hours were above the range.

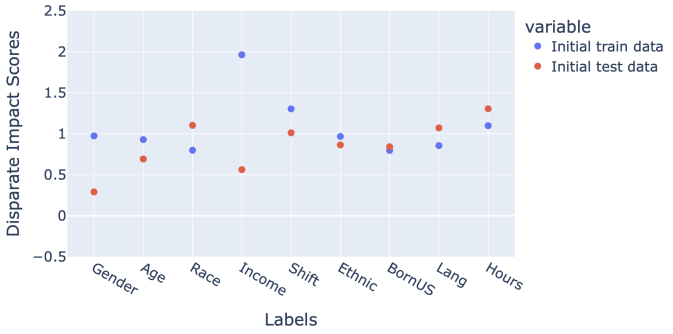

Fig. 4. Disparate Impact Scores of Initial Data

After running the base LSTM model, we obtained an accuracy score of 57.5% and an F1 score of 0.487. The baseline random performance for anxiety prediction was 49%, and the macro F1 score was 0.348.

A fairness analysis of the predictions showed us that the model does introduce some bias for some of the labels, specifically age, income, ethnicity, and bornUS, as shown in Figures 7 and 8. The scores for the other labels fall within the chosen range of 0.8 and 1.2.

Figure 5 shows the saliency map of anxiety produced by the base LSTM model. The x-axis represents the different features in our data, and the y-axis represents the 24 timesteps fed into the LSTM model to infer upcoming anxiety labels. From this figure, we can see that feature importance is relatively evenly distributed, with the most importance given to sdsd and pnni_20, which are time-domain-based heart rate variability features. In general, heart rate features (mean_hr, max_hr, min_hr, std_hr) seem to have the least importance.

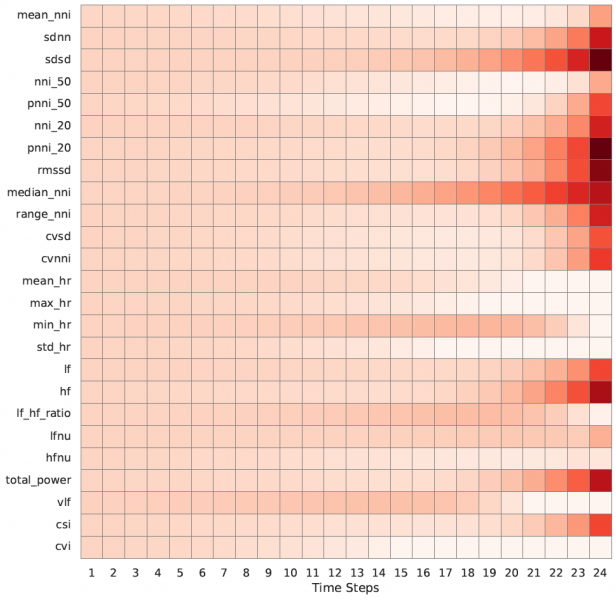

Fig. 5. Saliency Map of Base Anxiety Prediction (5 mins x 24 steps). The Y-axis indicates the feature names, where the X-axis represents the time steps of the sequences. Each time step represents the each feature point in the input sequences, and the number of the time steps is equal to the length of input sequences.

B. Proposed Method

Next, we implemented our method to predict anxiety while mitigating bias caused by age, income, ethnicity, and bornUS protected labels.

Figures 6, 7, and 8 show a comparison of performance (accuracy and F1) and fairness scores for the base model, our method, and an implementation of the reweighting method on our data. Reweighting involves applying appropriate weights to different tuples in the training dataset to make it discrimination free with respect to the protected label [24]. These figures show that our method was able to improve all the fairness scores, giving an ideal disparate impact ratio score of 1 for, and 0 for differences in False Positive and False Negative rates. This comes at a performance cost, as both our method and reweighting were not able to preserve the accuracy of the base model and F1 scores of the base model.

Figures 9 and 10 show the saliency maps for the age and bornUS experiments. Darker red color indicates more importance. The first map on each figure shows the feature importance on the demographic prediction when running the MTL model, the second shows that of anxiety from the MTL model before proposed method was implemented, and the third one for anxiety after method was fully implemented. Our method shifted weights from features that are important to the protected label to those that are less important. This ultimately decreased the importance of features associated with the protected label in the final anxiety prediction. For example, in Figure 9, the first map shows that features nni_20 and pnni_20 carried the most importance when it comes to the age label, and the second map shows that before our method

implementation, they also appeared to be very important for anxiety prediction. After our method was implemented, it is shown in the third map that these features carried less weight and ultimately are less important for anxiety prediction, which according to our hypothesis, reduced the possibility of the model being biased based on age.

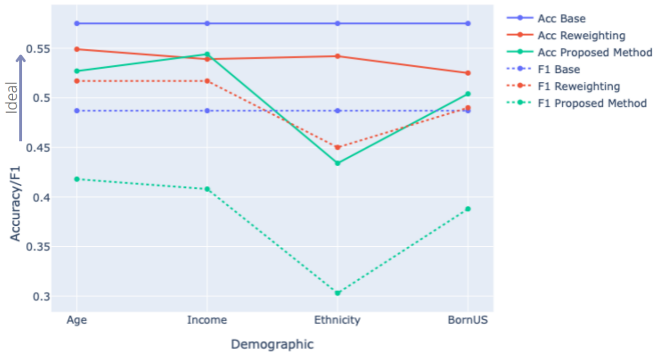

Fig. 6. Comparison of Model Accuracy and F1 Scores

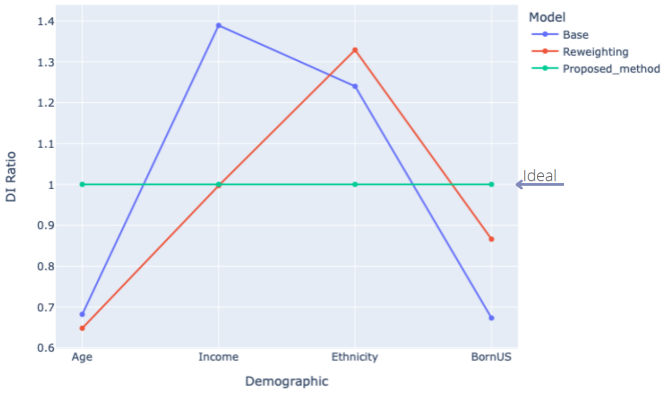

Fig. 7. Comparison of Disparate Impact Scores

V. Discussion and Conclusion

Seeing the ever growing effects of mental health on people's lives, and the gap between advances made in mental health prediction research, and development of effective bias analysis and mitigation methods, the goal of our study is to develop an effective technique to identify and mitigate bias in mental health prediction models. To do this, we analyzed bias in a physiological dataset, and an anxiety prediction model, and introduced a multi-task learning-based bias mitigation method where necessary. This method is based on the hypothesis that when a model is most uncertain about a particular protected label, the weights of that model will lack knowledge at certain regions of the feature space related to that label. Which means that using the weights of that model to predict anxiety will minimize the influence of the protected label on anxiety prediction. Our analysis found that the TILES dataset on its own was imbalanced by number of participants based on gender, age, race, income, shift, ethnicity, number of work hours, and whether or not the participant was born in the

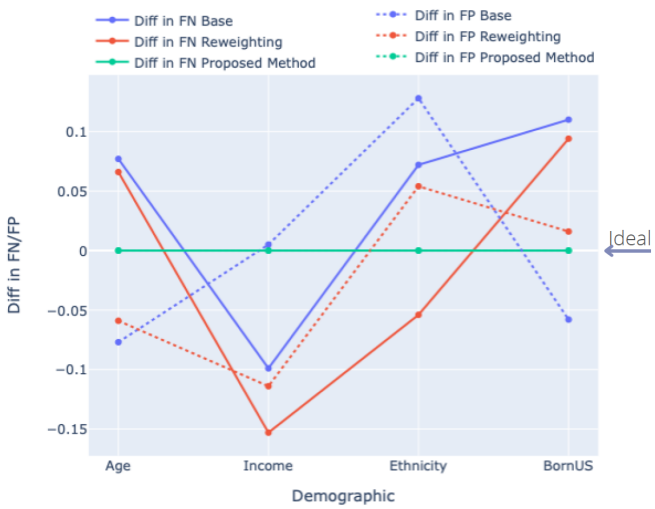

Fig. 8. Comparison of Differences in FN and FP Rates

US (bornUS), but it was not biased by these demographics when analyzed using the disparate impact ratio as a fairness metric. After running an LSTM model to predict anxiety, and analyzing the fairness in the prediction results, we discovered that the model introduced some bias by age, income, ethnicity, and bornUS.

We implemented our method to mitigate bias by each of these protected labels, and compared our results to a standard reweighting bias mitigation technique. Our results show that our method overall did better than reweighting on the fairness metric scores, but was not able to preserve the anxiety model performance. There was an average of 7% drop in accuracy, across all four experiments (age, income, ethnicity, bornUS), with the experiment on ethnicity causing the most depreciation by 14%.

To better understand the effect of different features on prediction, and how they are related to anxiety, and the different protected labels, we utilized a saliency map technique to visualize the importance of model weights on different prediction tasks. We found feature importance to be relatively evenly distributed for anxiety prediction but most correlated with 2 time-domain-based heart rate variability features (sdsd and pnni_20). Heart rate variability is the fluctuation of heart period over time, commonly measured by ECG, and is an important marker of psychological well-being [3]. One study found anxiety disorders to correspond with lower heart rate variability (HRV) [3], confirming why these features carry the most weight for anxiety prediction.

The saliency maps for the protected labels also gave us some interesting insight. We found that age correlated most with time domain-based HRV features (pnni_20, nni_20, sdsd, sdnn, cvi). Research has shown connections between HRV and a number of demographic factors such as age, race, ethnicity and gender or sex [6], [32], [41]. This was confirmed by the high correlation between HRV features with bornUS and ethnicity as well. Income on the other hand, correlated with

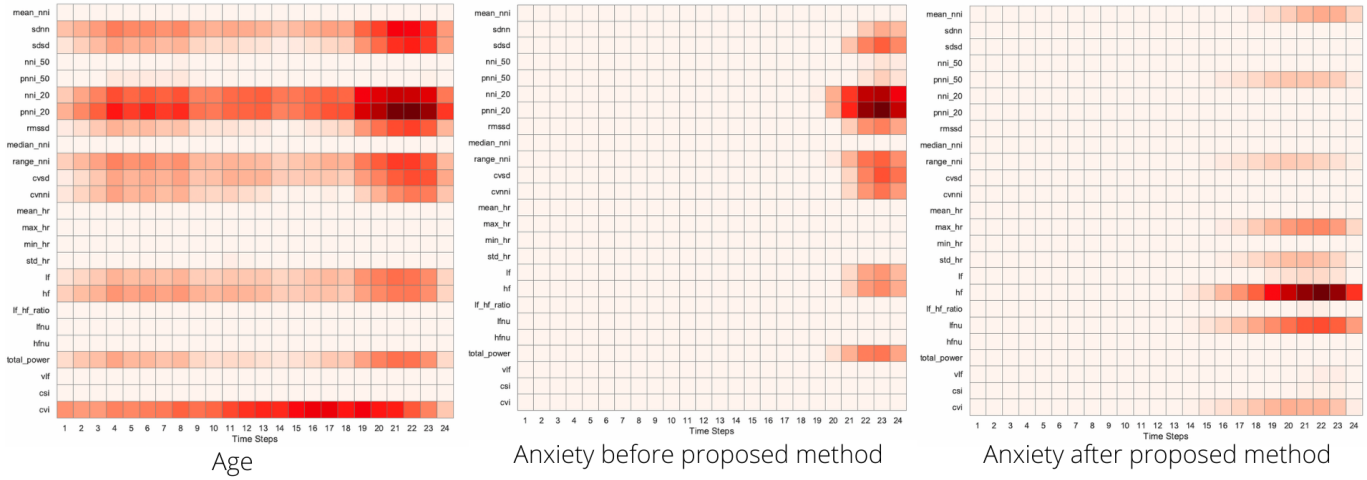

Fig. 9. Saliency Maps of Age Experiments

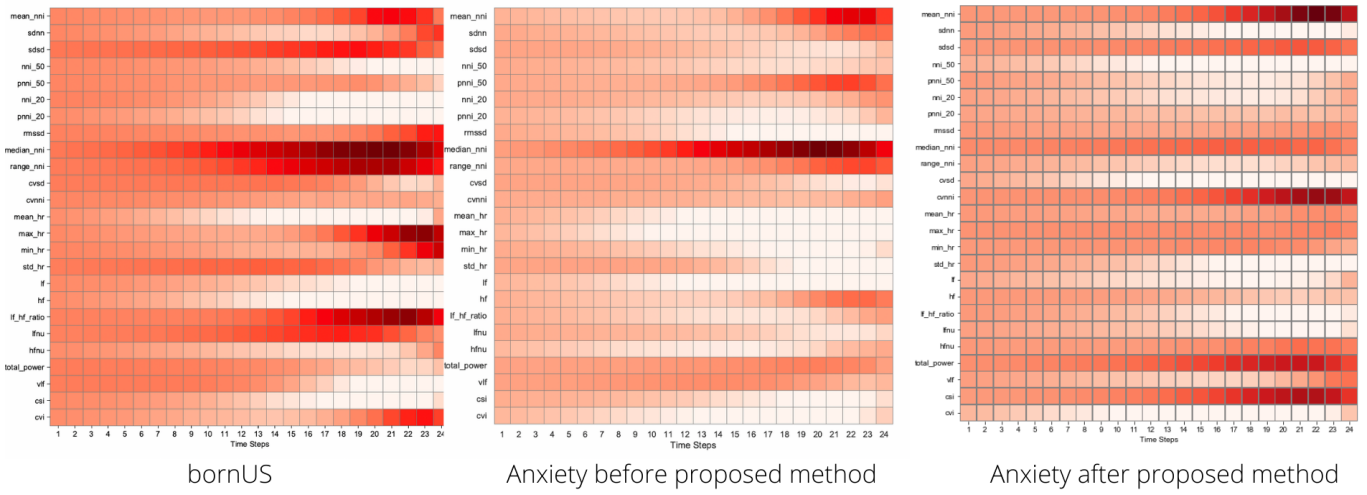

Fig. 10. Saliency Maps of bornUS Experiments

both frequency domain and time domain-based HRV features. Seeing as the considered protected labels and anxiety have high correlations with certain HRV features, it is easy to see how bias can be introduced to the model based on these protected labels.

The saliency maps also showed how our method shifted importance from features that highly correlate with a particular protected label, in order to reduce it's effect on the final anxiety prediction. Having stated that, we attribute our method's lack of ability to preserve accuracy and F1 scores to the fact that both anxiety and the protected labels have similar important features, which means that shifting weights from these features will undoubtedly have a significant effect on the model's ability to form a pattern for anxiety, ultimately affecting it's prediction. This method will benefit from further testing using datasets with more varying label-feature relationships.

In conclusion, it is important to acknowledge that data sources represent just one aspect of bias, which can be introduced through a model and certain technical metrics as well. The purpose of prediction algorithms is to influence the clinical decision-making process, and the biases of those developing and using them would have greater impact on whether they ultimately perpetuate inequality. For this reason, it is imperative to take into consideration, the impact of both human behavior and technical details when developing algorithms that make critical decisions such as those on mental health, especially in clinical settings.

Working with deep learning models for our prediction task, we have encountered a few limitations during the course of our study. First is the computational complexity of using monte carlo dropout. Using this method along with the aspect of saving and loading model weights results in a higher than normal computational cost. In the future, we aim to test our bias mitigation method on other datasets both physiological and non-physiological, and explore different ways to better preserve accuracy with it. We will explore developing a custom loss function that will utilize the relationship between uncertainty and weights to produce better performance.

REFERENCES

[1] F. Albertetti, A. Simalastar, and A. Rizzotti-Kaddouri. Stress detection with deep learning approaches using physiological signals. In *International Conference on IoT Technologies for HealthCare*, pages 95–111. Springer, 2020.

[2] D. Casacuberta and J. Vallverdú. Biases in assigning emotions in patients due to multicultural issues. In *Handbook of Artificial Intelligence in Healthcare*, pages 215–228. Springer, 2022.

[3] J. A. Chalmers, D. S. Quintana, M. J. Abbott, A. H. Kemp, et al. Anxiety disorders are associated with reduced heart rate variability: a meta-analysis. *Frontiers in psychiatry*, 5:80, 2014.

[4] R. Champseix. Heart rate variability analysis. https://github.com/Aura-healthcare/hrv-analysis, 2018.

[5] Y. Chen, X. Yang, T.-J. Cham, and J. Cai. Towards unbiased visual emotion recognition via causal intervention. *arXiv preprint arXiv:2107.12096*, 2021.

[6] J.-B. Choi, S. Hong, R. Nelesen, W. A. Bardwell, L. Natarajan, C. Schubert, and J. E. Dimsdale. Age and ethnicity differences in short-term heart-rate variability. *Psychosomatic medicine*, 68(3):421–426, 2006.

[7] R. Dai, C. Lu, L. Yun, E. Lenze, M. Avidan, and T. Kannampallil. Comparing stress prediction models using smartwatch physiological signals and participant self-reports. *Computer Methods and Programs in Biomedicine*, 208:106207, 2021.

[8] A. Domnich and G. Anbarjafari. Responsible ai: Gender bias assessment in emotion recognition. *arXiv preprint arXiv:2103.11436*, 2021.

[9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[10] A. Gaballah, A. Tiwari, S. Narayanan, and T. H. Falk. Context-aware speech stress detection in hospital workers using bi-lstm classifiers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8348–8352. IEEE, 2021.

[11] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[12] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane. Gender de-biasing in speech emotion recognition. In *INTERSPEECH*, pages 2823–2827, 2019.

[13] A. Howard, C. Zhang, and E. Horvitz. Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pages 1–7. IEEE, 2017.

[14] M. A. Jalal, R. Milner, T. Hain, and R. K. Moore. Removing bias with residual mixture of multi-view attention for speech emotion recognition. In *Interspeech 2020*, pages 4084–4088. ISCA-International Speech Communication Association, 2020.

[15] O. Kara, N. Churamani, and H. Gunes. Towards fair affective robotics: Continual learning for mitigating bias in facial expression and action unit recognition. *arXiv preprint arXiv:2103.09233*, 2021.

[16] Y. Li, S. Rao, A. Hassaine, R. Ramakrishnan, D. Canoy, G. Salimi-Khorshidi, M. Mamouei, T. Lukasiewicz, and K. Rahimi. Deep bayesian gaussian processes for uncertainty estimation in electronic health records. *Scientific reports*, 11(1):1–13, 2021.

[17] M. D. McCradden, S. Joshi, M. Mazwi, and J. A. Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, 2020.

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. arxiv 2019. *arXiv preprint arXiv:1908.09635*, 2019.

[19] H. F. Menezes, A. S. Ferreira, E. T. Pereira, and H. M. Gomes. Bias and fairness in face detection. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 247–254. IEEE, 2021.

[20] K. Mundnich, B. M. Booth, M. l'Hommedieu, T. Feng, B. Girault, J. L'hommedieu, M. Wildman, S. Skaaden, A. Nadarajan, J. L. Villatte, et al. Tiles-2018, a longitudinal physiologic and behavioral data set of hospital workers. *Scientific Data*, 7(1):1–26, 2020.

[21] L. Oneto and S. Chiappa. Fairness in machine learning. In *Recent Trends in Learning From Data*, pages 155–196. Springer, 2020.

[22] L. Oneto, M. Doninini, A. Elders, and M. Pontil. Taking advantage of multitask learning for fair classification." the 2019 aaai. In *ACM Conference (AIES)*, 2019.

[23] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 579–595, 2018.

[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*, 4(4):e213909–e213909, 2021.

[25] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[26] K. T. Rodolfa, H. Lamba, and R. Ghani. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3(10):896–904, 2021.

[27] H. R. Saloni Dattani and M. Roser. Mental health. *Our World in Data*, 2021. https://ourworldindata.org/mental-health.

[28] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[29] M. Srivastava, H. Heidari, and A. Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pages 2459–2468, 2019.

[30] Y. Su, B. Hu, L. Xu, H. Cai, P. Moore, X. Zhang, and J. Chen. Emotiono+: Physiological signals knowledge representation and emotion reasoning model for mental health monitoring. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 529–535. IEEE, 2014.

[31] S. Swati, M. Kumar, and S. Namasudra. Early prediction of cognitive impairments using physiological signal for enhanced socioeconomic status. *Information Processing & Management*, 59(2):102845, 2022.

[32] K. Umetani, D. H. Singer, R. McCraty, and M. Atkinson. Twenty-four hour time domain heart rate variability and heart rate: relations to age and gender over nine decades. *Journal of the American College of Cardiology*, 31(3):593–601, 1998.

[33] A. Valdivia, J. Sánchez-Monedero, and J. Casillas. How fair can we go in machine learning? assessing the boundaries of accuracy and fairness. *International Journal of Intelligent Systems*, 36(4):1619–1643, 2021.

[34] S. Vallor. Artificial intelligence and public trust. 2017.

[35] D. Wang, J. Yu, L. Chen, X. Li, H. Jiang, K. Chen, M. Zheng, and X. Luo. A hybrid framework for improving uncertainty quantification in deep learning-based qsar regression modeling. *Journal of cheminformatics*, 13(1):1–17, 2021.

[36] L. Xia, A. S. Malik, and A. R. Subhani. A physiological signal-based method for early mental-stress detection. *Biomedical Signal Processing and Control*, 46:18–32, 2018.

[37] T. Xu, J. White, S. Kalkan, and H. Gunes. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*, pages 506–523. Springer, 2020.

[38] S. Yan, D. Huang, and M. Soleymani. Mitigating biases in multimodal personality assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 361–369, 2020.

[39] H. Yu and A. Sano. Passive sensor data based future mood, health, and stress prediction: User adaptation using deep learning. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5884–5887. IEEE, 2020.

[40] K. Zanna, T. Neal, and S. Canavan. Clustering of physiological signals by emotional state, race, and sex. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 312–316, 2021.

[41] J. Zhang. Effect of age and sex on heart rate variability in healthy subjects. *Journal of manipulative and physiological therapeutics*, 30(5):374–379, 2007.

[42] Y. Zhang and Q. Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[43] M. J. Zvolensky, L. Garey, and J. Bakhshaie. Disparities in anxiety and its disorders, 2017.