Analysis of Extended Multitask Learning Model on Facial Pain Evaluation

Lehan Li
University of California, San Diego
Cognitive Science Honors Program 2021 - 2022
Committee: Virginia R. de Sa, Vijay Veerabadran

Introduction

Pain perception based on facial expression is highly individualized and heavily influenced by a person's prior knowledge and experiences, and it's vulnerable to individual biases related to race and ethnicity. A substantial number of medical practitioners hold false beliefs about biological differences between blacks and whites, and these false beliefs predict racial bias in pain perception and treatment recommendation accuracy (Hoffman, Trawalter, Axt & Oliver, 2016). Persistent racial and ethnicity biases exist in the healthcare system, with African Americans tend to receive lesser quality pain care compared to White Americans (Mathur, Richeson, Paice, Muzyka & Chiao, 2014; Anderson, Payne, 2009).

Various metrics are invented in order to properly measure the perceived pain level of individuals. Machine Learning model in facial expression recognition and prediction is developed in order to be used in examination of the perceived pain intensity of individual people. Algorithmic bias exists in the machine learning model. The machine learning model is able to imbibe bias in the dataset and produce unexplainable discriminatory outcomes and influence an individual's articulateness of system outcome due to the presence of racial bias features in datasets. (Sengupta & Srivastava, 2022). Machine Learning performances in face recognition and facial expression emotion recognition are significantly worse among African people. Such inequity in model performance could be caused by unbalanced dataset, inappropriate sampling techniques and the intrinsic nature of machine learning algorithms.

The aim of the project is to analyze racial bias in facial pain perception using Machine Learning Model. The computer vision model Extended Multitask Learning Model developed by Xu and de Sa examines a facial image and predicts relevant statistics and metrics in facial expression and pain evaluation. Three central questions are investigated:

1. If the Machine Learning model exhibit algorithmic racial biased in pain prediction in the same way as human racial bias
2. Which facial components / facial muscles yield greater discrepancies between race
3. Algorithmic racial bias is due to a) skin color b) morphology of the face related to race (such as the shape of the eyes, mouth, lip etc)

Moreover, the answer to the above three questions regarding the machine learning model might be able to cast some insights in human racial bias in pain perception as well, and to better understand the reason and rationale behind human racial bias in pain perception on facial expressions.

Related Work

The most commonly used method for pain measurement is Visual Analog Scale. Visual Analog Scale (VAS) is the subjective measure for pain based on self reported scores. It's widely used in healthcare professionals and physicians to evaluate perceived pain intensity of a patient and track the development of various symptoms in order to provide adequate treatment. A patient is asked to indicate his/her perceived pain intensity (most commonly) along a 100 mm horizontal line, and this rating is then measured from the left edge (=VAS score) (Myles, Troedel, Boquest & Reeves, 1999). It contains a continuous spectrum of intensity ranging from none to extreme, and is further divided into subcategories: none, mild, moderate and severe. A patient would rate their perceived pain intensity from 0 to 10. However, VAS scores are highly subjective and suffer from individual bias. It's useful when tracking the development of the pain for a patient, but less effective for pain intensity comparison between individuals.

Instead of solely based on self reported scores, more objective pain measurement metrics utilize human reflex response for pain prediction, especially for facial expressions. Two significant measurements in facial expression recognition for pain are the Facial Action Unit and Prkachin and Solomon Pain Intensity. Facial Action Unit (AU) is defined by the Facial Action Coding System (FACS) which offers a taxonomic decomposition of the facial expression. (Hjortsjo, 1970; Cohen, Ambadar, & Ekman, 2007). The Facial Action Unit divides and labels facial expressions into individual facial muscle movements with an activation level from 0% (no activation) to 100% (maximum activation). There are a total of 46 Main Action Units (coding from the top to the bottom of the face), 8 Head Movement Action Unit and 4 Eye Movement Action Unit. Prkachin and Solomon Pain Intensity (PSPI) is a facial expression measurement for pain, which is based on the Facial Action Unit. (Prkachin & Solomon, 2008) PSPI = AU4+max(AU6,AU7) +max(AU9,AU10) +AU43. The higher the PSPI value is, the more pain a person is experiencing.

AU detection through manual coding could be labor intensive and requires extensive training and professional experiences. Thus, machine learning and deep learning are used under these circumstances to provide fast and accurate predictions. In order to correctly estimate the pain intensity, a computer vision model (ExtendedMTL4Pain) is presented to conduct pain intensity detection for individuals from different racial and demographic backgrounds with distinct visual attributes. The Extended Multi-Task Learning (ExtendedMTL4Pain) model contains three stages. The first stage takes input images to predict AUs and PSPI. The second stage takes PSPI scores and accumulate over multiple images to obtain relevant statistics (min, max, mean, sd, 25th, 50th, 75th, 85th, 95th percentile) as input to predict VAS, OPR (Observers Pain Rating 0-5), AFF (Affective-motivational scale 0 - 15) and SEN (Sensory Scale 0 - 15).

And then apply OLC for each value of VAS, OPR, AFF and SEN across multiple models to produce the final VAS value (Xu et al., 2020).

Method

Data Generation

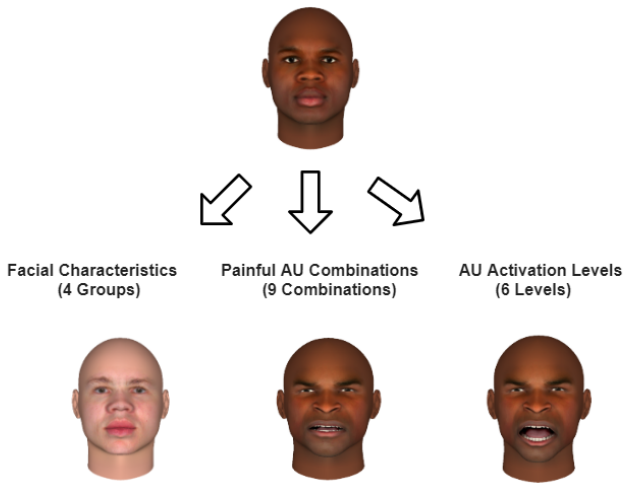
For the original Extended Multitask Learning Model, it's trained in real world dataset from UNBC-McMaster Shoulder Pain Dataset (Lucey et al., 2011) It's a publicly available dataset obtained by McMaster University and University of Northern British Columbia. Participants with shoulder pain health issues were recruited. The facial dataset is recorded when participants were conducting a series of actions with their affected and unaffected limbs. The dataset includes 200 video sequences containing spontaneous facial expressions from 25 participants.

For this project, artificial data is used for model performance analysis. Artificial data allows more accurate generation of Action Unit Activation Level, better facial characteristics manipulation, and better elimination of confounding variables. The Action UnitS which are tightly related to pain facial expressions are AU4, AU6, AU7, AU10, AU12, AU20, AU25, AU26 AND AU43 [Table 1]. Within the 9 Action Unit, 4 of them relate to the muscle movement around the eyes, and 5 of them relate to the muscle movement around the mouth. The artificial stimuli are generated from a software called FaceGen Modeller. The FaceGen Modeller software allows manipulation of faces regarding race, skin color, face texture, action unit and its activation level.

AUs	AU4	AU6	AU7	AU10	AU12	AU20	AU25	AU26	AU43
Definition	Brow Lower-er	Cheek Raiser	Lid Tighte-ner	Upper Lip Raiser	Lip Corner Puller	Lip Stretc-her	Lips part	Jaw Drop	Eyes Closed

[Table 1: Definition of Action Unit]

In general, the facial image dataset consists of three major components: Facial characteristics, painful Action Unit Combinations and Action Unit Activation levels also known as AU score or AU value [Figure 1]. And all the dataset is generated with a male face.



[Figure 1: Overview of Data Generation Process]

The first component is the basic facial characteristics [Figure 1]. It's the basic image of the face without any facial expression and muscle activation. It's determined by racial differences and individual variability, and leads to the difference in skin color and the facial features such as the relative position and shape of facial features. For the first components, a total

of 4 groups are generated: African baseline face, European baseline face, African baseline face with light skin, and European baseline face with dark skin. Firstly, a face image is generated with African facial features while other facial features are selected at random. Then only change the color of the African face into light skin while all the other characteristics stay the same. Then, a face image with European facial features is generated, and the skin color of the European face is manipulated into dark skin. Thus, regarding skin color, African faces and European faces with dark skin should have the same dark color skin, while European faces and African faces with light skin should have the same light color skin. Regarding facial features, African faces and African faces with light skin should look the same except for skin color, and European faces and European faces with dark skin should also look the same except for skin color. Since the first and third aim of the project are to analyze if the computer vision model exhibit racial bias in facial pain perception, and if such racial bias in computer algorithms is due to the skin color or facial morphology related to race and ethnicity, in the data generation process, these two independent features should be generated separately while other confounding variables stay constant. In the interest of exploring model performance in different racial groups, results for African and European groups could be compared to draw the inference. In the interest of exploring the impact of skin color on model performance, results could be compared in African vs African with light skin and European vs European with dark skin. In order to explore the impact of facial morphology on the CV model, results could be compared in African vs European with dark skin and European vs African with light skin.

African

African with Light Skin

European

European with Dark Skin

[Figure 2: Four Conditions of Random Face]

The second component is the painful Action Unit combination. Since the project is interested in pain perception in facial images, 9 different facial expressions are generated using the defined 9 Action Units with various Action Unit Activation Level. (AU4, AU6, AU7, AU10, AU12, AU20, AU25, AU26 AND AU43). The Activation Level of the 9 Action Units for 9 painful facial expressions are stored as baseline painful facial expressions, and they serve as the foundation for the third step [Table 2]. In this process, only the activations of facial muscles are stored, which are not related to facial features. And then, the baseline Action Unit Activation Level is mapped onto the 4 racial groups such that for each racial group, there are 9 painful Action Unit combinations, which gives us a total of 36 painful facial expressions as baseline images.

The third component is the Activation Level which could also be referred as Action Unit value or Action Unit score. It measures the degrees of activation of the facial muscles. In the FaceGen Modeller software, the range for Activation Level is from 0 (no activation at all) to 10 (full activation). Total of 6 different Activation Levels are selected, which are 0, 2, 4, 6, 8 and 10. In this process, 9 painful Action Unit combinations are used as baseline, and modify the Action Unit Activation Level for only one AU, each corresponding to one Action Unit at a time. For example, for the first set of AU combinations, Activation Level of AU 4 is manipulated, and the second set of AU combinations, Activation Level of AU6 is manipulated and so on. Thus, a total of 216 artificial faces are generated. [Fabi, Xu, de Sa, 2022]

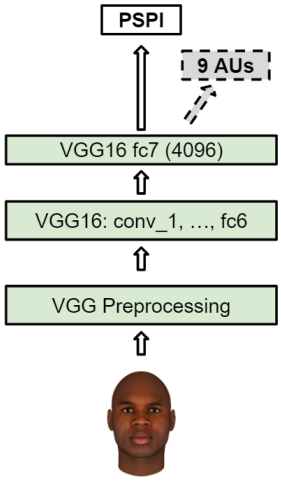
AU Combinations for 9 Painful Facial Expression

AU Score	AU4	AU6	AU7	AU10	AU12	AU20	AU25	AU26	AU43
Comb 1	8	8	6	9	5	8	5	3	0
Comb 2	6	10	8	7	1	6	2	3	0
Comb 3	7	8	6	8	3	2	3	3	0
Comb 4	8	8	7	8	3	4	2	6	0
Comb 5	5	6	9	5	2	6	5	1	0
Comb 6	6	10	7	10	1	2	1	1	0
Comb 7	5	9	4	8	3	5	4	2	0
Comb 8	5	9	9	7	1	4	8	2	0
Comb 9	4	10	7	7	1	5	6	2	0

[Table 2: AU Combinations for 9 Painful Facial Expression]

Computer Vision Model Deployment

The project is performed using the first stage of the computer vision model Extended Multitask Learning Model (Xu et al., 2020). The first stage of the model takes a dataset of facial images and generates predictions of PSPI values and 9 AU values. The preprocessing step implements a pretrained model VGG Matconvnet for facial detection with a bounding box factor 0f 0.1, then crop the white margins around the face and only retain the central facial image (Parkhi, Vedaldi & Zisserman). The cropped images are sent into the CV model which contains 6 convolutional layers from the original VGG16 model and a transfer learning regression layer trained over the UNBC-McMaster Shoulder Pain Dataset [Figure 3]. After obtaining the result, a general overview of the data is conducted first. A paired t test is conducted to analyze if there's significant difference in model performance between racial groups related to PSPI value and AU values. The project then analyzes the individual AU performances across different racial groups. Finally, trends of individual AU prediction results are discussed.



Modified from [Xu, X , Huang, J.S., & de Sa, V.R. (2019)]

[Figure 3: CV Model]

Result

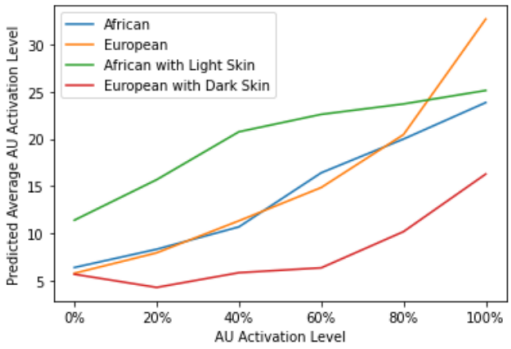
Five paired t-tests are conducted over different racial groups to analyze 1) if the model has significant performance difference across race and 2) if the skin color and/or facial morphology effects on CV models. For racial comparison, paired t-test is implemented in African vs European face. For skin color bias analysis, two paired t-tests are implemented in African vs African with light skin and European vs European with dark skin. For facial morphology bias, two paired t-tests are implemented in African vs European with dark skin and European vs African with light skin. Two resulting values are tested: 1) absolute difference in real and predicted PSPI score, and 2) absolute difference in real and predicted AU score.

Paired t-test Between Racial Groups	Absolute Difference in PSPI Score (P Value)	Absolute Difference in AU Score (P Value)
African vs European	0.23	0.24
African vs African with light skin	0.0062	0.92
European vs European with dark skin	0.54	0.026
African vs European with dark skin	0.025	0.36
European vs African with light skin	0.32	0.069

[Table 3: Paired t-test Statistics for Racial Groups]

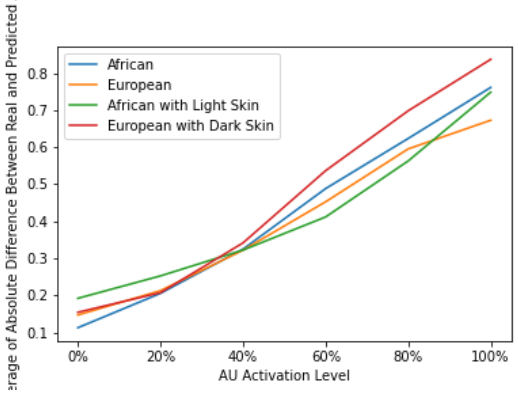
Racial bias in absolute difference in PSPI score and AU score is not detected in the model, however, there exists some statistically significant value regarding skin color bias and facial morphology bias.

Average AU Activation Prediction Over Race



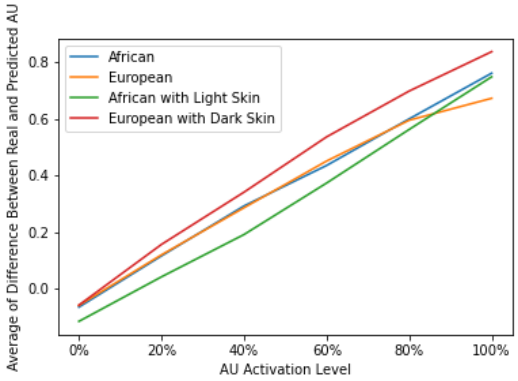
[Figure 4: Average AU Score Prediction]

Average of Absolute Difference of AU



[Figure 5: Avg Abs Diff AU Score Prediction]

Average of Difference of AU



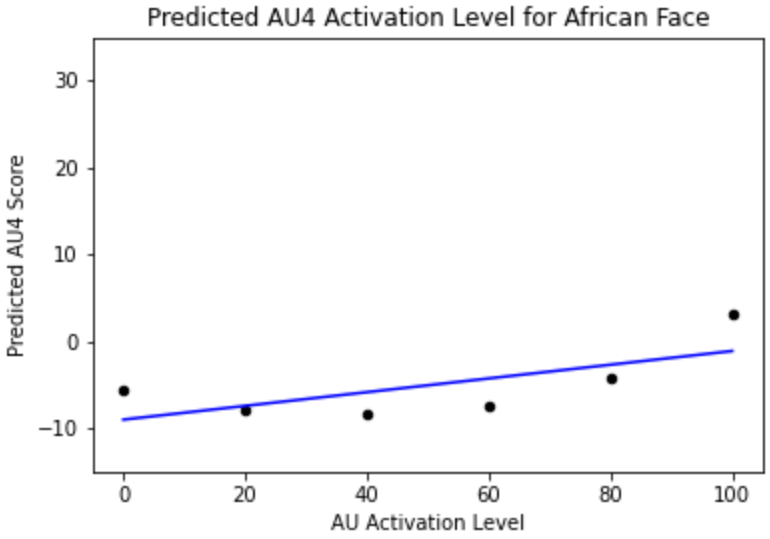
[Figure 6: Avg Diff AU Score Prediction]

AU prediction scores are in general positively correlated with the target AU score. Different racial groups seem to have different AU prediction performance. Africans with Light skin have a greater AU predicted score in general. The CV model performance on European dataset and African with light skin dataset is similar to each other, while the predicted AU for European with dark skin dataset is the lowest. [Figure 4]. In general, the average of absolute difference between real and predicted AU score is positively correlated with the target AU score. And there's a crosswalk between lines around 35% activation level. With greater target AU score, the error gets larger. And there's no significant difference on model performance between racial groups [Figure 5]. Similar pattern is observed in the average difference between real and predicted AU scores, but without a significant crosswalk in lines for racial groups. However, for all racial groups, the model has a negative AU activation level error when the real AU activation level is 0%.

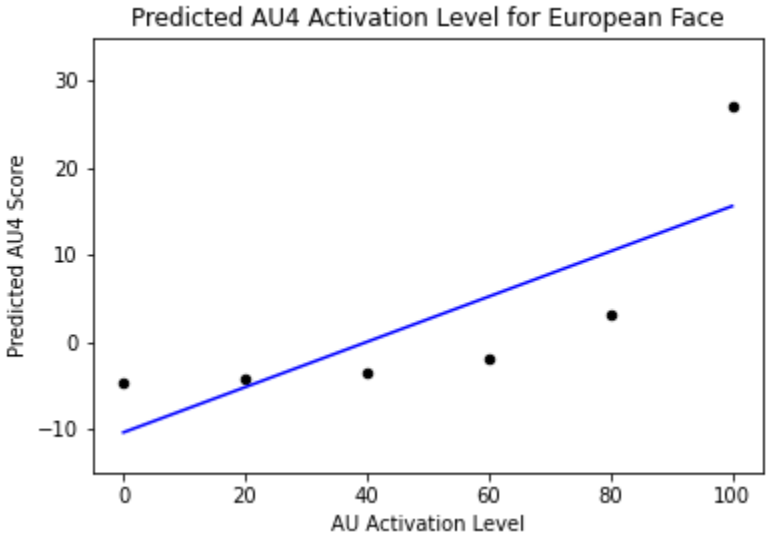
Individual AU performance has greater variability across race. AU4 and AU7 shows statistical significant discrepancy between African and European faces with both p values smaller than 0.001, which indicates potential racial bias in AU detection.

AU4: Brow Lowerer

The project examines AU4 (Brow Lowerer) and the corresponding model performances across different racial groups. The predicted AU scores are significantly higher for European faces compared to African faces. The model also shows significantly higher prediction for facial morphology difference in European faces with dark skin compared to African faces. However, the model fails to show a statistically significant difference in skin color difference.



[Figure 8: Predicted AU score for African Face]

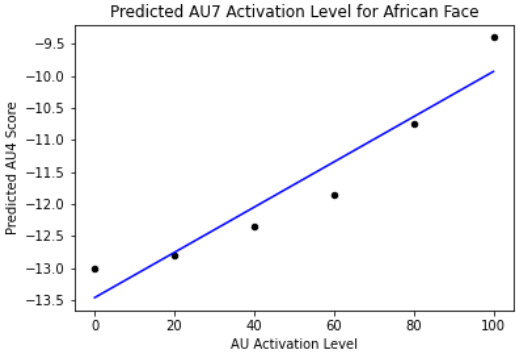


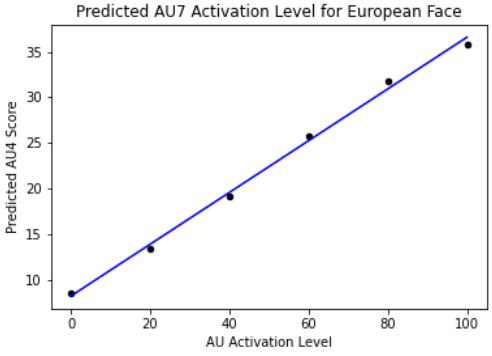
[Figure 9: Predicted AU score for African Face]

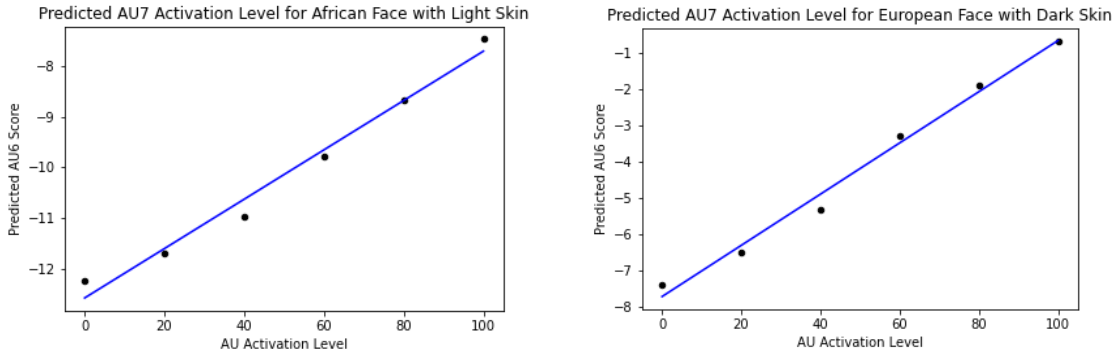
For AU4, there exists positive correlation between predicted and target AU values. The range for predicted AU4 value for African faces is from -8 to 3 [Figure 6], while predicted AU4 value for European faces is from -4.58 to -27.02 [Figure 7]. The predicted AU values for both African faces and European faces are relatively similar when the real AU4 activation level is at 0%. As the real AU4 activation level gets larger, the predicted AU4 value gets larger for European faces compared to African faces. The slope of the linear regression line for African faces is 0.078, which is significantly smaller compared to the slope of the linear regression line for European faces 0.26.

Next, the project tests the effect of skin color and facial morphology on model performance. There is a significant increase in model prediction accuracy in the African faces with light skin. The range of predicted AU4 value is from -0.33 to 57.58, with a linear regression line slope of 0.55. Compared to model performance on African faces, the overall prediction for Africans with light skin increases, with a greater unit increment as the target AU4 activation level increases. While for facial morphology conditions, the result is not significantly different from previous data. Thus, algorithmic bias for AU4 is likely to be a result of skin color difference instead of facial morphology difference.

AU7: Lid Tightener







[Figure 10: AU 7 Prediction on Racial Condition]

Next step is to analyze model performance on AU7. AU7 also yields statistically significant higher predictions for European faces compared to African faces. Similar positive correlation pattern is observed across all racial groups. AU7 prediction for African face is from -13.01 to -9.39 with linear regression line slope of 0.035, while AU7 prediction for European face is from 8.53 to 35.83 with linear regression line slope of 0.28. The AU7 prediction is at 0% activation level is different across racial groups, and the unit increase is also greater for European faces. For skin color impact, African facial groups seem not to be heavily influenced by the change in skin color. The range of AU7 prediction value is from -12.25 to 7.45 with a slope of 0.049. The statistics don't vary greatly with the change in skin color. However, European facial groups experienced a significant drop in model prediction scores when converting European faces into dark skin color. Prediction for European faces with dark skin ranges from -7.40 to 7.69 with a slope of 0.07. Even though the prediction value dropped significantly compared to original European faces, its value is still above original African values. Thus, for AU7, algorithmic racial bias might be due to both skin color impact and facial morphology impact.

Conclusion

In conclusion, the project seeks to understand algorithmic racial bias in pain perception based on facial images. The project is analyzed using artificial facial image dataset to better control the independent variables as well as confounding variables. The Extended Multitask Learning Model does not exert significant racial bias in overall PSPI score prediction, but exerts model performance difference in individual Action Unit Predictions with often overestimated AU activation level for European faces compared to African faces. There is great variability in model performance on individual AU across different racial groups. The model is not consistently better at detecting AU for one race, and the effect of skin color and facial morphology is not consistently observed in both African and European conditions. AU4 and AU7 have significantly higher prediction scores for European faces compared to African faces. For some AUs, such differences might be due to skin color difference, while others might be due to a combination of effects.

In general, Extended Multitask Learning Model does not show significant racial bias in pain perception using facial images. However, one should be cautious when using the model for other types of emotional detection. Since different emotions require different combinations of Action Units, slight bias in individual AU prediction might be caught in other emotional detection tasks.

Future work should be done on a more thorough analysis of individual AU predictions. And then train the model with additional dataset to improve AU prediction performances. Moreover, we only focus on the stage 1 of the Extended Multitask Learning Model, one possible

work is to analyze if the difference in AU prediction actually influences the final result VAS scores of the facial images.

Acknowledgement

Funding greatly acknowledged from UC San Diego Social Sciences (Advancing Racial Justice award), and the Sanford Institute for Empathy and Compassion (Center for Empathy and Technology award). We also thank the team of FaceGen Modeller for providing us with their software. I want to extend my sincere thanks to professor Virginia de Sa for advising me throughout this project, thanks for Sarah for her guidance on the research plan and method, thanks Mingze for his help in generating artificial faces, thanks William for his insights in artificial face generation procedure and his contribution on sona experiment, thanks for Dr. Xu and Vijay for their help in the model deployment, and thanks Ritik for his further facial expression analysis as well as the help I received from the entire de Sa lab. Moreover, I also want to thank Dr. Kutas and the entire Cognitive Science Honor Project Committee for their help and advice from formulating research plans to presentation preparation.

Reference

Hoffman, Kelly M., et al. “Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs about Biological Differences between Blacks and Whites.” Proceedings of the National Academy of Sciences, vol. 113, no. 16, 2016, pp. 4296–4301., https://doi.org/10.1073/pnas.1516047113.

Mathur, Vani A., et al. “Racial Bias in Pain Perception and Response: Experimental Examination of Automatic and Deliberate Processes.” The Journal of Pain, vol. 15, no. 5, 2014, pp. 476–484., https://doi.org/10.1016/j.jpain.2014.01.488.

Anderson, Karen O., et al. “Racial and ethnic disparities in pain: causes and consequences of unequal care.” The Journal of Pain, vol. 10, no. 12, 2009, pp. 1187–1204., https://doi.org/10.1016/j.jpain.2009.10.002.

Myles, Paul, Troedel, Boquest & Reeves. “The Pain Visual Analog Scale: Linear or Nonlinear?” Anesthesiology, vol. 100, no. 3, 2004, pp. 744–744., https://doi.org/10.1097/00000542-200403000-00042.

Hjortsjo, C.-H. (1970). ¨ Man’s face and mimic language. Lund: Studentlitteratur.

Cohen, J. F., Ambadar, Z., & Ekman, P. (2007). Observerbased measurement of facial expression with the facial action coding system. In J. A. Coan & J. J. B. Allen (Eds.), Handbook of Emotion Elicitation and Assessment. Oxford:Oxford University Press.

Prkachin, K. M., & Solomon, P. E. (2008). The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. Pain, 139(2), 267–274.

Sengupta & Srivastava. (2022) Causal effect of racial bias in data and machine learning algorithms on user persuasiveness & discriminatory decision making: An Empirical Study. https://doi.org/10.48550/arXiv.2202.00471

Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., & Matthews, I. (2011). Painful data: The UNBC-McMaster shoulder pain expression archive database. In IEEE International Conference on Automatic Face & Gesture Recognition (pp. 57–64).

Parkhi, Vedaldi & Zisserman. https://www.robots.ox.ac.uk/~vgg/software/vgg_face/

Fabi, Sarah., Xu, Xiaoxing., de Sa, Virginia (2022). Exploring the Racial Bias in Pain Detection with a Computer Vision Model. Oral presentation at CogSci 2022.