An Analysis of Gender Biases in Emotion Recognition Software Performance

Hannah Ashton
School of Engineering and Computer Science
Syracuse University
Syracuse, New York
hcashton@syr.edu

***Abstract*— Emotion recognition technology plays a critical role in various applications, from security systems to human-computer interactions. However, the presence of bias in these systems can have significant implications. While bias in other biometric technologies, such as facial recognition and voice recognition, has been well-documented, it is also present in emotion recognition systems. Previous research has highlighted accuracy issues across race and ages, though a comprehensive analysis in the fairness of emotion recognition remains limited.**

This study explores the performance bias of emotion recognition software between men and women. Using DeepFace, a widely-used deep learning-based emotion recognition tool, images from the Facial Expression Recognition 2013 dataset are labeled according to the emotion on the face in the image. The data is then assesses to determine the software's accuracy across different genders. The experiment found differences in emotion recognition performance between men and women, though the overall performance was similar. These findings contribute to the understanding of gender bias in emotion recognition and inform efforts to mitigate such biases.

Keywords*— emotion recognition, biometrics, detection, image processing, action units

I. Introduction

Humans rely heavily on emotion recognition through non-verbal cues in conversations and interactions. As artificial intelligence and technology increasingly emulate human characteristics, the application of computer emotion recognition is expanding across various domains, including commercial and sales environments, ensuring safer driving, filtering candidates in the hiring process, and within the criminal justice system. While emotion recognition has been studied through voice, gait, and physiological signals, this paper will specifically focus on emotion recognition through facial expressions.

With the growing use of biometric systems by the general public, attention has been drawn to the biases that can arise in their performance across different demographics, such as gender, skin color, race, and age. These biases can have significant consequences depending on the application. Although there has been substantial research into biases in facial detection and recognition, less is known about biases in facial emotion detection. This paper aims to address this gap by examining performance differences for men and women using the commonly-used facial emotion detection algorithm, DeepFace.

This paper acknowledges the diversity of gender identities and does not attempt to define or categorize these terms comprehensively. The terms "man" and "woman" are used based on the prediction provided by the DeepFace algorithm, and "gender" is used to refer to these categories, understanding that these labels may not truly match the individual's identity. It is also acknowledged that a binary gender classification does not encompass all human experiences, and is used for the purposes of this experiment to begin to measure the biases in the technology.

Following this introduction, Section 2 will review previous research on biases in biometrics and delve into the methodology and accuracy of emotion detection. Section 3 will describe the system tested and performance measurements. Section 4 presents and analyses the results of the emotion detection across the two genders. Section 5 concludes the paper with the key points and suggestions for continued work.

II. Previous Works

In recent years, the field of emotion detection has garnered significant attention due to its increasing integration into society, from customer service to security systems. To understand the implications of this technology, it is essential to review previous works that address the underlying biases and accuracy issues inherent in emotion recognition systems. This section will begin by exploring the broader context of biases in biometric technologies, providing a foundational understanding of how these issues may translate into emotion recognition. Emotion recognition is then defined and examined for its current accuracy metrics, highlighting both advancements and limitations. Finally, the focus will shift to works that have previously measured the disparities in accuracy of emotion recognition systems across different demographics. This comprehensive review aims to contextualize the current research landscape and identify key areas for further investigation and improvement.

A. Biases in Biometrics

Drozdowski et al highlights that biases in biometric systems like facial recognition, fingerprinting, iris, and even palm printing can arise from human design choices, the algorithms themselves, and the training datasets used [1]. The implications of these errors vary greatly, depending on the context: repeated false negatives when unlocking one's iPhone may lead to increased frustrations, but false positives in surveillance and policing could result in a wrongful conviction altering the path of one's life. The article discusses proposed solutions to mitigate these issues, including improving diversity in the training set and utilizing a "bias-aware approach" for recognition [1].

Other studies also explore ways to mitigate biases in facial recognition. In one such example, multiple facial recognition techniques were tested for accuracy across different races and found that Principal Component Analysis and Fourier Transforms were the most effective methods [2]. Another article found that the use of synthetically created facial images, specifically designed to increase the diversity of the dataset, ultimately reduced the bias of the facial recognition [3].

B. What is Emotion Recognition

Emotion recognition refers to the technology and methods used to identify and interpret human emotions based on various inputs. This can be achieved through analyzing vocal tone, gait and posture, or physiological signs such as heart rate and skin conductance [4]. Emotion detection systems can operate publicly and passively, allowing them to assess emotional states without requiring direct interaction from individuals. These systems typically focus on anger, fear, sadness, disgust, surprise, and happiness as the emotions identified, though more recently “neutral” has been added as the absence of an emotion [4].

The process of emotion detection generally involves three key phases: facial detection, feature extraction, and expression classification. Facial detection locates facial features, feature extraction analyzes these features to capture emotion-specific patterns, and expression classification interprets these patterns to determine the emotional state. The feature extraction and expression classification methods vary among models, but two common ones include the “k-nearest neighborhood” and the “eigenface algorithm” [4]. The k-nearest neighborhood uses some pre-selected “k” value to identify a close-enough image. Eigenfaces, originally used in facial recognition, rely on Principal Component Analysis to create a set of base images that all other images can be composed of. A similar method was used in Cai et al when measuring facial expressions across age, race, and gender [5].

Another method used in emotion detection involves the use of “action units”. Action units originated from the Facial Action Coding System, which was developed by psychologists in the 1970s in an effort to categorize facial images by emotion [6]. The system describes each miniscule facial movement (such as wrinkling the nose or raising the cheeks) as an individual action units; each emotion can then be described as a specific set of these action units [6]. In computer emotion recognition, these action units can be extracted through Gabor filters [7]. The action units identified in a face are then matched to the emotion from the Facial Action Coding System that is represented with the most similar action units.

Applications of emotion detection span several fields, including mental health and medicine for diagnosing and monitoring emotional well-being, product development for enhancing user experience, and security for improving surveillance and threat assessment [8]. Recent advancements have shown promising results, with some models particularly excelling in accurately predicting emotions like happiness [8].

C. Accuracy of Emotion Recognition

Overall, the accuracy of emotion recognition software is estimated to be between 64% and 88% or 85% and 97%, depending on the source [9, 4]. Various solutions have been suggested to increase the accuracy of software. One such solution is suggested by Huzaifa Shahbaz et al and involves more data points for each person or instance being studied. By using voice, text, and visual samples, an accuracy of 87.8% is achieved, higher than the other accuracy calculations reported in this paper [9].

Another suggested solution targets the difficult distinction between surprise and fear. Because these two emotions present similarly in the face, Cadayona et al suggest first classifying the image as a positive or negative emotion; then, the distinction between surprise (positive) and fear (negative) becomes simpler [10].

D. Biases in Emotion Recognition

Emotion detection software, not dissimilar to other biometrics, has come under scrutiny for its inherent biases, particularly with protected classifications such as age, skin color, and ability level. A study highlighted that these systems are less accurate in detecting faces and subsequently recognizing emotions in individuals with certain physical characteristics, such as older age or darker skin tones [4]. Further, distinct facial features like pale or thin lips and small eyes are associated with poorer emotion recognition performance, which may indicate further biases for certain ages or races. These variations can significantly impact the effectiveness of emotion detection, leading to potential misinterpretations or missed detections.

Another study that utilized professional basketball players’ photos as a dataset found that black players were more likely to be categorized as angry by emotion detection software than their white counterparts [11]. This discrepancy was found across both the softwares tested (Face++ and Microsoft’s AI), and existed even after controlling for the degree of the player’s smile [11]. This bias reflects a broader issue within the technology, where racial stereotypes are inadvertently perpetuated, causing unfair and inaccurate assessments of emotions based on skin color and facial structure. Such biases not only undermine the reliability of these systems but also raise concerns about their deployment in real-world settings where fairness and accuracy are critical.

Finally, research has shown that the facial action units, which play an important role in the emotion recognition field as described above, can vary significantly across different demographics, including gender, race, and age [12]. For example, women may smile more broadly than men, leading to variations in how happiness is detected. These differences mean that the same facial expression might be interpreted differently depending on the individual's demographic background, further complicating the accuracy of emotion recognition technologies. As these biases become more evident, it is increasingly clear that current emotion detection systems require significant refinement to ensure they are equitable and effective across diverse populations.

As these issues become increasingly public and criticized, various techniques have been explored to mitigate the impacts. One study found that specific softwares are less biased; YOLOv3, for instance, was found to “excel in detecting faces even when they were not directly facing the camera, irrespective of factors such as race, gender, color, and angle” [8]. Many also suggest that a more diverse training dataset will minimize the bias in the algorithm [8], since it will be taught to recognize emotions more broadly.

Suresh and Ong suggested that rather than teaching the system to avoid specific characteristics that may result in biased results, the algorithm should be taught which characteristics to specifically look for. The system is directed to the facial elements directly associated with the emotion rather than left to find an unintended correlation with facial elements that may align more closelybwith a specific race, age, or gender [13]. A concern with this method is the possibility that the system user (who directs the system what to specifically pay attention to) may have their own subconscious biases that would ultimately lead to biases in the system.

III. Experiment Design

The experiment aims to compare the emotion recognition accuracy for photos of men and women using DeepFace, a widely-used emotion detection software. The target emotions to be recognized include Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

A. Test Data Set

The dataset used in this paper will be the 2013 Facial Expression Recognition dataset provided by Kaggle (DS). According to Kaggle:

"The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image.

The task is to categorize each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set consists of 28,709 examples and the public test set consists of 3,589 examples."

This experiment will utilize the public test set, which as stated above, contains 3,589 photos. These photos are downloadable from Kaggle in sorted folders by emotion. To prepare the dataset for the experiment, each photo will be renamed to include the emotion it portrays, as determined by the folder name.

The gender of the individual photographed in each image will be predicted using DeepFace's "dominant gender". Any images from which DeepFace cannot detect a face will be ignored for the purposes of the experiment, since both the gender and the emotion will not be predictable.

B. Emotion Recognition Software

The emotion recognition for this experiment will be conducted utilizing DeepFace. DeepFace is a deep learning-based facial emotion recognition system that is widely used and open-source. It leverages several pre-trained models, including VGG-Face, Google FaceNet, OpenFace, Facebook DeepFace, DeepID, ArcFace, and Dlib, to achieve high accuracy in emotion recognition.

The stages of the emotion recognition will include detection of a face in the image, alignment of the face in the image, representation analysis of the expression, and verification to match the expression with the closest emotion. These stages are all conducted by DeepFace in the background.

The software calculates the percentages of each of the seven emotions (Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral) present in the image. The dominant emotion can be identified from these percentages, and will be used for the purposes of this experiment.

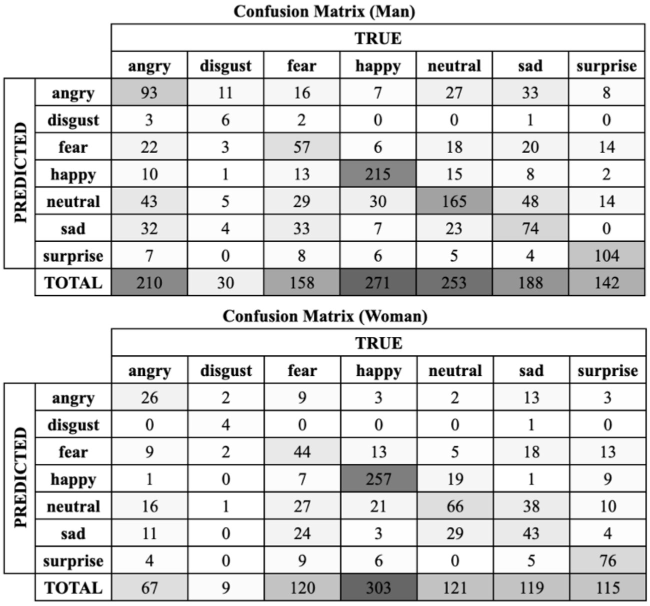
C. Analysis of Bias

After the emotion recognition software has predicted the dominant emotion for each photo, the prediction will be compared to the "real" emotion, as identified from the dataset. The image name, predicted gender, predicted emotion, real emotion, and whether or not the predicted emotion matches the real emotion will be printed out as a comma separated file. The results file will be used to construct a standard and a normalized confusion matrix for each of the genders. This confusion matrix will then be used to compare and contrast the accuracy for each emotion across genders.

IV. Results

A. Face Detection and Gender Distribution

Out of the 3,589 images analyzed using DeepFace, the software was able to detect a face in 2,106 of them, accounting for approximately 59% of the total images. Within this subset, 1,252 images were identified as containing men, while 854 images were identified as containing women. This indicates that just over 40% of the detected faces were women, highlighting a gender imbalance. This imbalance is prevalent across emotions as well; on average by emotion, 37% of the images contained women.



Confusion Matrix (Man)

		TRUE						
		angry	disgust	fear	happy	neutral	sad	surprise
PREDICTED	angry	93	11	16	7	27	33	8
	disgust	3	6	2	0	0	1	0
	fear	22	3	57	6	18	20	14
	happy	10	1	13	215	15	8	2
	neutral	43	5	29	30	165	48	14
	sad	32	4	33	7	23	74	0
	surprise	7	0	8	6	5	4	104
	TOTAL	210	30	158	271	253	188	142

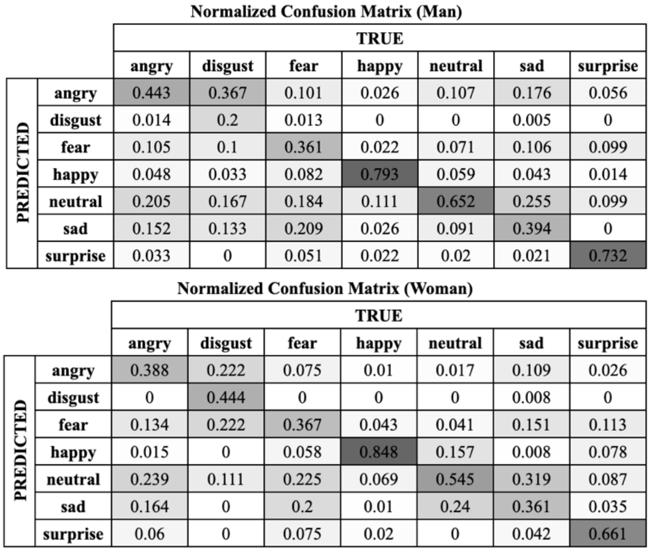
Confusion Matrix (Woman)

		TRUE						
		angry	disgust	fear	happy	neutral	sad	surprise
PREDICTED	angry	26	2	9	3	2	13	3
	disgust	0	4	0	0	0	1	0
	fear	9	2	44	13	5	18	13
	happy	1	0	7	257	19	1	9
	neutral	16	1	27	21	66	38	10
	sad	11	0	24	3	29	43	4
	surprise	4	0	9	6	0	5	76
	TOTAL	67	9	120	303	121	119	115

Fig. 1. Confusion matrix for men and women emotion detection accuracy using DeepFace.

B. Emotion Detection Accuracy

The overall accuracy for emotion detection in men was 51%, slightly lower than the 52% accuracy observed for women. Despite this marginal difference, a deeper analysis reveals some interesting trends. For men, three emotions—happiness, neutral, and surprise—were detected with over 50% accuracy, with happiness being the highest at 79%. In contrast, women had only one emotion, happiness, detected with an accuracy greater than 50%, reaching 85%.



Normalized Confusion Matrix (Man)

		TRUE						
		angry	disgust	fear	happy	neutral	sad	surprise
PREDICTED	angry	0.443	0.367	0.101	0.026	0.107	0.176	0.056
	disgust	0.014	0.2	0.013	0	0	0.005	0
	fear	0.105	0.1	0.361	0.022	0.071	0.106	0.099
	happy	0.048	0.033	0.082	0.793	0.059	0.043	0.014
	neutral	0.205	0.167	0.184	0.111	0.652	0.255	0.099
	sad	0.152	0.133	0.209	0.026	0.091	0.394	0
	surprise	0.033	0	0.051	0.022	0.02	0.021	0.732

Normalized Confusion Matrix (Woman)

		TRUE						
		angry	disgust	fear	happy	neutral	sad	surprise
PREDICTED	angry	0.388	0.222	0.075	0.01	0.017	0.109	0.026
	disgust	0	0.444	0	0	0	0.008	0
	fear	0.134	0.222	0.367	0.043	0.041	0.151	0.113
	happy	0.015	0	0.058	0.848	0.157	0.008	0.078
	neutral	0.239	0.111	0.225	0.069	0.545	0.319	0.087
	sad	0.164	0	0.2	0.01	0.24	0.361	0.035
	surprise	0.06	0	0.075	0.02	0	0.042	0.661

Fig. 2. Normalized confusion matrix for men and women emotion detection accuracy using DeepFace.

When analyzing the performance across specific emotions, notable differences emerged between the genders. Disgust was particularly challenging for the model, with women being identified with 44% accuracy, compared to only 20% for men, making it the lowest-performing emotion for men and more poorly detected than any emotion for women. This discrepancy could be partially attributed to the small sample size of images portraying disgust. Of the 39 total disgust images, a mere 9 of them were identified as women.

For both genders, happiness was the emotion with the highest accuracy, which may be due to the distinctiveness of a smile or the fact that happiness was the most represented emotion in the dataset, with 574 images of the 2106 detectable. Anger, neutral, and sadness were detected with higher accuracy in men, whereas disgust, fear, and happiness showed better accuracy for women. Of these top three emotions in women, fear and happiness had a relatively higher proportion of images of women (greater than 40% of the detectable images), suggesting that a larger dataset may improve prediction accuracy. Disgust does not follow this trend with an extremely small sample size for women, but surprise (while still more accurate in men) maintained a high accuracy for women at 66% and has the second largest proportion of women.

C. Data Quality and Model Limitations

An important aspect to consider is the large proportion of images (41%) where DeepFace was unable to detect a face. This could be due to several factors, including low image quality, a known limitation of DeepFace, and the minimal image preprocessing performed by the software. A visual inspection of the FER dataset revealed several images that likely contributed to this detection failure, including images of text, images without discernible faces, with cartoon faces, or with faces of infants where gender is indeterminate (Figure 3). Even further, some images display an emotion contrary to the FER label, according to this author's discernment (Figure 4). All of these discrepancies could limit the performance of the emotion detection.

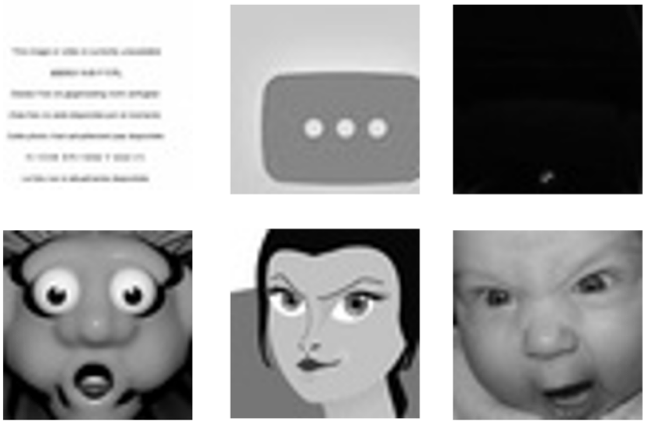


Fig. 3. Examples of photos from the FER dataset that may have been undetectable or have skewed the results of this study.

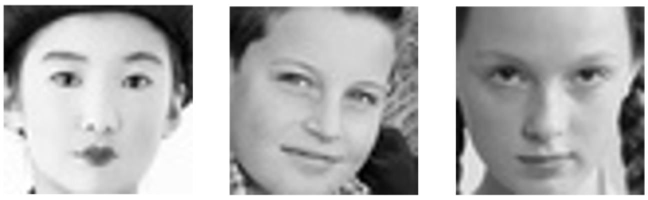


Fig. 4. Examples of photos from the FER dataset that are identified by FER as displaying an emotion (sad, neutral, and fear, from left to right) contrary to this author's interpretation (neutral, happy, and angry, from left to right).

Additionally, since gender was predicted using the same software as emotion detection, there is a possibility that inaccuracies in emotion prediction could also affect the gender identification accuracy, further complicating the interpretation of the results.

D. Summary

The results indicate that while DeepFace performs similarly across genders in terms of overall accuracy, certain emotions are more accurately detected in one gender over the other. The discrepancies observed, particularly in the accuracy of detecting emotions like disgust and fear, underscore the importance of considering dataset composition and quality when evaluating emotion detection models. Moreover, the substantial number of undetected faces and the overall accuracy rates hovering around 50% suggest limitations in DeepFace's ability to process this dataset effectively, highlighting the need for further refinement and preprocessing in future studies.

V. Conclusion

This study highlights the presence of gender-based disparities in the performance of DeepFace, a widely-used emotion recognition tool, when analyzing facial expressions. While the overall accuracy of emotion detection between men and women was relatively close—51% for men and 52% for women—our analysis revealed that the model's ability to detect specific emotions varied significantly between genders. Notably, emotions such as happiness and disgust were more accurately detected in women, while anger, neutral, and sadness were better detected in men.

These findings underscore the importance of evaluating and addressing potential biases in emotion recognition systems, particularly as these technologies become increasingly integrated into various societal applications. The imbalance in gender representation within the dataset, combined with the model's limitations in detecting certain emotions, suggests that more diverse and representative datasets are essential to improve accuracy and fairness. Moreover, the high rate of undetected faces (41%) indicates that the quality and preprocessing of images play a critical role in the efficacy of these systems.

The discrepancies observed in the detection of emotions like disgust, where the accuracy for men was particularly low, raise concerns about the reliability of emotion recognition tools in real-world applications, especially when the dataset is not sufficiently large or balanced. Given that disgust had the smallest sample size, future research should consider larger and more balanced datasets to ensure robust evaluations.

Overall, this study contributes to the growing body of research on bias in emotion recognition technology. It emphasizes the need for continued efforts to refine these systems, including better dataset curation, enhanced preprocessing techniques, and more sophisticated models that can equitably recognize emotions across different genders. As emotion recognition technology continues to evolve, it is crucial that these biases be addressed to ensure fair and accurate outcomes for all users.

REFERENCES

[1] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer and C. Busch, "Demographic Bias in Biometrics: A Survey on an Emerging Challenge," in IEEE Transactions on Technology and Society, vol. 1, no. 2, pp. 89-103, June 2020, doi: 10.1109/TTS.2020.2992344.

[2] M. Buntoun and F. Kuok, "Cross-race Effect in Face Recognition: An Observation on Southeast Asian," 2023 15th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Kuala Lumpur, Malaysia, 2023, pp. 75-80, doi: 10.1109/SKIMA59232.2023.10387297.

[3] P. Melzi et al., "Synthetic Data for the Mitigation of Demographic Biases in Face Recognition," 2023 IEEE International Joint Conference on Biometrics (IJCB), Ljubljana, Slovenia, 2023, pp. 1-9, doi: 10.1109/IJCB57857.2023.10449034.

[4] A. Sharma, V. Bajaj and J. Arora, "Machine Learning Techniques for Real-Time Emotion Detection from Facial Expressions," 2023 2nd Edition of IEEE Delhi Section Flagship Conference (DELCON), Rajpura, India, 2023, pp. 1-6, doi: 10.1109/DELCON57910.2023.10127369.

[5] J. Cai et al., "Identity-Free Facial Expression Recognition Using Conditional Generative Adversarial Network," 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 2021, pp. 1344-1348, doi: 10.1109/ICIP42928.2021.9506593.

[6] C. Wang, J. Zeng, S. Shan and X. Chen, "Multi-Task Learning of Emotion Recognition and Facial Action Unit Detection with Adaptively Weights Sharing Network," 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 56-60, doi: 10.1109/ICIP.2019.8802914.

[7] Wenfei Gu, Cheng Xiang, Y.V. Venkatesh, Dong Huang, Hai Lin, Facial expression recognition using radial encoding of local Gabor features and classifier synthesis, Pattern Recognition, Volume 45, Issue 1, 2012, Pages 80-91, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2011.05.006.

[8] A. Nethi et al., "Cohesive Group Emotion Recognition using Deep Learning," 2023 26th ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter), Taiyuan, China, 2023, pp. 264-269, doi: 10.1109/SNPD-Winter57765.2023.10466291.

[9] M. H. Shahbaz, Zain-Ul-Abidin, K. Mahboob and F. Ali, "Enhancing Contextualized GNNs for Multimodal Emotion Recognition: Improving Accuracy and Robustness," 2023 7th International Multi-Topic ICT Conference (IMTIC), Jamshoro, Pakistan, 2023, pp. 1-7, doi: 10.1109/IMTIC58887.2023.10178481.

[10] A. M. Cadayona, N. M. S. Cerilla, D. M. M. Jurilla, A. K. D. Balan and J. C. d. Goma, "Emotional State Classification: An Additional Step in Emotion Classification through Face Detection," 2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA), Tokyo, Japan, 2019, pp. 667-671, doi: 10.1109/IEA.2019.8715171.

[11] Rhue, Lauren, Racial Influence on Automated Perceptions of Emotions (November 9, 2018). Available at SSRN: https://ssrn.com/abstract=3281765 or http://dx.doi.org/10.2139/ssrn.3281765.

[12] Fan, Y., Lam, J.C.K. & Li, V.O.K. Demographic effects on facial emotion expression: an interdisciplinary investigation of the facial action units of happiness. Sci Rep 11, 5214 (2021). https://doi.org/10.1038/s41598-021-84632-9.

[13] V. Suresh and D. C. Ong, "Using Positive Matching Contrastive Loss with Facial Action Units to mitigate bias in Facial Expression Recognition," 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII), Nara, Japan, 2022, pp. 1-8, doi: 10.1109/ACII55700.2022.9953865.

[14] "Challenges in Representation Learning: A report on three machine learning contests." I Goodfellow, D Erhan, PL Carrier, A Courville, M Mirza, B Hamner, W Cukierski, Y Tang, DH Lee, Y Zhou, C Ramaiah, F Feng, R Li, X Wang, D Athanasakis, J Shawe-Taylor, M Milakov, J Park, R Ionescu, M Popescu, C Grozea, J Bergstra, J Xie, L Romaszko, B Xu, Z Chuang, and Y. Bengio. arXiv 2013.