Investigating Bias in Affective State Detection Using Eye Biometrics*

Yuxin Zhi[1,*], Bilal Taha[1] and Dimitrios Hatzinakos[1]

[1]University of Toronto, ON, Canada

Abstract

This study delves into the exploration of pupillometry as a modality for affect state recognition. It examines the propensity for bias in both feature-based and learning-based machine learning models that interpret affect through pupil responses. Our research lies at the intersection of affective computing and mental health, recognizing the paramount importance of accurately identifying affect states for effective mental health interventions. We rigorously evaluate the performance of these pupillometry-based models across diverse demographic groups, including variables such as ethnicity, gender, age, vision problems, and iris color. Our findings reveal notable disparities, particularly in gender and ethnicity. Bias levels are pronounced in both feature-based and learning-based models, with F1 score differentials reaching up to 36.28%. Additionally, our analysis uncovers a slight bias related to iris color, significantly impacting the efficacy of affect state recognition models that rely on pupil responses. This underscores the critical need for fairness and accuracy in developing machine learning models within affective computing. By highlighting these areas of potential bias, our study contributes to the broader discourse on creating equitable AI systems and advancing mental health care, education, and social robotics. It emphasizes the ethical imperative of developing unbiased, inclusive technologies in healthcare systems.

Keywords

Pupillometry, Affect state recognition, Mental health interventions, Bias, Fairness

1. Introduction

In the emerging field of affective computing and mental health, the intricate relationship between affect state recognition and cognitive and mental health outcomes presents a domain of significant research interest [1, 2]. Affect state recognition, central to understanding and managing various mental health disorders, encompasses the complex process of identifying and interpreting emotional states [3]. This process is crucial in disorders such as depression and anxiety, where impairments in emotional awareness and regulation are prevalent [4]. The advancement of cognitive and mental health therapies, including cognitive-behavioral therapy (CBT) and mindfulness-based strategies, hinges on the nuanced understanding and regulation of affect states [5, 6]. These emotional states profoundly influence core cognitive processes, including attention, memory, and decision-making [7]. This underscores the importance of affect state recognition in therapeutic interventions. Furthermore, the predictive nature of affect state recognition in mental health conditions paves the way for early and more effective intervention strategies [8].

The advent of technological solutions, such as recognition software and mood-predicting algorithms, has opened new avenues in the monitoring and treatment of mental health conditions [9]. However, this brings forth the challenge of bias in machine learning models [10]. The accuracy and reliability of these models in affect state recognition are paramount, as biases can lead to misinterpretations, potentially worsening mental health conditions or leading to inappropriate treatment methodologies.

Pupil response has been employed in diverse studies within psychiatry and psychology, particularly in assessing cognitive load for memory-based tasks [11]. It has also been utilized in analyzing the emotional impact of stimuli on individuals [12]. One investigation focused on the confounding effects of eye blinking in pupillometry and proposed remedies [13]. Additionally, the utility of pupillometry in psychiatry was reviewed, highlighting its role in understanding patients' information processing styles, predicting treatment outcomes, and examining cognitive functions [14]. A separate study employed pupillometry to assess atypical pupillary light reflexes and the LC-NE system in Autism Spectrum Disorder (ASD)[15]. The potential clinical use of pupillometry in diagnosing nonconvulsive status epilepticus (NCSE) has also been explored[16]. Although physiological responses such as pupillometry are generally considered less biased than other modalities, hidden biases can emerge from factors like stimuli selection and demographic influences [17, 18]. For instance, responses to visual stimuli may vary significantly across different cultural backgrounds, orientations, and age groups.

Machine Learning for Cognitive and Mental Health Workshop (ML4CMH), AAAI 2024, Vancouver, BC, Canada

*Corresponding author.

yuxin.zhi@mail.utoronto.ca (Y. Zhi); bilal.taha@mail.utoronto.ca (B. Taha); dimitris@comm.utoronto.ca\protect\protect\leavevmode@ifvmode\kern+.1667em\relax (D. Hatzinakos)

This work aims to investigate the bias that exists in affect state recognition models based on physiological signals, specifically pupillometry, which plays a significant role in understanding cognitive and mental health applications. The main goal of this study is to shed light on the potential bias that may exist in common learning methods. The structure of the paper is as follows: first, the methodology, which includes preprocessing and learning models, is explained. Then, the experiments and results are presented and validated using a dataset collected for this work. Finally, we discuss the findings and conclude at the end.

2. Methodology

The framework focuses on the use of pupillary responses and approaches the task of affect state recognition as a binary classification problem based on the targeted group. The first step is preprocessing the pupillometry data to mitigate the effect of noisy samples. Then, the data is used to develop the classification model either from handcrafted features specific to the pupillometry data or using a learned model. Finally, model training and testing are described at the end to investigate the different cases.

2.1. PreProcessing

The initial processing of the pupillometry data is paramount to remove any irrelevant and noisy samples that may impact pupil size analysis. The raw data can be contaminated with various outliers like system errors, blinks, eye-tracker glitches, and eyelid occlusion, which can be identified and eliminated during this stage. Previous studies [19, 20] have proposed a robust method for detecting such invalid samples, which we have used in our study. The method uses dilation speed as a metric to determine whether a data point is an outlier. If a data sample exhibits a dilation speed greater than a pre-defined threshold, it is removed as an anomaly. After that, to ensure the continuity of the data, the filtered data is modeled using a Gaussian process.

2.2. Feature-Based Models

The feature-based method is a common approach in machine learning where specific features are extracted from the data and used to train a the algorithm. In this study, the pupil responses for each participant were divided into 150 sets of sequences, with each sequence corresponding to the pupil response for each image. Each sequence has a length of 300 samples, which were used to extract the features.

Several features can be extracted from the pupil response, including mean and variance of the pupil response, maximum dilation, minimum contraction, dilation speed, dilation duration, contraction duration, and the difference between dilation and contraction. In total, 30 features were manually extracted and used to train a kernel SVM classifier. Different kernels were tested, and the Gaussian Kernel showed the best performance in general.

2.3. Learned-Based Model

The long short-term memory (LSTM) [21] model is commonly used in machine learning for modeling sequential data. In this approach, the LSTM model has been implemented as seen in Figure1 with 128 LSTM units, a dropout rate of 0.5, a 128-unit dense layer, and a rectified linear unit (ReLU) activation function. Finally, a dense layer at the end is added with a SoftMax function to produce the classification output. The cross-entropy loss function and RMSprop optimizer are used for training the model.

The use of deep learning methods such as LSTM for feature learning and affect state recognition is effective in various machine learning tasks. This approach can improve the performance of the model, as it can capture temporal dependencies and relationships in the data that might be missed by manual feature extraction.

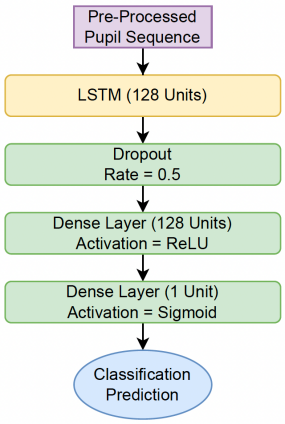

Figure 1: The LSTM structure used for modeling the learned-based approach.

2.4. Model Training

In both approaches, feature-based and learned-based, we divided the data into training and testing datasets, allocating 80% to training and 20% to testing, respectively. While constructing the model, we utilized data from all demographic groups with the intention of creating a model that

captures feature representations from all these groups. To assess the model's fairness and prevent bias towards any particular group, we further divided the testing data into subgroups during the evaluation phase and assessed the model's performance for each subgroup.

Due to the limited number of samples, we introduced augmentation to enhance the training data. This augmentation was applied later in the evaluation, allowing us to assess its impact on the results. The pupil data sequences were augmented using noise injection and time-shifting methods [22]. Specifically, we added white noise to the original pupil data and performed 50 sample shifts. Importantly, the augmentation was applied to samples from the non-dominant group to ensure that our findings were not influenced by this imbalance.

3. Experiments and Results

Bias can be seen as the disparity in performance metrics across different groups for a given task. Assuming we have $G = \{g_1, g_2, \dots, g_n\}$ be the set of groups for bias investigation. For each group g_i, we compute the performance metrics of a recognition model $M(g_i)$. Then, the bias B is identified for a pair of groups (g_i, g_j) as the absolute difference in their metrics:

$$B(g_i, g_j) = |M(g_i) - M(g_j)|$$

3.1. Data

To conduct a thorough assessment of bias in pupillometry affect state recognition, we collected a dataset that encompasses pupillometry data in response to visual stimuli, taking into account a diverse range of demographics. The study involved 35 university students aged between 18 and 40 years, with a mean age of 24.6 and a standard deviation of 5.17. Participants were required to have no history of vision disorders, and they were also asked about any medications they might be taking that could affect their responses, such as depression medication. The data collected from the participants is categorized into different cases based on various demographic factors:

- *Gender*: This case examines the algorithm's ability to fairly recognize emotional states in females versus males.
- *Ethnic Group*: This case assesses the model's ability to impartially detect emotional states based on participants' ethnic groups, including Asian (Chinese), White (North American or European), Black (African American or Caribbean), and South Asian (Pakistani or Indian) [6].
- *Age*: This case explores the impact of age on the model's ability to detect emotional states, considering age groups [17-24] versus [25-55].
- *Iris Color*: The eye color case is a unique factor relevant to models using pupillometry data for their applications. Eye color affects the precision of detecting pupils and measuring their dilation and contraction. Thus, we categorize the data into light (light brown, green, blue, hazel) versus dark (black, brown, dark brown) iris colors.
- *Vision*: This case evaluates the model's effectiveness in capturing emotional states in data from individuals wearing glasses versus those not wearing glasses.

3.2. Experimental Protocol

The proposed system was evaluated using a dataset collected at the University of Toronto. In the experiment, participants viewed a series of visual stimuli intended to elicit emotions spanning different valence and arousal values. The visual stimuli were selected from the International Affective Picture System (IAPS) dataset [23]. The IAPS database provides normative ratings of emotional valence and arousal for a large set of images. The rating scales are based on the Self-Assessment Manikin (SAM), a 9-point rating scale where a score of 9 represents a high rating (i.e., high pleasure, high arousal), a score of 5 indicates a neutral rating, and a rating of 1 represents a low rating (i.e., low pleasure, low arousal).

The selected visual stimuli elicit the emotions of interest, which include the two quadrants of the VA dimensional model (i.e., HA, LA, or HV, LV). Each of the aforementioned emotional states is achieved by displaying 30 images of the same emotional target for 5 seconds each. The images were selected to statistically produce the same response for different groups of people. All images were presented on a screen with a resolution of 1920 by 1080 pixels. Following the recommendations of the device manufacturers, the Gazepoint eye-tracking system was placed approximately 45 cm in front of the participant at an angle of around 30 degrees. The total number of participants was 35.

The data collection process was approved by the research ethics committee at the University of Toronto. All participants signed a consent form that clearly explained the data collection procedure and the privacy of their data. Furthermore, all participants received compensation in the form of a gift card.

3.3. Metrics:

In the evaluation process, two common metrics were employed: accuracy and F1 score. Accuracy gauges the proportion of correct predictions made by the algorithm. The F1 score, on the other hand, assesses the balance between precision and recall. It offers a more nuanced

evaluation of the algorithm's performance, especially when dealing with imbalanced datasets.

3.4. Results from the Feature-Based Model

We employed a feature-based algorithm for emotion recognition and assessed the presence of bias among different demographic groups, focusing on valence-based and arousal-based classifications. Our evaluation yielded results presented in Tables 1 and 2, along with Figures 2 and 3.

Notably, our findings reveal significant performance differences between males and females in both arousal and valence. Specifically, our analysis indicated that males scored 20.28% higher in arousal and 17.46% higher in valence compared to females. The F1 score exhibited a similar gender-based pattern of differences.

Further examination of the model based on ethnicity factors showed significant variations in accuracy and F1 scores across different groups. Notably, the Asian group, despite having the highest number of samples, displayed the lowest accuracy and F1 scores in terms of arousal classification. In contrast, the South Asian group, with the second-lowest number of samples, demonstrated the highest performance. The percentage difference between the highest-performing group (South Asian) and the lowest-performing group (Asian) was 28.93% in accuracy and 21% in F1 score for arousal classification. These findings suggest that obtaining accurate feature representations for the Asian group in terms of arousal classification may be more challenging based on the provided stimuli.

Regarding valence classification, our analysis revealed similar performance among the Asian, White, and South Asian groups, while the Black group exhibited significantly lower accuracy and F1 scores. Specifically, the percentage difference between the Black group and the group with the highest performance was 26.99% in accuracy and 46.71% in F1 score, respectively.

Table 1
Arousal Result of SVM for the different Ethnic Groups.

Ethnic Group	Testing %	Accuracy	F1 Score
Asian	44.24%	51.48%	0.667
White	35.86%	67.88%	0.790
South Asian	11.78%	68.89%	0.816
Black	8.12%	64.5%	0.784

3.5. Results from the LSTM Model

We employed an LSTM-based approach to investigate bias across different demographic groups. The results of our analysis are presented in Tables 3 and 4, and Figures 4 and 5.

Consistent with the findings of the feature-based model, we observed significant performance differences between genders and ethnic groups. Specifically, our results revealed a significant 24.12% bias toward females

Table 2
Valence Result of SVM for the different Ethnic Groups.

Ethnic Group	Testing %	Accuracy	F1 Score
Asian	45.28%	52.4%	0.615
White	37.47%	51.1%	0.600
South Asian	9.43%	54.3%	0.667
Black	7.82%	41.4%	0.414

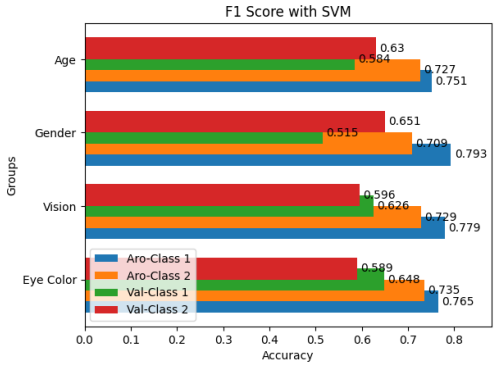

Figure 2: SVM F1 results for the remaining groups

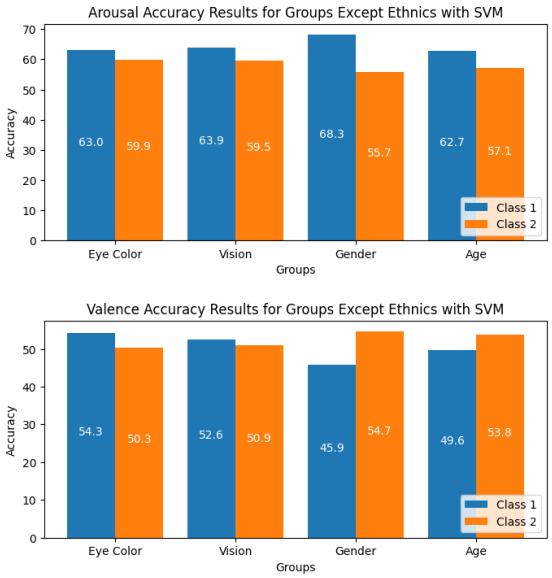

Figure 3: SVM Accuracy results for the remaining groups.

in arousal accuracy and a 10.15% bias toward males in valence accuracy. Concerning ethnic groups, accuracy exhibited substantial variations across different ethnicities, as depicted in Tables 3 and 4. In terms of arousal, the Asian group had the lowest performance, while the White group achieved the highest accuracy, resulting in a significant 20.90% advantage favoring the White group. The other ethnic groups showed similar performance. In terms of valence, the Black group displayed the lowest performance, while the South Asian group achieved the highest, with a difference of 36.28%. In the remaining cases, there were no significant differences between individual groups, suggesting that these factors share common representations that can be captured by the algorithms.

Table 3
LSTM Arousal Result for different Ethnic Groups.

Ethnic Group	Accuracy	F1 Score
Asian	52.7%	0.413
White	65.0%	0.581
South Asian	62.2%	0.546
Black	64.5%	0.506

Table 4
LSTM Valence Result for different Ethnic Groups.

Ethnic Group	Accuracy	F1 Score
Asian	54.7%	0.404
White	50.4%	0.345
South Asian	54.3%	0.382
Black	37.9%	0.209

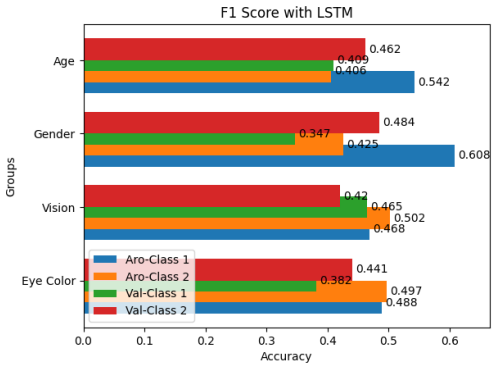

Figure 4: F1 results for the LSTM model for the remaining demographic groups.

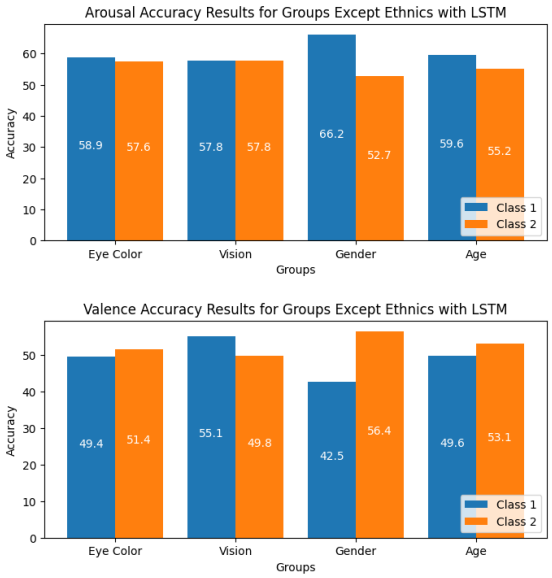

Figure 5: LSTM Accuracy results for the remaining demographic groups

3.6. Bias and Fairness

Based on the results presented above, it is evident that both models exhibit significant differences in accuracy and F1 scores concerning ethnic groups and gender. This indicates that these two factors play a pivotal role in the development of affect recognition from pupillometry data, as the models struggled to find effective representations for them. In contrast, the other four cases displayed minor differences in terms of accuracy and F1 scores, suggesting that these factors share common representations across all groups and do not adversely affect the data's quality. For example, iris color had a limited impact on recognition performance, albeit not as pronounced as with gender and ethnic groups.

Despite the dataset including diverse groups during model training, the quality of the representations failed to adequately capture the diverse group responses within the studied population. We acknowledge that the unbalanced number of samples in each group might contribute to the bias observed in the results. To address this potential issue, we implemented data augmentation techniques (see 2.4) for the non-dominant groups (groups with fewer samples) to increase their sample size. Subsequently, we followed the same procedure as in the original case. However, our results demonstrated that even with the implementation of data augmentation, the performance did not change significantly. The bias in performance persisted in both the ethnic groups and gender-based cases, while the remaining cases exhibited similar performance.

4. Conclusion

In this study, we investigated the performance of feature-based and learned-based affect recognition models across various group factors, including ethnicity, gender, vision, iris color, and age, focusing on pupillometry as a the modality. Our research, involving a dataset from 35 diverse participants, revealed significant gender and ethnic biases in standard affect recognition algorithms, impacting both arousal and valence-based classifications. We also identified minor biases related to other factors, such as iris color. These findings emphasize the potential bias in affect recognition systems, highlighting the need for more inclusive and representative training data, rigorous fairness evaluation, and enhanced transparency in model development. Our study not only sheds light on the inherent biases in affective computing but also underscores the importance of considering demographic factors in the development of more equitable and effective affect recognition technologies, particularly given their direct relation to cognitive and mental health.

References

[1] R. Assabumrungrat, S. Sangnark, T. Charoenpattarawut, W. Polpakdee, T. Sudhawiyangkul, E. Boonchieng, T. Wilaiprasitporn, Ubiquitous affective computing: A review, IEEE Sensors Journal 22 (2021) 1867–1881.

[2] S. Greene, H. Thapliyal, A. Caban-Holt, A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health, IEEE Consumer Electronics Magazine 5 (2016) 44–56.

[3] R. A. Calvo, K. Dinakar, R. Picard, P. Maes, Computing in mental health, in: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2016, pp. 3438–3445.

[4] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, M. Berk, Affective and content analysis of online depression communities, IEEE transactions on affective computing 5 (2014) 217–226.

[5] C. Zucco, B. Calabrese, M. Cannataro, Sentiment analysis and affective computing for depression monitoring, in: 2017 IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE, 2017, pp. 1988–1995.

[6] M. A. Kirk, B. Taha, K. Dang, H. McCague, D. Hatzinakos, J. Katz, P. Ritvo, A web-based cognitive behavioral therapy, mindfulness meditation, and yoga intervention for posttraumatic stress disorder: Single-arm experimental clinical trial, JMIR Mental Health 9 (2022) e26479.

[7] J. S. Lerner, D. Keltner, Beyond valence: Toward a model of emotion-specific influences on judgement and choice, Cognition & emotion 14 (2000) 473–493.

[8] R. e. Kaliouby, R. Picard, S. Baron-Cohen, Affective computing and autism, Annals of the New York Academy of Sciences 1093 (2006) 228–248.

[9] M. Nouman, S. Y. Khoo, M. P. Mahmud, A. Z. Kouzani, Recent advances in contactless sensing technologies for mental health monitoring, IEEE Internet of Things Journal 9 (2021) 274–297.

[10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.

[11] E. Granholm, R. F. Asarnow, A. J. Sarkin, K. L. Dykes, Pupillary responses index cognitive resource limitations, Psychophysiology 33 (1996) 457–461.

[12] M. M. Bradley, L. Miccoli, M. A. Escrig, P. J. Lang, The pupil as a measure of emotional arousal and autonomic activation, Psychophysiology 45 (2008) 602–607.

[13] K. Yoo, J. Ahn, S.-H. Lee, The confounding effects of eye blinking on pupillometry, and their remedy, Plos one 16 (2021) e0261463.

[14] S. Graur, G. Siegle, Pupillary motility: bringing neuroscience to the psychiatry clinic of the future, Current neurology and neuroscience reports 13 (2013) 1–9.

[15] G. Lynch, Using pupillometry to assess the atypical pupillary light reflex and lc-ne system in asd, Behavioral Sciences 8 (2018) 108.

[16] S. Hocker, Pupillometry for diagnosing nonconvulsive status epilepticus and assessing treatment response?, Neurocritical Care 35 (2021) 304–305.

[17] K. Yang, C. Wang, Y. Gu, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, J. Goncalves, Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition, IEEE Transactions on Affective Computing (2021).

[18] H.-C. Yang, C.-C. Lee, Annotation matters: A comprehensive study on recognizing intended, self-reported, and observed emotion labels using physiology, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2019, pp. 1–7.

[19] M. E. Kret, E. E. Sjak-Shie, Preprocessing pupil size data: Guidelines and code, Behavior research methods 51 (2019) 1336–1342.

[20] B. Taha, M. Kirk, P. Ritvo, D. Hatzinakos, Detection of post-traumatic stress disorder using learned time-frequency representations from pupillometry, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 3950–3954.

[21] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[22] T. Ko, V. Peddinti, D. Povey, S. Khudanpur, Audio augmentation for speech recognition, in: Sixteenth annual conference of the international speech communication association, 2015.
[23] P. J. Lang, International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual, Technical report (2005).