UNIVERSITY OF TORONTO

Investigating Bias in Affective State Detection Using Eye Biometrics

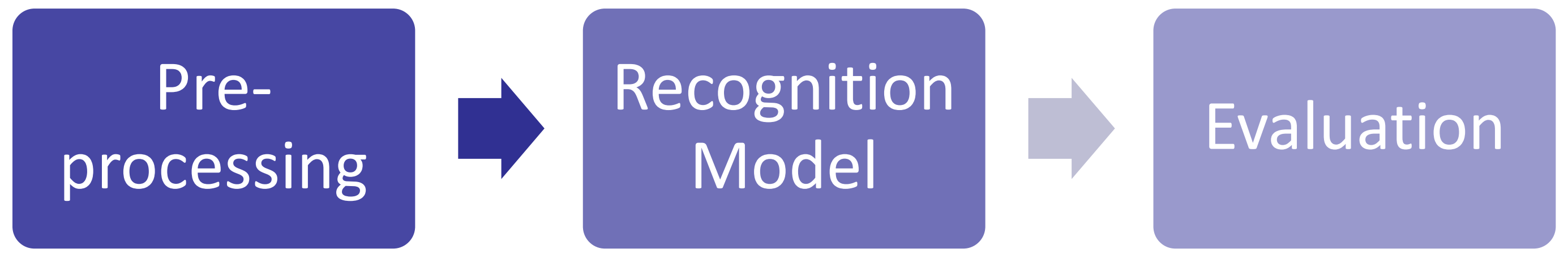
Yuxin Zhi, Bilal Taha, Dimitrios Hatzinakos

Introduction

- In the emerging field of affective computing and mental health, the interplay between affective states and cognitive health has sparked considerable research interest.
- Recognizing affective states is crucial for understanding and addressing mental health disorders, such as depression and anxiety
- The success of cognitive and mental health therapies, like cognitive behavioral therapy (CBT), hinges on the comprehension and management of affective states. These states significantly influence cognitive processes.
- Emotion recognition algorithms offer promising avenues for mental health monitoring and treatment. However, they also present challenges related to bias in machine learning models.
- Pupillometry, a widely used tool in psychiatry and psychology, holds the potential to enhance our understanding of cognitive functions and diagnose diseases.
- This study aims to explore biases in affective state recognition models, with a focus on pupillometry, to understand the factors that affect the performance of such models.

Methodology

We focused on the use of pupillary responses as a binary classification problem based of valence and arousal levels.


Pre-processing → Recognition Model → Evaluation

Pre-processing

- Data Cleaning: Remove noisy samples.
- Interpolation: Ensure the continuity of the data.

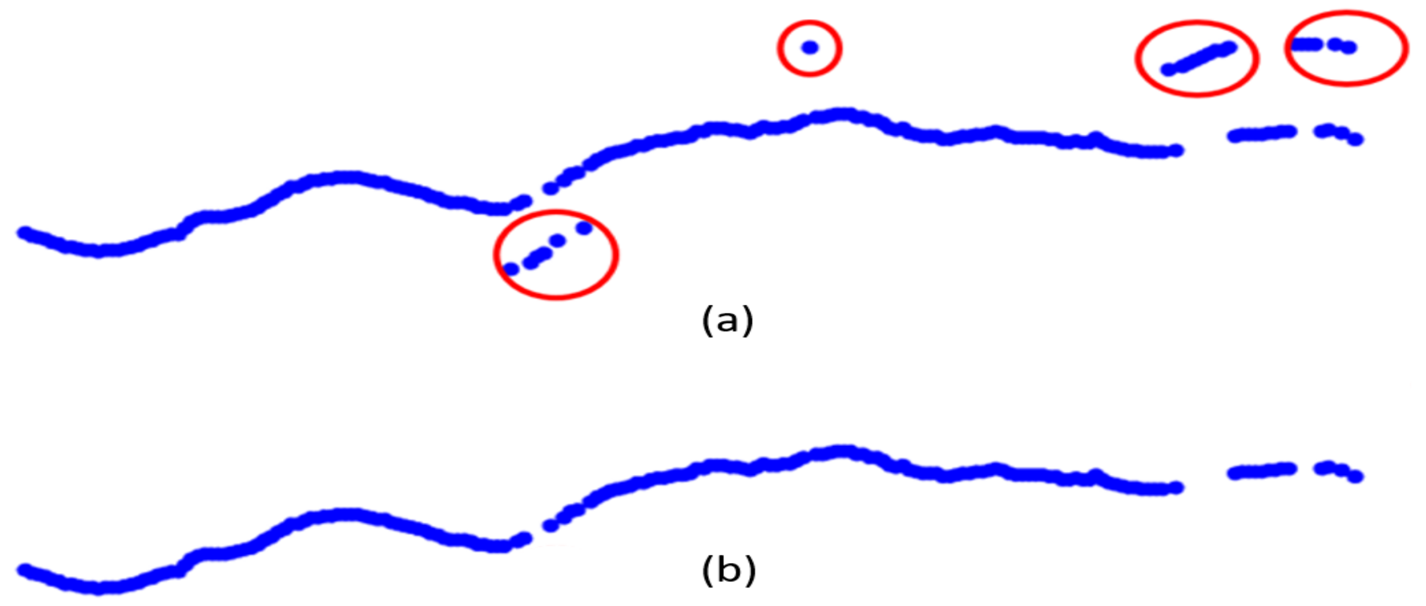


Figure 1: The output of the data cleaning stage, with (a) depicting the data before and (b) after the cleaning process.

Recognition Models

- Feature-Based Models

- In total, 30 features were manually extracted and used to train a kernel SVM classifier.
- Several features have been 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.
- Different kernels were tested, and the Gaussian Kernel showed the best performance in general.

- Learned-Based Model

- The long short-term memory (LSTM) model is used to model the pupillary responses. The use of deep learning methods such as LSTM for feature learning and affect state recognition is effective in various machine learning tasks.

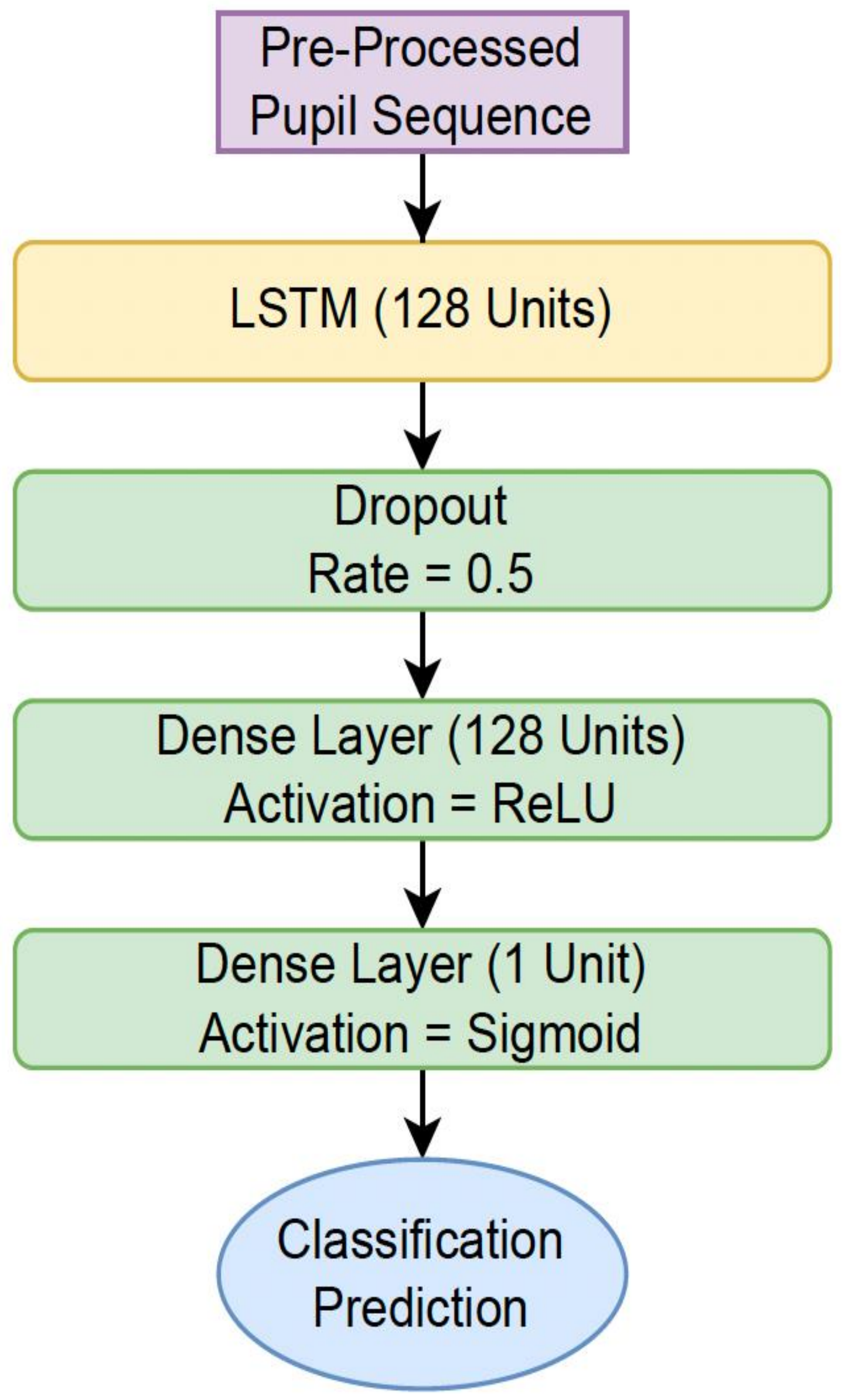

Pre-Processed Pupil Sequence → LSTM (128 Units) → Dropout Rate = 0.5 → Dense Layer (128 Units) Activation = ReLU → Dense Layer (1 Unit) Activation = Sigmoid → Classification Prediction

Figure 2: The LSTM structure used for modeling the learned-based approach.

Experiments and Results

Bias can be seen as the disparity in performance metrics across different groups for a given task. Assuming we have G = $\{g1, g2, \ldots, gn\}$ as the set of groups for bias investigation, for each group gi, we compute the performance metrics of a recognition model $M(gi)$. Then, the bias B is identified for a pair of groups (gi, gj) as the absolute difference in their metrics.

$$B(gi, gj) = |M(gi) - M(gj)|$$

Metrics:

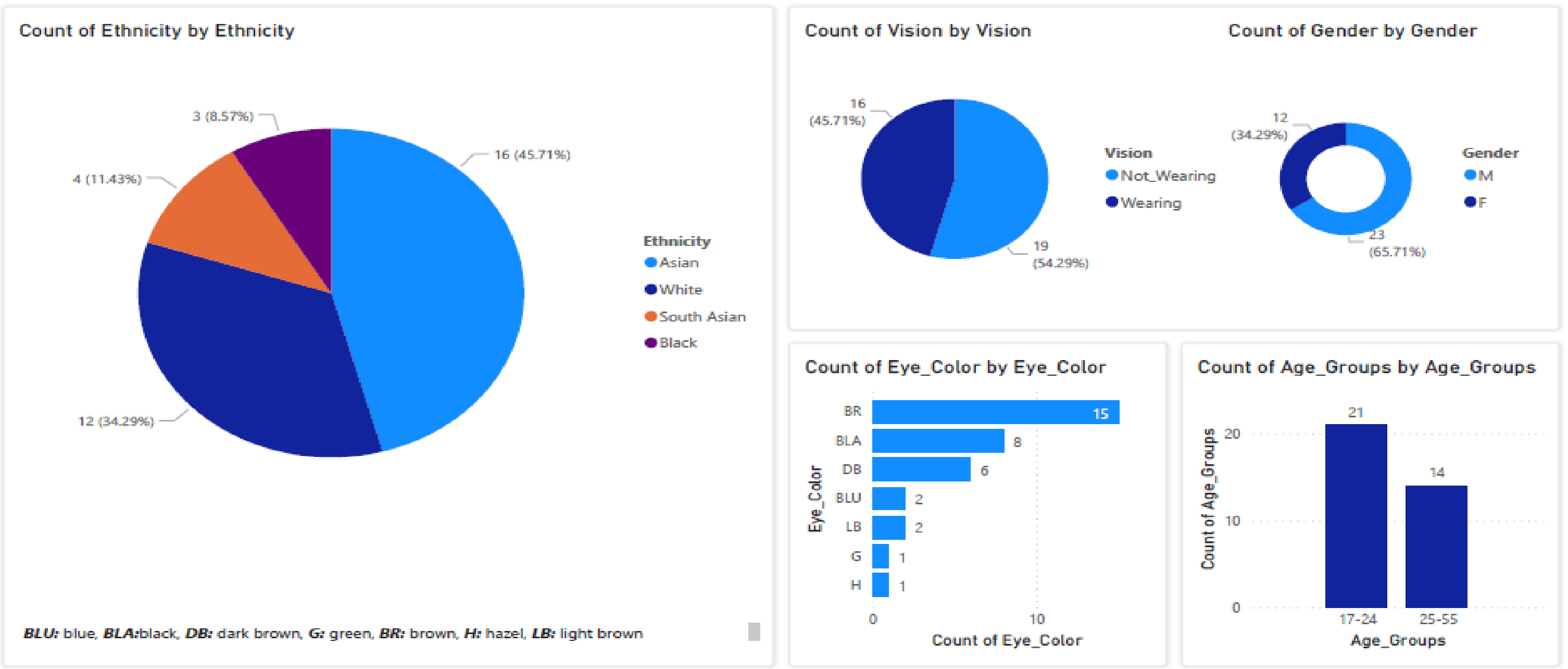
Two primary metrics, accuracy, and F1 score, were utilized for evaluation.

$$Accuracy = (number\ of\ correct\ predictions) / (total\ number\ of\ predictions)$$

$$F1\ score = 2 * (precision * recall) / (precision + recall)$$

Data

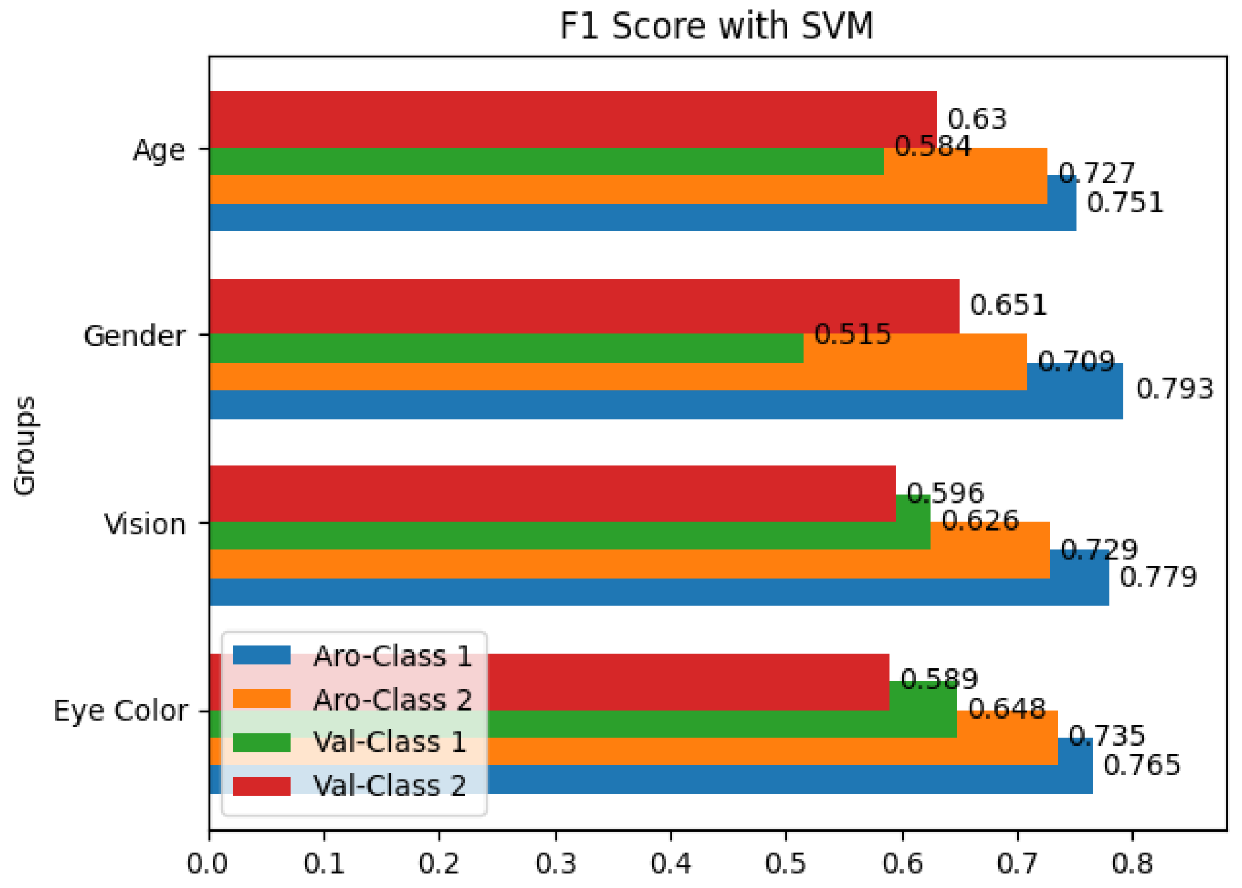
A comprehensive dataset of pupillometry affect state recognition was collected, covering a range of demographic factors, from 35 participants.



BLU: blue, BLA:black, DB: dark brown, G: green, BR: brown, H: hazel, LB: light brown

Results from the Feature-Based Model

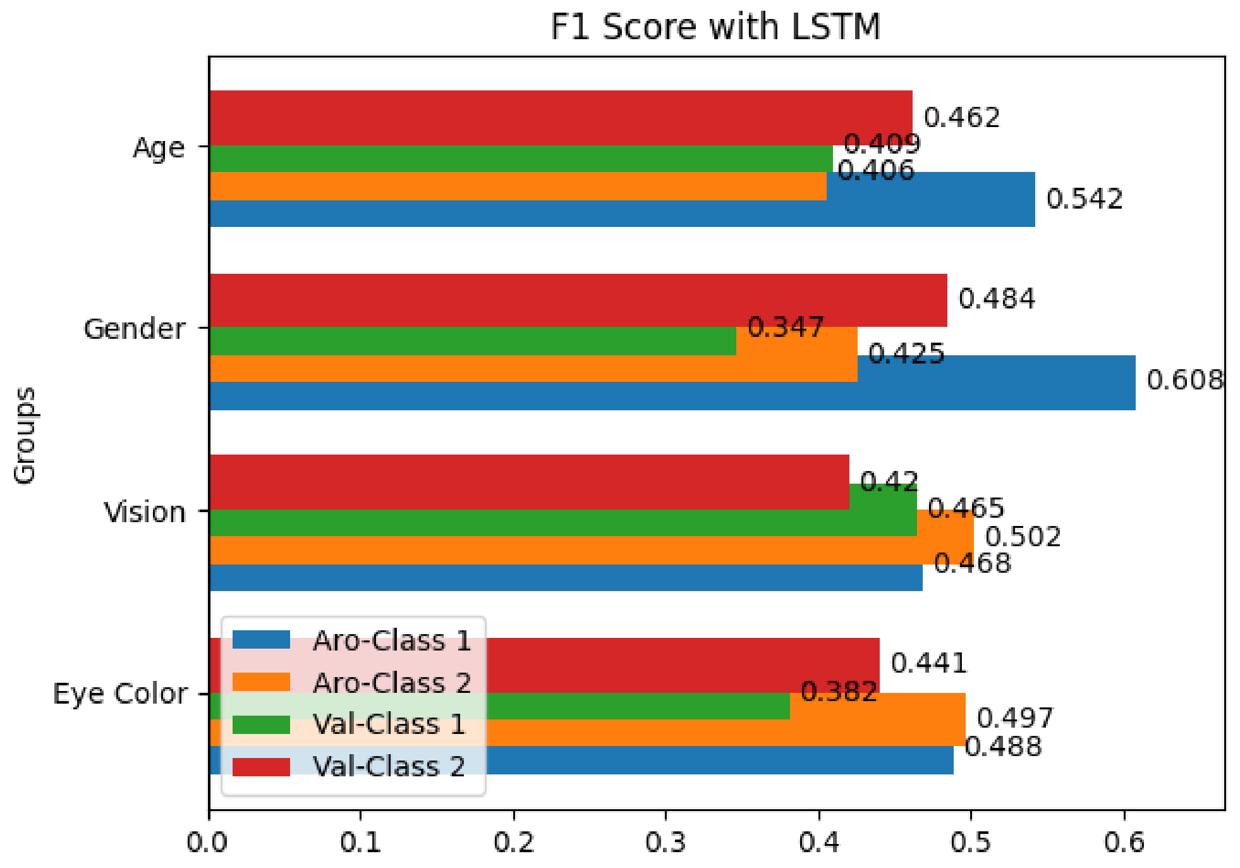
- Results from using the feature based model on different demographic groups in terms of F1 score.
- The gender group shows a notable difference in performance.


F1 Score with SVM

Groups	Aro-Class 1	Aro-Class 2	Val-Class 1	Val-Class 2
Age	0.751	0.727	0.584	0.63
Gender	0.793	0.709	0.515	0.651
Vision	0.779	0.729	0.626	0.596
Eye Color	0.765	0.735	0.648	0.589

Results from the LSTM Model

- Results from using the LSTM model on different demographic groups in terms of F1 score.
- The gender group shows a notable difference in performance.


F1 Score with LSTM

Groups	Aro-Class 1	Aro-Class 2	Val-Class 1	Val-Class 2
Age	0.542	0.406	0.409	0.462
Gender	0.608	0.425	0.347	0.484
Vision	0.468	0.502	0.465	0.42
Eye Color	0.488	0.497	0.382	0.441

Conclusion

- Our research has uncovered biases related to gender and ethnicity in standard affect recognition algorithms, affecting both arousal and valence classification.
- We identified smaller biases associated with other factors, including iris color.
- These findings underscore the presence of potential biases in affect recognition systems and highlight the importance of using more inclusive and representative training data.