Bias and Fairness in Multimodal Emotion Recognition Across Cultures

Advait Rane[1], Armaghan Asghar[1], and Nghi Le[1]

[1]Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA

1 Problem Definition

Machine learning algorithms increasingly play a bigger role in society's decision-making processes, for example, hireability assessment, personality assessment, recidivism prediction. The harm that a biased machine learning algorithm can cause, thus, becomes more widespread and dangerous. A well-known case of bias involves the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software that was used by courts to measure recidivism. A research found the software to be biased against African-Americans (Mattu (n.d.)). Hence, bias and fairness in machine learning is an important research direction going forward.

This project aims to study and quantify the biases learned by emotion recognition models on the multimodal cross-cultural SEWA dataset, and contribute a way to reduce biases in the models. We analyse the data to identify sources of bias the model may learn. We quantify bias and unfairness in the predictions of baseline deep learning emotion recognition models trained on different modalities. We further analyse the effects of late fusion on prediction bias and evaluate debiasing strategies at the late fusion step. Our work thus performs a holistic analysis of bias in the entire pipeline including data, model, and result fusion. The pipeline is illustrated in Figure 1.

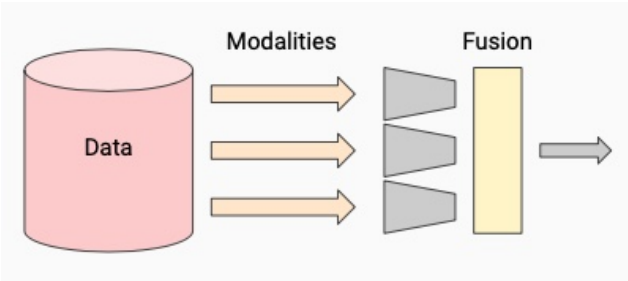


Figure 1: We analyse biases in the data, across modalities, and after late fusion.

2 Literature Review

2.1 Source of Biases in Machine Learning

The current literature identified three main sources of unfairness: biases introduced from the data to the algorithm (representation bias, sampling bias, etc.), biases introduced from the algorithm to the user (algorithmic bias, user interaction bias, etc.) and biases introduced by the user back into the data (historical bias, self-selection bias, etc.) (Mehrabi, Morstatter, Saxena, Lerman, and Galstyan (2022)). The three sources of biases constituted a feedback cycle. The biases from the data leads to biases in the model's outputs, which then leads to biases in the interaction between the user and the system, which finally generates new data that contains biases.

For our project, we are focusing on two types of biases: the biases inherent in the data and the biases arising after the machine learning model is trained on the data.

2.2 Abstract Fairness Definition and Fairness Metrics

Fairness has many definitions, some of which are adapted by ML research in different contexts (Hutchinson and Mitchell (2019)). There are several metrics defined to quantify fairness (Caton and Haas (2020), Mehrabi et al. (2022)). At present, there is no universal means to measure fairness and no guidelines on which measure is the "best". Authors in (Barocas, Hardt, and Narayanan (2019)) defined 3 abstract criteria for fairness. Consider S as the sensitive variable (group or cultural identifier), Y as the target variable, and R is the classification score produced by the classifier, the three abstract criteria of fairness are as follows:

- **Independence**: this criterion aims to make classifiers scores independently of the group membership:

$$R \perp S \tag{1}$$

  This criterion does not take into account the correlation between the group and the predicted variable. From another perspective, this can be considered unfair for some groups even though the criterion itself is satisfied. An example of a fairness metric that focus on this criterion is Statistical Parity (Caton and Haas (2020)).

- **Separation**: this criterion is an extension of the Independence criterion to address the unfairness concern above:

$$R \perp S \mid Y \tag{2}$$

  This criterion looks at the independence of the score and the sensitive variable conditional on the value of the target variable Y. An example of a fairness metric that targets this criterion is Equal Opportunity (Caton and Haas (2020)).

- **Sufficency**: this criterion looks at the independence of the target Y and the sensitive variable S conditional for a given score R:

$$Y \perp S \mid R \tag{3}$$

This criterion is related to calibration-based metrics (Barocas et al. (2019)).

In recent surveys, fairness metrics are typically divided into group based metrics and individual metrics. Group based metrics (Hutchinson and Mitchell (2019)) enforce fair outcomes across groups and are aligned with our work. We explore two types of these group-based metrics for our project:

- **Parity-based**: These metrics consider the predicted positive rates across groups (Dwork, Hardt, Pitassi, Reingold, and Zemel (2011), Caton and Haas (2020)). For example, statistical parity defines fairness as the equal probability of predicting the positive label across all groups:

$$\Pr(\hat{y} = 1 \mid g_i) = \Pr(\hat{y} = 1 \mid g_j) \tag{4}$$

  This metric targets the Independence fairness criterion. However, it does not consider inherent differences between groups.

- **Confusion-matrix based**: These metrics consider the true label. They compare aspects like True Positive Rate (Hardt, Price, Price, and Srebro (2016)) or Overall Accuracy (Berk, Heidari, Jabbari, Kearns, and Roth (2021)) across different groups. Hence, differences between groups are taken into account unlike parity-based metrics. An example is the Equal Opportunity metric:

$$\Pr(\hat{y} = 1 \mid y = 1 \& g_i) = \Pr(\hat{y} = 1 \mid y = 1 \& g_j) \tag{5}$$

  Under this metric, an algorithm is considered fair if the True Positive Rate is the same for both groups. This enforces Separation.

Furthermore, fairness is usually defined in classification but our problem is regression. Fair regression has been explored with convex (Berk et al. (2017)) and non-convex (Komiyama, Takeda, Honda, and Shimao (2018)) optimisation by adding certain regularization terms that incentivize fair outcomes.

2.3 Bias and Fairness in Affective Computing

Bias and fairness has been studied in ML applications for Affective Computing. (Li and Deng (2020)) explores dataset bias in facial expression recognition, and (Sagha, Deng, and Schuller (2017)) explores the effects of personal traits on speech valence recognition. (Yan, Huang, and Soleymani (2020)) identify biases in multi-modal personality assessment and explore de-biasing strategies. (Raghavan, Barocas, Kleinberg, and Levy (2020)) studies biases and debiasing strategies in algorithmic hiring based on actual vendor practices. Bias in multi-modal machine learning for assessing hireability from automated video interviews is detailed in Booth et al. (2021). We build on this literature by evaluating bias and fairness in emotion recognition across cultures.

2.4 De-biasing Strategies

There has been extensive research on addressing biases in machine learning. There are three categories of de-biasing methods (Mehrabi et al. (2022)):

- **Pre-processing**: Pre-processing the data to remove biases before training.
- **In-processing**: These methods reduce bias by modifying the model itself, like adding a regularizer to incentivize fairness (Berk et al. (2017)).
- **Post-processing**: Post-processing method try to reassign the labels given by the model.

In this project, we will explore a pre-processing method for de-biasing by dropping the feature sets that is most predictive of culture.

3 Dataset

SEWA is a database of annotated audio and 2D Visual dynamic behavior (Kossaifi et al. (2021)). It contains over 2000 minutes of audio-visual data of 398 participants belonging to six different cultural backgrounds, British, German, Hungarian, Greek, Serbian and Chinese. We restrict ourselves to the German and Hungarian cultures.

To collect the data, participants' standard webcams and microphones were used. Participants have been divided into pairs based on cultural background, age and gender. There were two experiments conducted to collect the data. Experiment Setup Part 1: Watching Adverts. Each person watched the same four advert videos each in 60 seconds length. The advert is chosen to elicit amusement, empathy, liking and boredom. After watching the advert, the person is asked to report his/her emotional state and sentiment.

Experiment Setup Part 2: While discussing Adverts in video Chat. After watching the 4th video, the volunteer pair had a 3 minute discussion with each other to elicit reactions and emotions about the advert and the advertised product. After the discussion each volunteer is asked to fill a questionnaire self-reporting his/her emotional state and sentiment towards the discussion.

The SEWA dataset has 198 recording sessions, with a total of 398 participants with a male/female ratio of 1.020. The participants are divided into 5 age groups, where participants aged 18-29 form the majority. The data has been annotated differently for each type of feature. Some have been labeled through manual annotation while others have been annotated semi-autonomously.

4 Methods

4.1 Feature Extraction

We followed the AVEC 2019 CES guidelines to extract features from the raw audio and video data in the SEWA database. We extract three types of features organised as follows:

- **Low Level Descriptors**: for visual LLDs, the intensities of 17 FAUs are extracted for each video frame using OpenFace (Baltrusaitis (2022)). Additionally a confidence measure, and descriptors for pose and gaze are extracted. For audio LLDs, OpenSMILE (*audeering/opensmile* (2022)) is used to extract two feature sets. These are the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features and MFCCs 1-13 (and their first- and second-order derivatives). The LLD features are further processed by calculating statistics over a sliding window.

- **Bag-of-Words**: OpenXBOW (openXBOW (2022)) is used to extract Bag-of-Word features from the LLD features. This provides us with three more modes of data.

- **Deep Representations**: The ResNet-50 (He, Zhang, Ren, and Sun (2015)) pretrained on the Affwild dataset (Zafeiriou et al. (2017)) is used to extract a 2048-dimensional visual feature vector for each frame. To extract audio features, DeepSpectrum (Amiriparian and Gerczuk (2022)) is used for two pretrained CNN models, the DenseNet (Huang, Liu, van der Maaten, and Weinberger (2018)) and the VGG16 (Simonyan and Zisserman (2015)). These models output a 1024- and a 4096-dimensional feature vector respectively.

4.2 Model

To evaluate bias and unfairness in model predictions, we use the baseline predictions provided in the AVEC 2019 CES. There are 10 baseline outputs - one for each extracted feature set and one that combines all modalities using late fusion with Support Vector Regression.

The predictions are made using a two layer LSTM-RNN model. For each emotion dimension (i.e., arousal, valence, liking) a dense layer outputting a value for each timestep is stacked on this base. Thus, the LSTM model solves a regression task over the three dimensions.

4.3 Fairness Metrics

We evaluate fairness in the baseline results using the following 4 metrics for group fairness:

- **Difference in Mean**: this metric measures the difference between the mean of the target variable of different groups. The lower the value the more fair the predictions are. This metric reflects the Independence criterion.

- **Difference in Concordance Correlation Coefficient**: this metric measures the difference in the CCCs between different groups. The lower the value the more fair the results are. This metric reflect the Separation criterion.

- **Difference in Mean Absolute Error**: The Mean Absolute Error of each group can also be compared similarly to the difference in CCC metric. This metric also reflects the Separation Criterion.

- **Jensen-Shannon Divergence**: The Jensen-Shannon divergence compares the difference in the distribution of the target variable between the two groups. This metric has more information than just comparing the mean and reflect the Independence criterion.

These metrics are calculated on the development set of the CES which has both the true values and the predicted values of the baseline LSTM.

4.4 Statistical Test Analysis

Student's T-test is also used in order to find differences in the feature means across different groups, which can indicate biases. We applied the T-test on the SEWA database training set. The motivation is to identify using p-value the data labels that can introduce bias in model training. From the SEWA database we have performed an analysis on the arousal, valuene and liking labels for the German and Hungarian Train dataset. Additionally, we do the same analysis for the extracted audiovisual LLD features.

4.5 Culture Predictivity of Modalities

We evaluate the culture predictivity of each feature set to quantify the model's ability to learn information about culture from the features. This approach is similar to the use of a random forest model to predict gender from the features in Booth et al. (2021). We use the same model described in section 4.2 to evaluate the biases it can learn. A dense layer is stacked on the 2 LSTM layers to predict between two culture classes.

To increase the amount of labelled data, we split the features recorded for an entire session into windows of 20 timesteps. Each of these windows has the label of that subject's culture. Thus, we predict a subject's culture from 20 timestep windows of a feature set.

Ideally, the model should not learn any biases about the cultures from the feature sets. Thus, the predictive accuracy of the model should be 50% ,i.e., the same

as random prediction. A predictive accuracy greater than 50% indicates that the model learns information about culture from the feature set.

4.6 Late Fusion Debiasing

Late Fusion combines the predictions of different modalities into a multimodal prediction using Support Vector Regression. We explore a debiasing technique at the late fusion step based on culture predictivity. We drop those feature sets that have a high culture predictivity and encode culture information.

We hypothesise that dropping highly predictive features will result in more fair predictions which will not be biased by the features that encode culture information. Thus the final multimodal predictions will be independent of the subject's culture. More specifically Late Fusion Debiasing by dropping highly culture predictive feature sets will give a better performance on the independence fairness metrics.

5 Results

5.1 Fairness Metrics

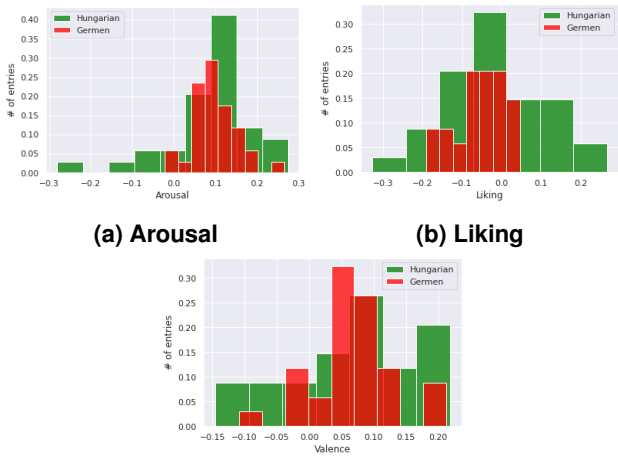
First, we evaluated the fairness in our development set using the four metrics detailed in Section 4.3. The results are summarized in Figure 2. Each graph in the figure is the results for a different metric. There are three bars group for three different emotions in every graph. And every bar represent a different feature sets.

As we can see from the Figure 2, the results are inconsistent. For example, for the MFCC feature set, when predicting Arousal, the MAE Difference is low while the CCC difference is high. This pattern can be seen elsewhere such as when predicting Arousal using DeepSpectrum features. For Mean Difference and JS Divergence graphs, the pattern also appears for eGeMAPS BOW feature set for Arousal prediction and others.

There are some possible reasons for this inconsistency. First, this could be due to the inherent limitation of these metrics; as in, they don't accurately capture fairness or biases in the data. Secondly, this could also mean that considering feature sets separately is not useful in detecting biases, as biases can emerged more clearly when combining different modalities together. A novel multi-modal fairness metric may be needed to capture such information.

5.2 Statistical Test Analysis

For the Valence, Liking and Arousal labels separated by culture, we found Valence label bias to be statistically insignificant with a p-value of 0.79, where as Liking and Arousal label bias were statistically significant with a p-vale < 0.05. The histograms in figure 3 show the labels averaged per subject for better visualization. These plots show that the label values for German subjects have a lower magnitude on average.



(a) Arousal (b) Liking

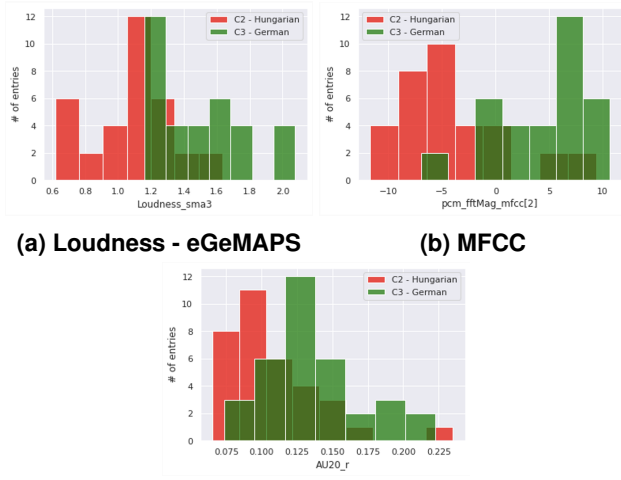
(c) Valence

Figure 3: Histograms of German and Hungarian labels. Arousal, Liking - significant; Valence - insignificant.

For the statistical analysis of extracted features from eGeMAPS, MFCC, and visual FAU activations we have identified multiple labels which are statistically significant.

Features	Statistically Significant
Visual Features	AU17_r, AU20_r, AU26_r
MFCC	pcm_fftMag_mfcc[1], pcm_fftMag_mfcc [2], pcm_fftMag_mfcc[3], pcm_fftMag_mfcc[4], pcm_fftMag_mfcc[6], pcm_fftMag_mfcc[7], pcm_fftMag_mfcc[8], pcm_fftMag_mfcc[9], pcm_fftMag_mfcc[12], pcm_fftMag_mfcc_de_de[2], pcm_fftMag_mfcc_de_de[5]
eGeMAPS	Loudness_sma3, alphaRatio_sma3, hammarbergIndex_sma3, slope500-1500_sma3, logRelF0-H1-H2_sma3nz, logRelF0-H1-A3_sma3nz,F1frequency_sma3nz, F1bandwidth_sma3nz

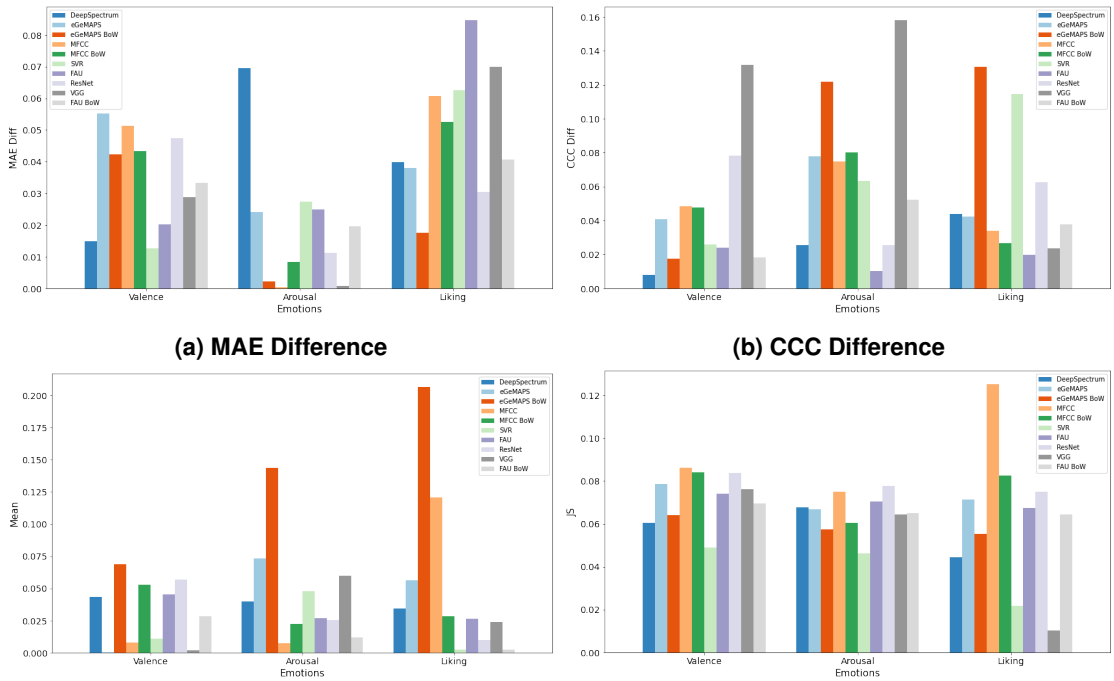
Figure 4: Statistically significant extracted labels from Visual and Audio Data



(a) Loudness - eGeMAPS (b) MFCC

(c) AU20 - Visual Feature

Figure 5: Histograms of German and Hungarian features. Loundess is in eGeMAPS, MFCC and AU20 in visual features are some of the significant labels



(a) MAE Difference **(b) CCC Difference**

(c) Mean Difference **(d) Jensen-Shannon Divergence**

Figure 2: Fairness metrics on different features set and across different emotions.

5.3 Cultural Predictivity of Modalities

We evaluate the classification accuracy of the LSTM culture prediction model for each modality. The accuracy across modalities are shown in figure 6.

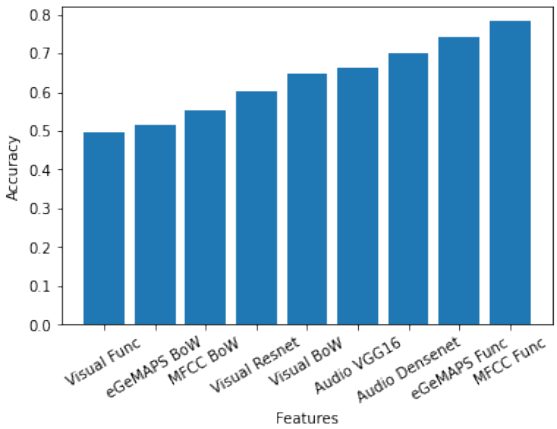


Figure 6: Classification accuracy of the LSTM culture prediction model on different feature sets.

Based on the accuracy values, we can split the modalities into three sets as follows-

- **Feature Set 1**: These feature sets are not predictive of culture. The accuracy value is 50-55% indicating that the model does not learn to classify culture from these features. This set includes the visual LLD features and the audio eGeMAPS BoW features.
- **Feature Set 2**: These feature sets are moderately predictive of culture. They achieve 55-65% accuracy indicating that the model learns to classify culture but not very well. This set includes the visual BoW features, the audio MFCC BoW features, and the deep visual Resnet features.
- **Feature Set 3**: These feature sets are highly predictive of culture. Models trained on these features can differentiate between cultures as seen by the accuracy value of 65-80%. This includes the audio MFCC features, audio eGeMAPs features, and audio DeepSpectrum features.

We learn that overall the audio feature sets encode more information about culture than the visual features. Furthermore, the visual deep features encode more information about culture than visual LLDs and audio deep features are highly predictive of culture. This indicates that deep learning features tend to be more biased. As many recent multi-modal approaches tend to turn towards deep features, the higher bias encoded by these model is a cause for concern.

5.4 Late-Fusion Debiasing

Late-Fusion Debiasing attempts to make predictions independent of culture. We primarily evaluate it using the independence fairness metrics - JS Divergence and Mean Difference.

Since feature set 3 is the most predictive followed by feature set 2 we first drop only feature set 3 and then drop feature sets 2 and 3. The plots for the fairness metric values can be seen in figure 7.

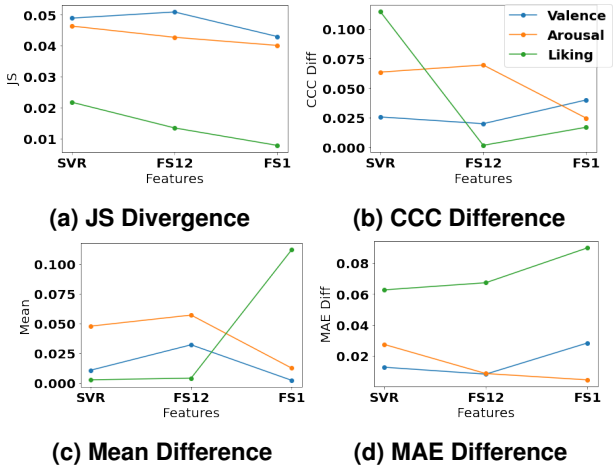


Figure 7: Late Fusion Debiasing Fairness metric performance. SVR: All feature sets, FS12: Feature sets 1 and 2, FS1: Only feature set 1.

We see a clear improvement in JS divergence (figure 7a) as we go from all feature sets, to sets 1 and 2, and finally only set 1. The results are more independent of culture membership as we drop highly predictive feature sets. Late-fusion debiasing successfully improves fairness in results with respect to the JS divergence. Results are not as unambiguous for the other independence metric, mean difference (figure 7c). Feature set 1 performs the best (lowest value) for valence and arousal. However for liking the value is already very low for SVR and it increases when we drop features.

The separation metric, CCC difference (figure 7b) also shows better performance on dropping highly culture predictive features for arousal and liking with a significant improvement in liking on dropping set 3. MAE difference 7d shows a consistent improvement when dropping predictive features for arousal but the contrary for valence and liking.

6 Conclusions

We quantified fairness in the baseline predictions using different modalities using four fairness metrics that reflect different fairness criteria. The results indicate that although the baseline model is trained on balanced data from both cultures, there is a disparity in performance for different cultures which varies by modality and metric. Thus, when solving a task it is crucial to pick a definition of fairness appropriate to that task and use the modalities which are unbiased by that definition. Furthermore, late-fusion might be helping improve fairness in the model predictions as these values are consistently better than the individual modalities in figure 2.

We employed the Student's T-test to identify bias in the training data which can be the source of biased predictions. The results show that there is significant label bias present for the arousal and liking labels, with labels for German subjects having a lower magnitude on average for both dimensions. We also identify several features that encode bias between the cultures in the LLD features. Generally, the audio features encode more information about culture than the visual features. Thus although audio features perform better on these tasks, they have the risk of producing biased results

This finding is reinforced by evaluating the culture predictivity of the features sets. Here too the audio LLD features are most predictive of culture followed by the audio deep features, and the visual LLD features are least predictive. Overall, the deep features tend to be more predictive of culture and should be used carefully for future multi-modal research. Finally, dropping highly culture predictive features during late-fusion led to an improved performance on the JS divergence metric. While our results across all four metrics are not as conclusive, with respect to JS divergence Late-Fusion Debiasing is a simple and effective method to reduce bias in model predictions.

7 Contributions

Advait Rane	Literature Review, Feature Extraction (LLDs, Deep Features), Culture Predictivity, Late-Fusion Debiasing
Nghi Le	Fairness Metrics Analysis, Literature Review, Experimentation with Various Debiasing Methods
Armaghan Asghar	Dataset Exploration, Feature Extraction(BoW), Statistical Test Analysis

Table 1: Teammate contributions

References

Amiriparian, S., & Gerczuk, M. (2022, February). *Deepspectrum github*. GitHub. Retrieved 2022-05-07, from https://github.com/DeepSpectrum/DeepSpectrum (original-date: 2018-05-30T18:17:58Z)

audeering/opensmile. (2022, May). audEERING GmbH. Retrieved 2022-05-07, from https://github.com/audeering/opensmile (original-date: 2020-10-15T13:44:10Z)

Baltrusaitis, T. (2022, May). *OpenFace 2.2.0: a facial behavior analysis toolkit.* Retrieved 2022-05-07, from https://github.com/TadasBaltrusaitis/OpenFace (original-date: 2016-03-05T20:08:49Z)

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. , 253.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., ... Roth, A. (2017, June). A Convex Framework for Fair Regression. *arXiv:1706.02409 [cs, stat]*. Retrieved 2022-05-04, from http://arxiv.org/abs/1706.02409 (arXiv: 1706.02409)

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021, February). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, *50*(1), 3–44. Retrieved 2022-05-04, from https://doi.org/10.1177/0049124118782533 (Publisher: SAGE Publications Inc) doi: 10.1177/0049124118782533

Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021, November). Integrating Psychometrics and Computing Perspectives on Bias and Fairness in Affective Computing: A case study of automated video interviews. *IEEE Signal Processing Magazine*, *38*(6), 84–95. (Conference Name: IEEE Signal Processing Magazine) doi: 10.1109/MSP.2021.3106615

Caton, S., & Haas, C. (2020, October). Fairness in Machine Learning: A Survey. *arXiv:2010.04053 [cs, stat]*. Retrieved 2022-05-04, from http://arxiv.org/abs/2010.04053 (arXiv: 2010.04053)

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011, November). Fairness Through Awareness. *arXiv:1104.3913 [cs]*. Retrieved 2022-05-04, from http://arxiv.org/abs/1104.3913 (arXiv: 1104.3913)

Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems* (Vol. 29). Curran Associates, Inc. Retrieved 2022-05-04, from https://papers.nips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

He, K., Zhang, X., Ren, S., & Sun, J. (2015, December). Deep Residual Learning for Image Recognition. Retrieved 2022-05-07, from https://arxiv.org/abs/1512.03385v1 doi: 10.48550/arXiv.1512.03385

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018, January). Densely Connected Convolutional Networks. *arXiv:1608.06993 [cs]*. Retrieved 2022-05-07, from http://arxiv.org/abs/1608.06993 (arXiv: 1608.06993)

Hutchinson, B., & Mitchell, M. (2019, January). 50 Years of Test (Un)fairness: Lessons for Machine Learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58. Retrieved 2022-05-04, from http://arxiv.org/abs/1811.10104 (arXiv: 1811.10104) doi: 10.1145/3287560.3287600

Komiyama, J., Takeda, A., Honda, J., & Shimao, H. (2018, July). Nonconvex Optimization for Regression with Fairness Constraints. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2737–2746). PMLR. Retrieved 2022-05-04, from https://proceedings.mlr.press/v80/komiyama18a.html (ISSN: 2640-3498)

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., ... Pantic, M. (2021, March). SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(3), 1022–1040. Retrieved 2022-05-05, from http://arxiv.org/abs/1901.02839 (arXiv: 1901.02839) doi: 10.1109/TPAMI.2019.2944808

Li, S., & Deng, W. (2020). A Deeper Look at Facial Expression Dataset Bias. *IEEE Transactions on Affective Computing*, 1–1. Retrieved 2022-05-04, from http://arxiv.org/abs/1904.11150 (arXiv: 1904.11150) doi: 10.1109/TAFFC.2020.2973158

Mattu, L. K. S. J. A., Jeff Larson. (n.d.). *Machine Bias.* Retrieved 2022-05-06, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=gl4jHLt-6ZxkcB55q8h_B25ydpK2Tm56

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022, January). A Survey on Bias and Fairness in Machine Learning. *arXiv:1908.09635 [cs]*. Retrieved 2022-05-04, from http://arxiv.org/abs/1908.09635 (arXiv: 1908.09635)

openXBOW. (2022, April). *openXBOW/openXBOW*. Retrieved 2022-05-07, from https://github.com/openXBOW/openXBOW (original-date: 2016-05-21T23:32:39Z)

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020, January). Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481. Retrieved 2022-05-04, from http://arxiv.org/abs/1906.09208 (arXiv: 1906.09208) doi: 10.1145/3351095.3372828

Sagha, H., Deng, J., & Schuller, B. (2017, October). The effect of personality trait, age, and gender on the performance of automatic speech valence recognition. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 86–91). (ISSN: 2156-8111) doi: 10.1109/ACII.2017.8273583

Simonyan, K., & Zisserman, A. (2015, April). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*. Retrieved 2022-05-07, from http://arxiv.org/abs/1409.1556 (arXiv: 1409.1556)

Yan, S., Huang, D., & Soleymani, M. (2020, October). Mitigating Biases in Multimodal Personality Assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 361–369). New York, NY, USA: Association for Computing Machinery. Retrieved 2022-05-03, from https://doi.org/10.1145/3382507.3418889

Zafeiriou, S., Kollias, D., Nicolaou, M. A., Papaioannou, A., Zhao, G., & Kotsia, I. (2017, July). Aff-Wild: Valence and Arousal 'In-the-Wild' Challenge. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1980–1987). Honolulu, HI, USA: IEEE. Retrieved 2022-05-07, from http://ieeexplore.ieee.org/document/8014982/ doi: 10.1109/CVPRW.2017.248