Analyzing Biases in AU Activation Estimation Toward Fairer Facial Expression Recognition

Miguel Monares
mmonares@ucsd.edu
UC San Diego
La Jolla, CA, USA

Yuan Tang
yutang@ucsd.edu
UC San Diego
La Jolla, CA, USA

Ritik Raina
rraina@ucsd.edu
UC San Diego
La Jolla, CA, USA

Virginia R. de Sa
desa@ucsd.edu
UC San Diego
La Jolla, CA, USA

ABSTRACT

Facial expression recognition plays a prominent role in numerous applications, from emotion detection to human-computer interaction. However, these models are often subject to different biases. This study explores potential racial biases in facial expression analysis using synthetically generated faces. We specifically investigate disparities in the performance of an action unit estimation network across different skin tones. This research highlights the presence of skin color biases in an action unit estimation network and demonstrates the impact and importance of dataset diversity and variety in achieving robust models. Furthermore, we show that these biases vary across different action units and skin tones and these model biases interact with the biases caused by dataset differences. This work is an important step towards the eventual goal of understanding the basis of these combined biases and removing them from facial expression models.

CCS CONCEPTS

• **Computing methodologies** → *Computer vision.*

KEYWORDS

facial expression recognition, affective computing, computer vision

ACM Reference Format:
Miguel Monares, Yuan Tang, Ritik Raina, and Virginia R. de Sa. 2023. Analyzing Biases in AU Activation Estimation Toward Fairer Facial Expression Recognition. In *Proceedings of August 6-10, 2023 (KDD '23).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Facial expression recognition (FER), an integral component of affective computing, is increasingly shaping a myriad of applications across diverse domains today. Such models discern facial expressions to interpret emotions, which has been useful in developing intelligent systems capable of automated analysis of human emotion and experience. For example, FER models are harnessed in healthcare to assist in the monitoring of patient care and diagnosis of medical or behavioral conditions [5]. In advertising and consumerism, FER models enable real-time customer sentiment analysis, providing valuable insights into product and service reception [3]. In education, models help facilitate personalized learning by identifying student engagement, boredom, or confusion [4]. Further, user engagement and interaction in sectors such as entertainment, gaming, and interpersonal gaming are being transformed through immersive, responsive, and emotionally intelligent experiences [7].

Given the widespread employment of automated FER, it is critical that models perform consistently and equitably across diverse populations. This consistency demands fairness in recognizing expressions in faces across different, genders, ages, ethnicities, and skin colors. However, existing public FER models demonstrate biases in the faces of diverse populations. Raina et al. [11] revealed racial biases in several publicly available models using synthetically generated faces. They revealed that these models for both emotion and action unit detection were biased across skin color and facial morphology. Fabi et al. [2] used artificially generated faces to explore racial biases in pain-related facial expressions using a pain-estimation model [12]. They revealed that the network's activation of facial AUs was subject to different biases in performance for different skin colors and races and that these biases were not solely better for the faces of the majority race and skin color.

In this work, we analyze the biases and performance of a facial action unit activation network from a computer vision pain estimation model [12]. Our method involves the generation of an artificial facial expression dataset, which enables precise control over facial parameters to isolate the impact of distinct manipulations on our model under evaluation and better understand the nuanced dependencies and biases within FER models.

This study serves two primary functions. Firstly, it conducts a targeted investigation of skin color biases in an AU Estimation network using synthetic faces, revealing the presence of skin color biases and highlighting the complexity and non-linearity of such biases. Secondly, it studies the impact of skin color distribution in the training set, highlighting the importance of dataset diversity and distribution.

These insights are valuable in the pursuit of developing more fair, accurate, and empathetic AI systems in the future.

KDD '23, August 6-10, 2023, Long Beach, CA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2 METHODS

In this section, we describe the facial expression dataset we contribute and use for model training and evaluation. We also describe the model under evaluation.

2.1 Synthetic Facial Action Unit Dataset

Our work makes use of artificially generated images of faces using the Character Creator 4 (CC4) software, a platform for customizing and generating realistic character assets. We create a dataset of 940 facial expression images of a European male and female face varying in skin color and action unit activations.

The facial expressions of our synthetic faces were crafted by manipulating activation levels for various facial activation units (AUs), based on the Facial Action Coding Systems (FACS) [1]. FACS is a system that categorizes AUs on an anatomical basis, linked to specific facial muscle movements that result in perceptible changes in facial expressions.

We systematically adjusted the facial expressions using ten specific AUs, controlled by facial morphing options in CC4. These action units and their corresponding options in CC4 are described in Table 1. These AU mappings are simulated based on [1].

Table 1: Action Units and Intensity Levels

Action Unit	Description	CC4 Facial Morphing Options
AU4	Brow Lowerer	Brow Drop L/R (30, 60, 90, 120, 150)
AU6	Cheek Raiser	Cheek Raise L/R (30, 60, 90, 120, 150)
AU7	Lid Tightener	Eye Squint L/R (30, 60, 90, 120, 150)
AU9	Nose Wrinkler	Nose Sneer L/R (30, 60, 90, 120, 150)
AU10	Upper Lip Raiser	Nose Nostril Raise L/R (30, 60, 90, 120, 150), Nose Crease L/R (20, 40, 60, 80, 100), Mouth Shrug Upper (30, 60, 90, 120, 150)
AU12	Lip Corner Puller	Mouth Smile L/R (30, 60, 90, 120, 150)
AU20	Lip Stretcher	Mouth Stretch L/R (30, 60, 90, 120, 150)
AU25	Lips Part	Mouth Shrug Upper (16, 32, 48, 64, 80), Mouth Drop Lower (16, 32, 48, 64, 80)
AU26	Jaw Drop	Jaw Open (10, 20, 30, 40, 50)
AU43	Eyes Closed	Eye Blink L/R (100)

For each face, only one action unit is activated, with the others remaining unactivated. The process of generating facial expressions is repeated for 10 different skin colors and for 2 genders, male and female. The 10 skin tones are derived from the Monk Skin Tone Scale [8], a skin tone scale that aims to be more representative and inclusive of a broader spectrum of skin tones toward better representation in datasets and ML models. The total number of samples is 940 = (9 AUs * 5 intensities + 1 AU43 + 1 face no activated AUs) * 10 skin tones * 2 genders).

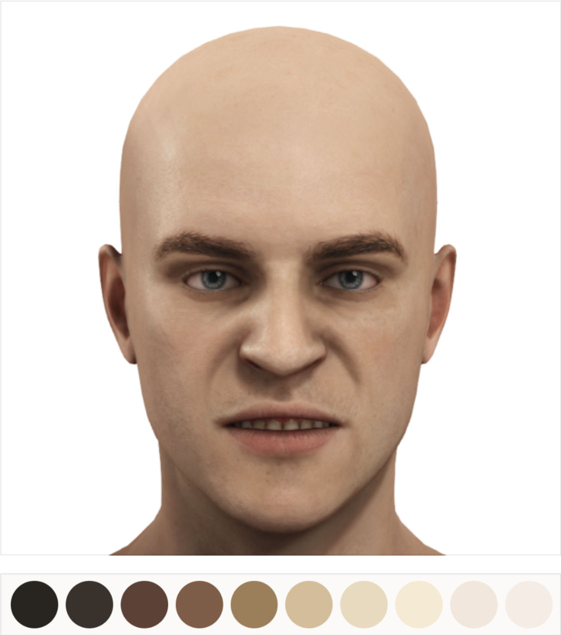

Figure 1: Top: Sample Face Created by CC4 from the Synthetic FAU Dataset: European Male, Skin Tone 4, AU10 Max Activated. Below: 10 Monk Skin Tone Scale (10 darkest, 1 lightest, scale taken from [8])

2.2 Extended MTL Model for Pain-Estimation

The model we use to conduct our experiments and investigations follows the research done by [2] and [11], investigating the performances and biases of the initial phase of the Extended Multi-Task Learning (MTL) pain estimation neural network of Xu et al. [12].

This model achieves state-of-the-art accuracy on the UNBC-McMaster Shoulder Pain Expression dataset [6]. This dataset is publicly accessible and comprises 200 face videos from 25 patients experiencing varying levels of shoulder pain during different movements. Each frame of these videos was annotated with 11 facial AU intensities, a corresponding PSPI score, and 66 AAM landmarks. The Prkachin and Solomon Pain Intensity score [10], also known as PSPI, is a pain metric derived from a unique combination of pain-related AU intensities. This metric is defined as:

$$PSPI = AU4 + max(AU6, AU7) + max(AU9, AU10) + AU43$$

The first stage of this tri-phased model is built on top of the VGGFace network [9], pre-trained on classifying 2622 faces of largely Caucasian celebrities. This network was further trained on the UNBC-McMaster Shoulder Pain Dataset, tuning the network to detect and score 10 pain-related AUs (4, 6, 7, 9, 10, 12, 20, 25, 26, 43) and determine the PSPI score of a facial image frame. The extended MTL model includes two additional stages that predict whole-video segment pain scores based on the output from the first stage. However, our experiments solely utilize the first stage of this model, which we’ll refer to as the “AU Estimator” model henceforth. This model is publicly available.

3 EXPERIMENTS

In this section, we describe our experiments aimed at uncovering skin color biases in the AU Estimation network and understanding how representation in training datasets impacts the network's behavior.

3.1 Exploring Color Bias with Paired T-Tests

To investigate potential biases in our AU Estimator model, we designed an experiment that deliberately focuses on differences across skin tones. Our synthetic faces maintain identical attributes in all aspects such as facial expression, morphology, and pose, varying only in skin color. Therefore, if our model is free from bias, it should assign equivalent AU activation and, consequently, PSPI scores across different skin tones.

To conduct our experiment, we first trained our AU Estimator on a subset of our synthetic faces, excluding the faces with skin tone 1 and skin tone 10. We then run the faces with skin tones 1 and 10 through the tuned AU Estimator model and employ paired two-sided t-tests on the model's outputs for each AU and PSPI. The input images for training and testing are reshaped to 256×256, center-cropped to 224×224, and their color channels are adjusted to match the VGGFace network. Given the controlled nature of our synthetic faces (identical in all aspects except skin color), any statistically significant difference in the output of the model can be ascribed to a potential skin color bias. The results of our paired t-tests can be seen in Table 2.

Table 2: Paired T-Tests of Skin Tone 1 vs Skin Tone 10 AU Activation and PSPI Estimation

Column	p-value	t-statistic
PSPI	5.1595e-19	-14.713
AU4	4.72062e-07	-5.85827
AU6	1.11425e-06	-5.60835
AU7	1.03877e-16	-12.7535
AU9	0.00455051	-2.98338
AU10	0.00135073	-3.41276
AU12	0.112856	-1.61635
AU20	0.346493	-0.951162
AU25	0.341532	-0.96109
AU26	1.71813e-09	7.48454
AU43	5.19768e-15	11.4107

Our results reveal a statistically significant difference between the faces of skin tones 1 and 10, indicative of a skin color bias. Specifically, the p-values associated with PSPI, AU4, AU6, AU7, AU9, AU10, AU26, and AU43 are all below the significance threshold of 0.05. This suggests that there are statistically significant differences in the model's estimation of these measures between the two skin tones. In particular, the substantially low p-values associated with PSPI, AU4, AU6, AU7, AU26, and AU43 demonstrate a strong level of statistical significance, further highlighting the potential bias in the model's outputs for these measures.

The t-statistics provide additional context. For instance, the t-statistics for PSPI (-14.713), AU7 (-12.7535), and AU43 (11.4107) are particularly high, indicating that the differences in the model's outputs for these measures are not only statistically significant but also practically significant, implying consistent disparities between skin tones in this testing environment.

3.2 Skin Color Biases in AU Activation

3.2.1 Profiling AU4 and AU10 Activation Biases.

To further investigate the skin color biases in the AU Estimation model, we run a comparative study on the model's performance in tracking the activation of AUs across skin tones. We choose AU4 and AU10 for their contrasting behavior in model estimation activation, demonstrated later. For this experiment, we employ three different models. The first model (ModelAll) is the model further trained on all of the European Male (EM) faces from skin tones 1 to 10.

For each skin tone in our European Women faces, we plot the model's predicted activation level against the true activation level of the AU. An unbiased performance would correspond to overlapping straight diagonal lines of slope 1, calculated by least squared regression, indicating that the predicted activation exactly matches the true activation for all skin tones. A deviation from this performance, especially if it systematically varies by skin tone, would indicate that the model's AU tracking performance is influenced by skin color, a suggestion of racial bias.

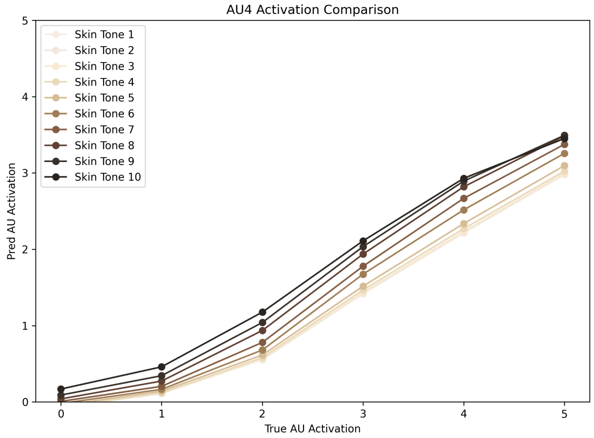

Figure 2: ModelAll AU Activation Comparison for AU4.

The results of ModelAll on AU4, brow lowerer, are presented in Figure 2, a plot of the Predicted vs True AU4 activation across the skin tones, and Table 3, showing the slope of the predicted AU4 activation for each skin tone, as well as the Mean Absolute Difference (MAD) between each tone's predicted and true AU activation.

Firstly, we observed a relatively consistent slope across all skin tones. This suggests that ModelAll is fairly consistent in tracking AU4 activation across different skin tones. Specifically, the slope values range from around 0.641 to 0.739, indicating that the model is somewhat effective at increasing its predicted activation level as the true activation level increases. However, the MAD between the model's estimations and the true AU4 activations increases as the skin tone get lighter. We observe in Figure 2 that the model's

Table 3: Slope and MAD of ModelAll AU4 Activation.

Skin Tone	Slope	MAD from True AU Activation
1	0.641013	1.300973
2	0.642171	1.294835
3	0.645249	1.288051
4	0.651103	1.267905
5	0.663155	1.224223
6	0.698193	1.121689
7	0.720832	1.033993
8	0.739004	0.931749
9	0.733068	0.881097
10	0.707294	0.840436

prediction for AU4 activation is higher for darker skins than lighter ones, suggesting a bias in ModelAll’s AU4 estimation.

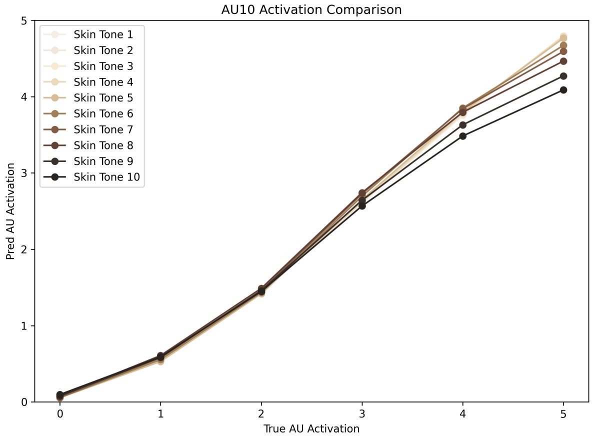

Figure 3: ModelAll AU Activation Comparison for AU10.

Table 4: Slope and MAD of ModelAll AU10 Activation

Skin Tone	Slope	MAD from True AU Activation
1	0.984719	0.322923
2	0.986277	0.320114
3	0.987638	0.320306
4	0.987191	0.317218
5	0.988527	0.315205
6	0.977794	0.308262
7	0.962636	0.310052
8	0.936256	0.329571
9	0.893738	0.415716
10	0.850137	0.486333

We run the same experiment of ModelAll on AU10, upper lip raiser, shown in 3 and Table 4.

As shown in Table 3, there is a slightly decreasing trend in the slope of the estimated activation as the skin tone darkens from 1 to 10, indicating that the model is less sensitive to AU10 activation as skin tone darkens. The MAD between the model’s estimations and the true AU activations also shows variance based on skin tone. More specifically, the MAD is noticeably greater in the estimation for the darkest skin tones (9 and 10), indicating a decrease in model accuracy for darker skin tones. Additionally, we notice that, at higher levels of AU10 activation, the model’s estimation of lighter faces is greater than estimations of darker faces. This is in contrast with the results of AU4, where the estimation of the AU4 intensity in darker faces was greater than in lighter ones. These contrasting results between the AUs speak to the complex nature of biases within the model, suggesting that the bias may not be uniformly distributed or predictable across different AUs or skin tones.

Next, we build on these experiments to explore the effects of training the models on select skin tone ranges. We focus on how model performance might be affected when trained exclusively on either lighter or darker faces, to observe how the training data’s skin tone distribution may influence the performance and potential biases of facial expression models.

The second model (ModelLighter) is the AU Estimator only tuned on lighter EM faces of skin tones 1 to 5. The third model (ModelDarker) is tuned only on darker skin tones 6 to 10. To test the models’ ability to track AU4 activation, we run the European Woman (EW) faces with AU4 activated from 0 to 5, across the range of skin tones.

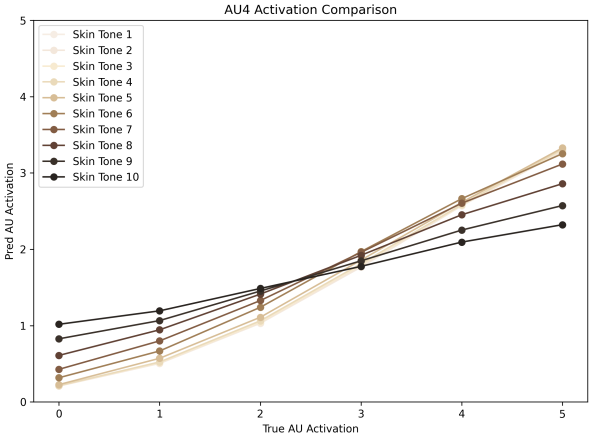

Figure 4: ModelLighter AU Activation Comparison for AU4.

3.2.2 Training Models on Select Skin Tone Ranges: Lighter Faces.

The results from ModelLighter on AU4 can be seen in Figure 4 and Table 5 above. For lighter skin tones (1 to 5), the slope remains relatively stable at around 0.63. This suggests that ModelLighter is adept at tracking the increase in AU4 activation for lighter skin tones. However, as the skin tone progresses toward the darker end of the spectrum, the slope notably decreases. The decreasing trend of the slope, particularly from skin tone 6 onwards, suggests a diminished capability of ModelLighter in effectively tracking AU4 activation for darker skin tones. By skin tone 10, the slope has dropped to approximately 0.27, significantly below the value for lighter skin tones. This could indicate a poorer performance of ModelLighter on darker skin tones, suggesting that the models trained

Table 5: Slope and MAD of ModelLighter AU4 Activation.

Skin Tone	Slope	MAD from True AU Activation
1	0.637202	1.009849
2	0.637675	1.004843
3	0.639741	0.997269
4	0.640499	0.987021
5	0.639594	0.957669
6	0.611360	0.920826
7	0.556935	0.936673
8	0.465135	1.003463
9	0.362362	1.127907
10	0.271869	1.255451

on exclusively lighter faces might lead to suboptimal performance on potentially darker skin tones.

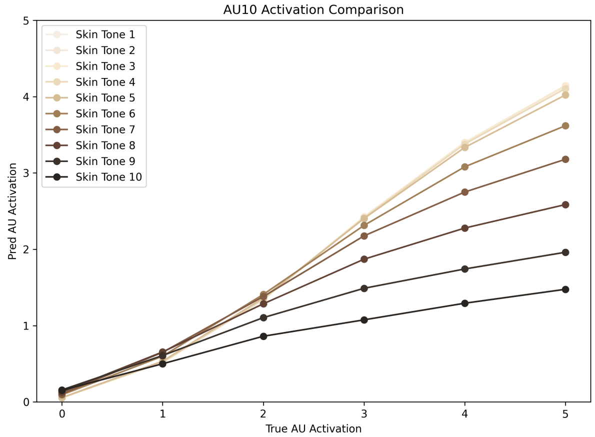

Figure 5: ModelLighter AU Activation Comparison for AU10.

Table 6: Slope and MAD of ModelLighter AU10 Activation.

Skin Tone	Slope	MAD from True AU Activation
1	0.862417	0.539269
2	0.860513	0.541064
3	0.862264	0.537313
4	0.854459	0.548893
5	0.836208	0.567076
6	0.742145	0.679075
7	0.641837	0.828398
8	0.506661	1.076060
9	0.365806	1.375715
10	0.263096	1.658199

We run the same experiment of ModelLighter on AU10, upper lip raiser, shown in 5 and Table 6.

As with the previous experiment, there is a distinct decreasing trend in the slope of estimated AU10 activation as the skin tone increases, suggesting that ModelLighter's ability to track the activation of AU10 diminishes for darker skin tones. Regarding the MAD, the data reveals that the error in AU10 estimation increases substantially as the skin tone gets darker, further supporting that ModelLighter shows diminishing performance on darker tones.

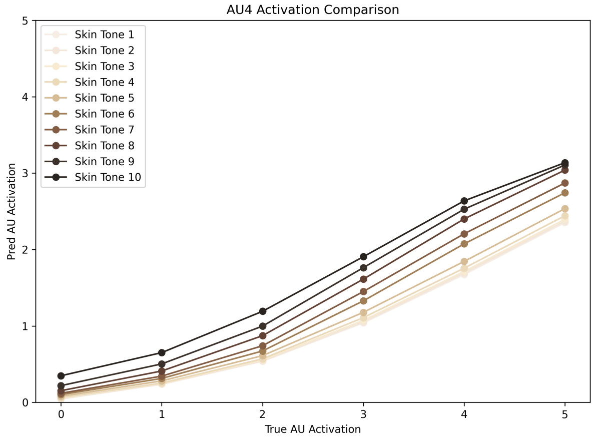

Figure 6: ModelDarker AU Activation Comparison for AU4.

Table 7: Slope and MAD of ModelDarker AU4 Activation.

Skin Tone	Slope	MAD from True AU Activation
1	0.467312	1.533550
2	0.470848	1.524444
3	0.475049	1.512461
4	0.483988	1.488791
5	0.500818	1.438385
6	0.546932	1.328365
7	0.573018	1.251257
8	0.604129	1.136052
9	0.607842	1.053450
10	0.588989	0.970898

3.2.3 Training Models on Select Skin Tone Ranges: Darker Faces.

We then proceeded to assess the performance of ModelDarker, trained exclusively on faces of skin tones 6-10, for its accuracy in tracking AU4. The results are presented in Figure 6 and Table 7.

We can observe an increasing slope trend as the skin tone becomes darker. This suggests an improved ability of ModelDarker to track the activation of AU4 as the skin tone darkens. This contrasts with the behavior of ModelLighter, which showcased a diminishing capacity to track AU activation with darker skin tones.

We see the same trend from ModelAll of decreasing MAD in darker faces, but we notice that the errors in estimation for both lighter and darker faces are higher. These differences in error are larger for lighter faces than darker faces, indicating that ModelDarker demonstrates diminished performance in estimating AU4 activation on lighter skin tones.

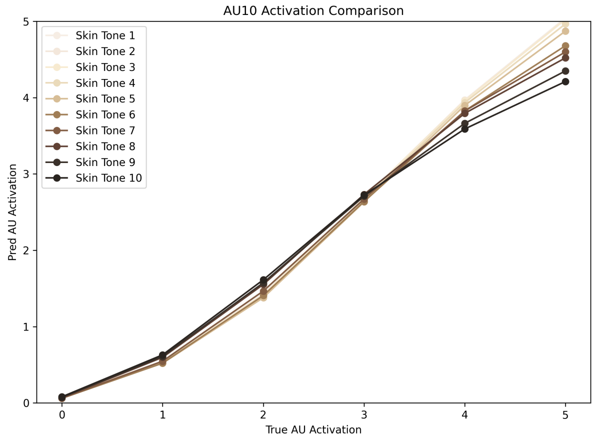

Figure 7: ModelDarker AU Activation Comparison for AU10.

Table 8: Slope and MAD of ModelDarker AU10 Activation.

Skin Tone	Slope	MAD from True AU Activation
1	1.038681	0.249587
2	1.036923	0.249347
3	1.036318	0.249880
4	1.027152	0.263725
5	1.012150	0.293630
6	0.978398	0.331304
7	0.963501	0.325799
8	0.943696	0.312326
9	0.904796	0.361936
10	0.875376	0.385905

Subsequently, we evaluated ModelDarker’s accuracy in tracking AU10 across skin tones, shown in Figure 7 and Table 8. In contrast to AU4, the slope trend for AU10 decreases as skin tone darkens, indicating a reduced capability in tracking AU10 activation for darker skin tones. Additionally, the MAD has an increasing trend for darker faces. We can attribute these trends to the behavior we noticed before, where the model’s estimation of AU10 is higher for lighter faces than darker faces.

While the shape of the AU10 activations for ModelDarker looks similar to the AU10 activations for ModelAll, seen in Table 4, we notice that the MAD is lower across all of the skin tones, suggesting how the distribution of skin tones in the training set may impact the performance and biases of the AU Estimation network.

The contrasting behavior of ModelLighter and ModelDarker highlights the influence of the training data’s skin tone distribution on the models’ performance. While ModelLighter showed diminished performance on darker skin tones, ModelDarker had smaller and more varied performance changes, with overall slightly increased performance on AU10, but overall slightly decreased performance on AU4 relative to ModelAll.

While these outcomes alone do not conclusively prove racial bias, they provide an indication that the model’s performance may vary based on skin color. This study suggests that there may be room for improvement in ensuring equitable performance across different skin tones, contributing to ongoing discussions around bias in facial expression models.

4 DISCUSSION

The aim of this project was to investigate the presence of skin color bias in FER, specifically in relation to the AU recognition. We explored this through a series of experiments using synthetic faces that differ only in skin color and systematically varying the activation of specific AUs. Our results reveal information into how model training, with respect to the distribution of skin tone, can influence performance and potentially contribute to bias.

Our first experiment, using the AU Estimator model trained on synthetic faces, exhibited indications of a skin color bias. The paired t-tests analysis showed statistically significant differences in AU and PSPI scores between light and dark-skinned faces that were identical in all aspects except skin color. This finding suggests that the model may include biases related to skin tone, affecting its accuracy and fairness in processing faces with varying skin tones. Subsequent experiments on AU4 and AU10 activation further supported the existence of a skin color bias in our AU Estimator, demonstrated contrasting color biases across AUs, and showed that these biases may not be uniform or predictable across skin tones and AUs. ModelLighter, trained on lighter tones, demonstrated reduced effectiveness in tracking activation as skin tones got darker, while ModelDarker demonstrated smaller and more varied changes in performance. The interaction between model and dataset biases reflected in the ModelDarker and ModelLighter results indicates the complexity of the skin-tone bias issue.

There are several limitations to acknowledge. Firstly, the size of the dataset was relatively small (940 faces). Secondly, the dataset lacked diversity in terms of ethnicity, morphology, pose, lighting, etc. We only included European faces of the same morphology, so our results may not extend to faces of other ethnicities or characteristics. Thirdly, we only evaluate a single model and a limited selection of AUs. The biases of other FER models may differ, which suggests the need for testing a broader range of models. Additionally, our use of synthetic faces may not fully capture the complexity, nuances, and variability of real human faces. Consequently, the biases observed may not fully reflect those in real-world scenarios.

Our findings highlight the complexity of skin color model biases and the impact of training distribution for FER models. Our research also demonstrates the utility of synthetic faces as a means for systematic evaluation of FER models, facilitating controlled, targeted investigations of bias, and collecting data, especially in the context of facial data, where data collection may be slow, expensive, and/or sensitive. In future work, we will leverage synthetic images to help identify and mitigate the root of racial biases in FER models.

ACKNOWLEDGMENTS

This work was supported by UC San Diego Social Sciences (Advancing Racial Justice Award) and the Sanford Institute for Empathy and Compassion (Empathy and Technology Award). We thank Xiaojing Xu for creating the model and Sarah Fabi for her inspirational work.

REFERENCES

[1] Paul Ekman and Wallace V Friesen. 1976. Measuring facial movement. *Environmental psychology and nonverbal behavior* 1, 1 (1976), 56–75.

[2] Sarah Fabi, Xiaojing Xu, and Virginia R. de Sa. 2022. Exploring the Racial Bias in Pain Detection with a Computer Vision Model. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, J. Culbertson, A. Perfors, H. Rabagliati, and V Ramenzoni (Eds.). 358–365.

[3] M.Rosario González-Rodríguez, M.Carmen Díaz-Fernández, and Carmen Pacheco Gómez. 2020. Facial-expression recognition: An emergent approach to the measurement of tourist satisfaction through emotions. *Telematics and Informatics* 51 (2020), 101404. https://doi.org/10.1016/j.tele.2020.101404

[4] Yifei Guo, Jian Huang, Mingfu Xiong, Zhongyuan Wang, Xinrong Hu, Jihong Wang, and Mohammad Hijji. 2022. Facial expressions recognition with multi-region divided attention networks for smart education cloud applications. *Neurocomputing* 493 (2022), 119–128. https://doi.org/10.1016/j.neucom.2022.04.052

[5] Marco Leo, Pierluigi Carcagnì, Pier Luigi Mazzeo, Paolo Spagnolo, Dario Cazzato, and Cosimo Distante. 2020. Analysis of Facial Information for Healthcare Applications: A Survey on Computer Vision-Based Approaches. *Information* 11, 3 (2020). https://doi.org/10.3390/info11030128

[6] Patrick Lucey, Jeffrey F. Cohn, Kenneth M. Prkachin, Patricia E. Solomon, and Iain Matthews. 2011. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*. 57–64. https://doi.org/10.1109/FG.2011.5771462

[7] Paris Mavromoustakos Blom, Sander Bakkes, Chek Tan, Shimon Whiteson, Diederik Roijers, Roberto Valenti, and Theo Gevers. 2021. Towards Personalised Gaming via Facial Expression Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 10, 1 (Jun. 2021), 30–36. https://doi.org/10.1609/aiide.v10i1.12707

[8] Ellis Monk. 2019. Monk Skin Tone Scale. https://skintone.google

[9] O. M. Parkhi, A. Vedaldi, and A. Zisserman. 2015. Deep Face Recognition. In *British Machine Vision Conference*.

[10] Kenneth M Prkachin and Patricia E Solomon. 2008. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain* 139, 2 (2008), 267–274.

[11] Ritik Raina, Miguel Monares, Mingze Xu, Sarah Fabi, Xiaojing Xu, Lehan Li, Will Sumerfield, Jin Gan, and Virginia R. de Sa. 2022. Exploring Biases in Facial Expression Analysis. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*. https://openreview.net/forum?id=jrddoq9NDP6

[12] Xiaojing Xu, Jeannie S Huang, and Virginia R de Sa. 2020. Pain Evaluation in Video using Extended Multitask Learning from Multidimensional Measurements. In *Proceedings of the Machine Learning for Health ML4H NeurIPS Workshop (Proceedings of Machine Learning Research, Vol. 116)*, Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones (Eds.). PMLR, 141–154. https://proceedings.mlr.press/v116/xu20a.html

Received 28 May 2023; Accepted 20 June 2023