

# Beyond Questionnaires: Video Analysis for Social Anxiety Detection

Nilesh Kumar Sahu  
IISER Bhopal  
Bhopal, India  
nilesh21@iiserb.ac.in

Nandigramam Sai  
Harshit  
IISER Bhopal  
Bhopal, India  
saiharshit3105@gmail.com

Rishabh Uikey  
IISER Bhopal  
Bhopal, India  
rishabhuikey@gmail.com

Haroon R. Lone  
IISER Bhopal  
Bhopal, India  
haroon@iiserb.ac.in

## Abstract

Social Anxiety is an anxiety disorder that affects millions worldwide, often leading to significant personal and social challenges. However, existing detection methods—primarily interviews and self-reported assessments—are resource-intensive, subjective, time-consuming, and prone to recall bias, creating a significant gap in timely mental healthcare. To address this gap, we explore the potential of smartphone cameras combined with multimodal machine learning for non-intrusive anxiety detection. We collected video data from 92 participants delivering impromptu speeches in an anxiety-provoking setting and extracted fine-grained behavioral features capturing head movement, eye gaze, upper body posture, and facial action units. We trained machine learning and deep learning models on various unimodal and multimodal feature combinations and achieved a best-performing accuracy of 74% in distinguishing SAD from non-SAD participants. Our work demonstrates the feasibility of leveraging ubiquitous smartphone cameras for non-intrusive, real-time SAD detection with non-verbal cues, with implications for scalable early screening and intervention.

## ACM Reference Format:

Nilesh Kumar Sahu, Nandigramam Sai Harshit, Rishabh Uikey, and Haroon R. Lone. 2018. Beyond Questionnaires: Video Analysis for Social Anxiety Detection. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 6 pages. <https://doi.org/XXXXXXX.XXXXXXX>

## 1 Introduction

Social Anxiety Disorder (SAD) is a prevalent mental health condition affecting 17% of youths globally [25]. Characterized by fear of negative evaluation in social settings, SAD can significantly impair relationships, education, and work performance — yet often goes undiagnosed or untreated [1].

Current SAD assessments rely on clinician-administered interviews or self-reported questionnaires such as Depression Anxiety Stress Scale (DASS), Social Interaction Anxiety Scale (SIAS), and

Social Phobia Inventory (SPIN). However, these methods face challenges of recall bias, subjective interpretation, and limited scalability beyond clinical settings [2, 24].

Video-based behavioral analysis offers a promising, non-invasive alternative for scalable SAD detection. Prior studies have shown success in detecting mental health conditions like depression [11, 16] and stress [12, 27] from behavioral cues. However, SAD presents unique challenges due to its context-driven, socially triggered manifestations. While isolated video features (e.g., eye gaze) have been explored [5, 17, 26, 28], the potential of fusing multiple behavioral modalities for SAD detection remains underexplored.

In this work, we propose a video-based framework for momentary anxiety detection that leverages multi-modal behavioral feature fusion. We recorded participants' videos during a speech-based stress task modeled after the Trier Social Stress Test (TSST), using a low-cost smartphone camera setup. Additionally, we developed a custom questionnaire specifically designed to assess anxiety experienced during the task, focusing on overall anxiety, perceived performance, avoidance behavior, feeling scrutinized, and fear of negative evaluation, with responses rated on a 5-point Likert scale. Unlike existing trait-based SAD questionnaires (e.g., SIAS, DASS, SPIN), which do not capture momentary fluctuations in anxiety, our questionnaire is designed to assess state anxiety. From the video recordings, we extracted a variety of non-verbal behavioral features, including head pose, body posture, facial action units, and eye gaze. After performing feature correlation analysis to reduce redundancy, we evaluated several machine learning models, achieving 74% classification accuracy in distinguishing SAD from non-SAD participants.

Our key contributions are: (i) We present a multi-modal video-based framework for SAD detection, fusing behavioral cues such as eye gaze, head pose, body posture, and facial action units. (ii) We demonstrate the use of a low-cost, everyday smartphone camera for anxiety detection.

## 2 Related work

Individuals experiencing mental distress often exhibit changes in non-verbal behaviors that can be captured through visual sensing technologies [13]. Prior research has explored the use of facial and behavioral cues to assess mental health conditions, including stress, depression, and anxiety, across settings such as interviews, public speaking, conversations, and task-based evaluations [27]. For example, controlled laboratory studies have induced stress through external stressors and analyzed participants' behavioral responses using video recordings [12]. Other work has examined stress assessment from facial videos collected during online meetings and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06

<https://doi.org/XXXXXXX.XXXXXXX>

social interactions [7, 16]. Similarly, Horigome et al. [16] analyzed upper-body behaviors during conversations and reported significant differences in head movement between the depressive and control groups.

Recent studies have increasingly adopted multimodal approaches that combine multiple visual cues. Sun et al. [27] integrated facial action units, eye gaze, and head pose features for stress estimation and demonstrated the benefits of combining multiple behavioral modalities. Similarly, Zhang et al. [31] utilized facial expressions and body motion features for video-based stress detection. Beyond handcrafted features, several researchers have employed machine learning and deep learning techniques to model visual indicators of mental distress. For example, Zhou et al. [32] proposed a deep regression framework that analyzed multiple facial regions for depression recognition. In contrast, Liu et al. [21] developed a multimodal deep neural network that combined facial expressions and body movements captured using commodity cameras. Other studies have leveraged facial action units [11], multimodal behavioral features [30], and spatiotemporal deep learning architectures [8, 33] to assess mental health conditions.

While prior research has demonstrated the utility of visual behavioral cues for assessing mental distress, much of the existing work has focused on stress and depression. In contrast, relatively little work has explored the detection of SAD using video-based behavioral cues. To address this gap, we investigate the feasibility of identifying SAD from low-cost smartphone-recorded videos using head posture, body posture, facial action units, and eye gaze.

### 3 Data Collection

**Participants:** Student participants were recruited after obtaining IRB approval. For each study session, three participants (P1, P2, and P3) who were unfamiliar with one another were invited and seated around a round table (Figure 1). The study procedure was then explained to them. After obtaining consent, video recording was initiated using Android smartphones (Redmi 9A), with one phone positioned to record each participant (C1 for P1, C2 for P2, and C3 for P3). To account for participant no-shows, two additional participants were kept on standby. The study began with a 2-minute baseline period during which participants sat quietly. Following this, the Research Assistant (RA) explained the task, which involved delivering a speech. Public speaking is a well-established component of the Trier Social Stress Test and is widely used to induce social anxiety [10, 14]. Participants were assigned topics such as ‘Artificial Intelligence and Unemployment’ or ‘Agriculture versus Manufacturing Industry.’ They were given 30 seconds to prepare and were then required to speak for at least 2.5 minutes in front of the RA and the other participants while being recorded. After completing the speech, participants self-reported their anxiety levels using a questionnaire. This procedure was carried out sequentially for P1, followed by P2 and P3.

Over a period of four months, we conducted 40 study sessions and collected data from 111 participants. After excluding corrupted video recordings, usable data from 92 participants were retained for analysis. The demographic details of these participants are presented in Table 1.

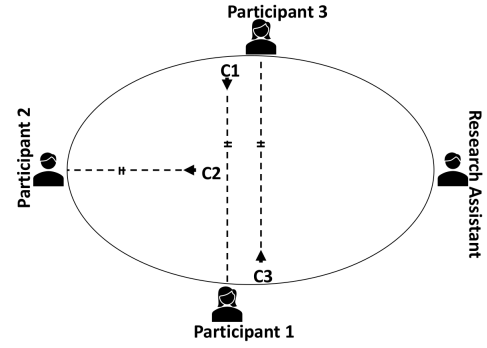


Figure 1: Study setup. C1, C2, and C3 represent cameras 1, 2, and 3, respectively

Table 1: Participants demographic information. SR - Self-reported State anxiety

|              | #  | SR Score<br>( $\mu$ , $\sigma$ ) | Age<br>( $\mu$ , $\sigma$ ) | Gender<br>(Male, Female) |
|--------------|----|----------------------------------|-----------------------------|--------------------------|
| All          | 92 | (14.5, 4.6)                      | (21, 2.7)                   | (61, 31)                 |
| Only SAD     | 45 | (18.4, 2.7)                      | (20.8, 2.6)                 | (29, 16)                 |
| Only Non-SAD | 47 | (10.7, 2.5)                      | (21.2, 2.8)                 | (32, 15)                 |

**Self-reported state anxiety questionnaire:** We developed a custom questionnaire to assess state anxiety during the speech task. The questionnaire consisted of five items targeting different aspects of anxiety during the task: overall anxiety (Q1), perceived performance (Q2), avoidance behavior (Q3), feeling scrutinized (Q4), and fear of negative evaluation (Q5). Participants rated each item on a 5-point Likert scale, with 5 indicating the highest level of distress. The questionnaire items were developed through a literature review and consultations and validated with 18 clinical psychiatrists. After completing the speech activity, participants completed the questionnaire to report their anxiety during the task.

**Data pre-processing:** *Total social anxiety score:* Participants’ responses to the questionnaire were aggregated to obtain a total score ranging from 5 to 25, with higher scores indicating greater anxiety. Based on the total score, participants were categorized into SAD (score  $\geq 15$ ) and non-SAD (score  $< 15$ ) groups. The threshold of 15 corresponds to the midpoint of the possible score range and resulted in a balanced distribution of participants, with 45 in the SAD group and 47 in the non-SAD group.

*Video data:* Although the intended duration of the speech activity was 2.5 minutes, the actual durations varied, with some participants speaking for less time. To ensure consistency in the analysis, we retained participants who spoke for at least 1.5 minutes and excluded those who spoke for less than 1.5 minutes.

### 4 Feature Extraction

From the processed video data, we extracted features related to posture: Head Posture (HP) and Body Posture (BP); facial cues: Facial Action Units (FAU); and eye behavior: Eye Gaze (EG). Analysis of the temporal patterns of these features showed that HP, BP, and

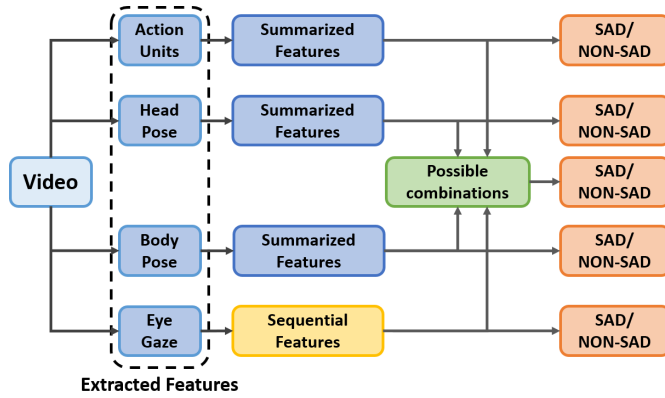


Figure 2: Methodology flowchart

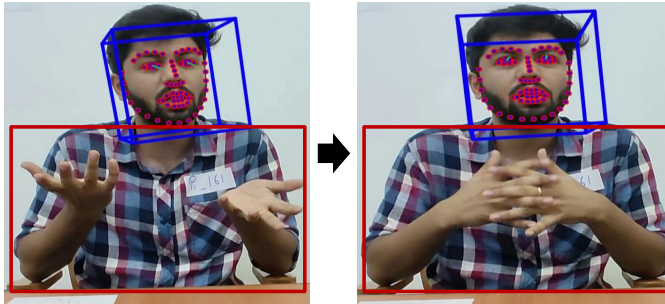


Figure 3: Examples showing the extracted features from continuous video frames: Head pose (Blue colored bounding boxes), Body pose (Red colored bounding boxes), Action units (Pink colored dots around the face), and Eye gaze (green line near eyes) [Best viewed in color].

FAU remained relatively stable over time, whereas EG exhibited noticeable fluctuations. Therefore, we used summarized features for posture and facial cues (Section 4.1) and sequential features for eye gaze (Section 4.2). Figure 3 presents an overview of the methodology.

#### 4.1 Summarized features

We extracted summarized features from head pose, body pose, and facial action units, as their sequential order during the speech activity did not influence their descriptive power. These features were computed at frame level as shown in Figure ??.

**i. Head posture (HP):** HP refers to the position and orientation of the head in three-dimensional space. To compute participants' HP, we used the Py-Feat library [6], which estimates head pose and provides measurements of pitch (up-down movement), yaw (left-right movement), and roll (head tilt). For each of these signals, we computed the corresponding velocity and acceleration. We also derived a combined head-movement signal from pitch, yaw, and roll and computed its velocity and acceleration. From all resulting signals, we extracted statistical features, including the mean, median, mode, minimum, maximum, range, standard deviation,

skewness, and kurtosis. This process resulted in a  $92 \times 72$  feature matrix, corresponding to 92 participants and 72 HP features.

**ii. Body posture (BP):** BP refers to the body's orientation and the spatial arrangement of body parts (e.g., shoulders, arms, and hands). Since participants were seated during the study, we considered only body landmarks above the waist. We used the MediaPipe library [22] to extract the three-dimensional coordinates of body landmarks, including the eyes, nose, mouth, shoulders, elbows, wrists, and fingers. Using these coordinates, we computed motion-related measures, including position, velocity, and acceleration, for each landmark. We then extracted statistical features, including the mean, median, mode, minimum, maximum, range, standard deviation, skewness, and kurtosis, from each landmark signal. This process resulted in a feature matrix of size  $92 \times 181$ , corresponding to 92 participants and 181 BP features.

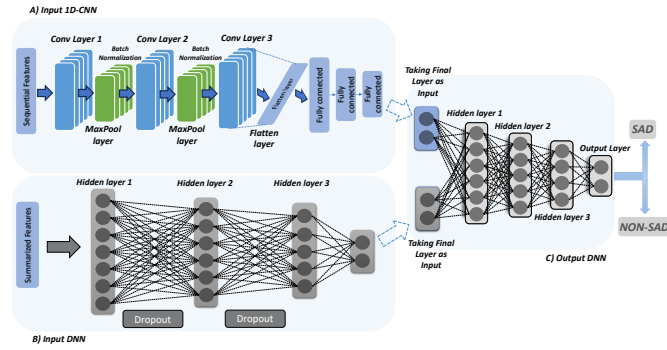
**iii. Action Units (AU):** Facial Action Units (FAUs) represent the fundamental actions of individual facial muscles or muscle groups. We used OpenFace 2.0 [3], a widely used facial behavior analysis toolkit that follows the Facial Action Coding System (FACS) and assigns action units to specific facial muscle movements. For each video frame, OpenFace 2.0 outputs the intensity of 20 action units on a scale from 0 to 5, where 0 indicates no activation, and 5 indicates maximum activation. For each action unit, we computed statistical features, including the mean, median, mode, minimum, maximum, range, standard deviation, skewness, and kurtosis. These features resulted in a 180-dimensional feature vector per participant, yielding a  $92 \times 180$  feature matrix.

#### 4.2 Sequential features

Sequential features preserve the temporal order of the input data, enabling the analysis of behavioral changes over time. We treated eye gaze (EG) as a sequential feature because its temporal variations can provide insights into participants' behavior during the speech task. Eye gaze refers to the orientation of the eyes (see Figure ??) and indicates the direction in which an individual is looking. We used OpenFace 2.0 [3] to extract eye-gaze coordinates and derived features such as pitch, yaw, roll, and the gaze-direction change rate. To capture temporal patterns while preserving sequence information, we computed rolling statistical measures, including the mean, median, standard deviation, minimum, and maximum values.

#### 4.3 Dataset

The final dataset was constructed by combining the HP, BP, AU, and EG feature sets for all 92 participants. This resulted in a high-dimensional feature space relative to the sample size (i.e., participant count), increasing the risk of overfitting. To address this issue, we applied Pearson correlation-based feature reduction to the summarized features and removed redundant features. The feature reduction process was performed separately for the HP, BP, and AU feature sets. Following prior work [23], we used a correlation threshold of 0.75 to identify and discard highly correlated features. After feature reduction, the number of AU, BP, and HP features was reduced to 53, 12, and 25, respectively. We did not apply feature reduction to sequential features because the deep learning eye-gaze model can automatically identify relevant features.



**Figure 4: Hybrid 1D CNN-DNN Architecture: Sequential EG features are processed through “input 1D-CNN”, and summarized features are handled by the “input DNN”. The output from both layers is combined, and inputs to “output DNN” for the final classification.**

## 5 Models

We implemented classical supervised Machine Learning (ML) and supervised Deep Learning (DL) for the classification of the participants (SAD versus non-SAD) using the extracted features. Now, we will be discussing them briefly with respect to the input features.

**i. Models for summarized features:** We used standard machine learning algorithms, including Decision Tree (DT), Gaussian Naive Bayes (GNB), Gradient Boost (GBoost), and Random Forest (RF). Additionally, we developed a custom DNN model illustrated in Figure 4 (see **Input DNN** architecture). It takes summarized features derived from head pose (C2), body pose (C3), and action units (C1) as input. We compared the DNN’s performance, primarily accuracy, against the classical ML models.

**ii. Models for sequential features:** Classical ML models often struggle with time-series data, while DL models can effectively capture temporal patterns and hierarchical representations. Thus, we employed DNN [4], 1D-CNN [18], and LSTM [20] models to classify participants into SAD and non-SAD categories using sequential eye gaze features.

**Parameter settings:** We used eye gaze data from 69 participants for training, each with 1978 time points and 30 features, resulting in an input size of (69, 1978, 30). All models preserved the temporal sequence of the data. (i) **DNN:** Three hidden layers with 64, 32, and 2 neurons. ReLU activation for the first two layers and sigmoid for the output layer. (ii) **1D-CNN:** Three convolutional layers (optimal feature maps: 64 and 32), each followed by ReLU activation, batch normalization, and max-pooling (after first two layers). Flattened output passed to fully connected layers with 32, 16, and 2 neurons. (iii) **LSTM:** Two LSTM layers with 64 and 32 neurons, followed by two fully connected layers with 16 and 2 neurons. All models used Adam optimizer, a batch size of 32, and categorical cross-entropy as the loss function.

**iii. Model for combined summarized and sequential features:** We used an ensemble approach [9] to combine summarized and sequential features and created a 1D-CNN-DNN architecture shown in Figure 4. The model comprises two main sub-modules: input (1D-CNN and DNN) and output (DNN). Once the input 1D-CNN and

DNN models are independently trained, we freeze their weights [19] and extract feature vectors from their final fully connected layers. These vectors, despite their differing shapes, are concatenated into a single feature vector and passed to an output DNN for final classification. Notably, during this phase, backpropagation [15] is limited to the output DNN, ensuring the input models remain unchanged. This final DNN predicts participant labels as SAD or Non-SAD.

**Evaluation:** We used stratified k-fold cross-validation [29] to train and evaluate our models. The ML and DNN models were trained on summarized features, while the 1D-CNN model was trained on sequential features. The best-performing models and their weights were saved and later used by the combined 1D-CNN-DNN model to make predictions leveraging both summarized and sequential features. We used Precision, Recall, F-1 score, and Accuracy to assess the performance of the classification models.

## 6 Results

For brevity, we use abbreviations (i.e., C1–C15) for features and their combinations throughout the remainder of the paper. Configurations evaluated using ML and DNN models included C1 (AU), C2 (HP), C3 (BP), C5 (AU + HP), C6 (AU + BP), C7 (HP + BP), and C8 (AU + HP + BP). Configuration C4 (EG) was evaluated using 1D CNN, LSTM, and DNN models. The EG-based feature combinations, namely C9 (HP + EG), C10 (BP + EG), C11 (AU + EG), C12 (AU + EG + HP), C13 (AU + EG + BP), C14 (EG + BP + HP), and C15 (AU + HP + BP + EG), were evaluated using 1D CNN and DNN models.

**Summarized features:** Figure 5a presents the classification accuracies obtained with GNB, DT, RF, and GBoost, and DNN models on summarized features, C1, C2, C3, C5, C6, C7, and C8. The highest and lowest accuracy achieved by DNN were 74% (on C8) and 52% (on C7), respectively. The proposed DNN outperformed classical ML models on all summarized feature combinations except for C7. This suggests that the DNN was more adept at learning patterns compared to classical ML methods.

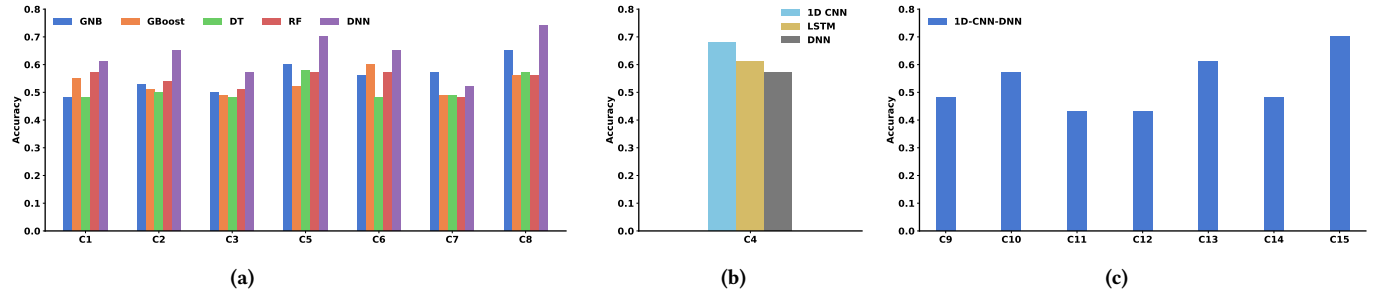
**Sequential features:** Figure 5b illustrates the accuracies achieved with 1D-CNN, LSTM, and DNN exclusively using Eye gaze (C4) features. The 1D-CNN model, yielded the highest accuracy of 68% along with superior precision, recall, and F1-score than the LSTM and DNN as shown in the Table 2b.

**Combined summarized and sequential features:** Figure 5c and Table 2b present the accuracy of the 1D-CNN-DNN model across various combinations (C9, C10, C11, C12, C13, C14, and C15) of summarized and sequential features. In this setup, the 1D-CNN architecture was employed to train the sequential features, while the DNN was utilized to train the summarized features. These features were subsequently fused and fed into another DNN model. The results indicate that the model achieved its highest classification accuracy of 70% on the C15 combination, with a precision and recall of 0.69 and 0.65, respectively.

## 7 Discussion

### 7.1 Classification accuracy

Our findings revealed that the DNN trained on standalone summarized features outperformed classical ML models. The DNN trained on HP (C2) achieved the highest accuracy of 65%, surpassing AU



**Figure 5: Accuracies (%) on (a) Summarized features with Gaussian Naive Bayes (GNB), Decision Tree (DT), Random Forest (RF), and Gradient Boost (GBoost), Deep Neural Network (DNN), (b) Eye gaze features with 1D-CNN, LSTM, and DNN, and (c) Combined summarized and sequential features with 1D-CNN-DNN.**

**Table 2: (a) Performance of best ML classifiers and DNN on *summarized* feature combinations (C1 - C3, C5 - C8); (b) Performance of 1D-CNN and 1D-CNN-DNN on eye-gaze *sequential* feature combinations (C4, C9 - C15).**

| (a) Results for C1 to C8 |      |      |      |      |      |      |      |      |      |      |      |      |      | (b) Results for C4 and C9 to C15 |     |           |        |            |      |      |      |      |      |      |  |
|--------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|----------------------------------|-----|-----------|--------|------------|------|------|------|------|------|------|--|
|                          |      | C1   |      | C2   |      | C3   |      | C5   |      | C6   |      | C7   |      | C8                               |     | Features  | C4     | C9         | C10  | C11  | C12  | C13  | C14  | C15  |  |
|                          |      | RF   | DNN  | GNB  | DNN  | DT   | DNN  | DT   | DNN  | GB   | DNN  | GNB  | DNN  | GNB                              | DNN | Model     | 1D-CNN | 1D-CNN-DNN |      |      |      |      |      |      |  |
| F1-Score                 | 0.58 | 0.60 | 0.56 | 0.63 | 0.51 | 0.56 | 0.59 | 0.67 | 0.55 | 0.63 | 0.56 | 0.51 | 0.63 | 0.71                             |     | F1-score  | 0.77   | 0.45       | 0.54 | 0.36 | 0.43 | 0.60 | 0.42 | 0.65 |  |
| Precision                | 0.58 | 0.60 | 0.61 | 0.63 | 0.52 | 0.61 | 0.60 | 0.68 | 0.56 | 0.63 | 0.58 | 0.51 | 0.67 | 0.73                             |     | Precision | 0.63   | 0.59       | 0.74 | 0.70 | 0.35 | 0.75 | 0.71 | 0.69 |  |
| Recall                   | 0.57 | 0.60 | 0.58 | 0.63 | 0.52 | 0.60 | 0.60 | 0.68 | 0.58 | 0.63 | 0.57 | 0.51 | 0.64 | 0.71                             |     | Recall    | 0.63   | 0.55       | 0.64 | 0.54 | 0.38 | 0.68 | 0.57 | 0.65 |  |
| Accuracy                 | 0.57 | 0.61 | 0.58 | 0.65 | 0.53 | 0.57 | 0.60 | 0.70 | 0.58 | 0.65 | 0.57 | 0.52 | 0.64 | 0.74                             |     | Accuracy  | 0.68   | 0.48       | 0.57 | 0.43 | 0.43 | 0.61 | 0.48 | 0.70 |  |

**Table 3: SAD versus non-SAD classification accuracies and their increase/decrease on transitioning from standalone to the fusion of two, three, and four feature combinations. ↑ and ↓ represent increase and decrease in accuracy, respectively.**

| Standalone | Acc (%) | Fusion (Two)                  | Acc (%)                    | Fusion (Three)                               | Acc (%)                    | Fusion (Four)     | Acc (%) |
|------------|---------|-------------------------------|----------------------------|----------------------------------------------|----------------------------|-------------------|---------|
| AU         | 61      | AU + HP<br>AU + BP<br>AU + EG | 70 (↑)<br>65 (↑)<br>43 (↓) | AU + HP + BP<br>AU + HP + EG<br>AU + BP + EG | 74 (↑)<br>43 (↓)<br>61 (↓) | AU + HP + BP + EG | 70 (↓)  |
| HP         | 65      | HP + BP<br>HP + EG            | 52 (↓)<br>48 (↓)           | HP + BP + EG                                 | 48 (↓)                     |                   |         |
| BP         | 57      | BP + EG                       | 57 (↓)                     |                                              |                            |                   |         |
| EG         | 68      |                               |                            |                                              |                            |                   |         |

(61%) and BP (57%). Furthermore, this suggests that the DNN model was able to discern discriminating patterns in head movements between SAD and non-SAD participants more appropriately than AU and BP. Similarly, the 1D-CNN model trained on eye gaze (C4) achieved the highest accuracy of 68% in discriminating between SAD and non-SAD participants.

The fusion of features (C5 to C15) displayed a varied impact on classification accuracy, with some combinations showing an increase and others a decrease. The highest accuracy of 74% was achieved with C8 (AU + HP + BP) using DNN. Regarding summarized features, we observed increased accuracy with the fusion of two or more summarized features, except for the combination of HP and BP. Table 3 illustrates the change in accuracy, mostly showing an increase in the case of summarized features. For instance, AU, HP, and BP alone yielded accuracies of 61%, 65%, and 57%, respectively. However, when AU and HP were fused, the accuracy increased to 70%. Similar improvements were observed for AU and

BP. Furthermore, when AU, HP, and BP were fused together, the highest accuracy of 74% was achieved. Additionally, we noted that the inclusion of EG led to a drop in accuracy, suggesting that the sequential feature (i.e., EG alone) is effective but not when combined with summarized features. The combined visual features AU, BP, and HP were identified as the most prominent feature fusions for discriminating SAD from non-SAD participants.

## 7.2 Eye gaze

It is noteworthy that prior studies have utilized high-end eye-tracking devices, such as Kim et al. [17] used Arrington Research ViewPoint EyeTracker Systems, and Weeks et al. [28] used D6 high-speed eye tracking systems. In contrast, our approach utilized low-end smartphone camera and OpenFace library to detect the X, Y, and Z coordinates of the left and right eye, from which various EG features were derived, as discussed in section 4.2.

Our findings indicate that EG features alone outperform the standalone AU, BP, and HP features. We achieved an accuracy of 68% with 1D-CNN using eye gaze features alone. However, combining eye gaze with summarized features yielded poor performance, resulting in a drop in accuracy in classifying SAD versus non-SAD participants. In summary, the eye gaze behavior played a crucial role in distinguishing between SAD and non-SAD participants.

## 7.3 Implication and limitations

Our findings offer several key implications. First, eye gaze emerged as the most effective individual feature for distinguishing between SAD and non-SAD participants, consistent with prior research [5, 17, 28]. While other features such as action units, body pose, and head pose also contributed to SAD detection, combining them with

eye gaze surprisingly reduced classification performance, possibly due to feature redundancy or noise.

The primary limitation of our study is controlled laboratory setting, where anxiety was induced through a structured speech task. While this approach ensured consistency across participants, it may not fully capture the complexity or intensity of anxiety experienced in real-world scenarios. Future research can address this limitation by validating our approach in more ecologically valid settings, such as daily-life environments or in-the-wild data collection.

## 8 Conclusion

In this work, we explored the effectiveness of using video-based behavioral cues—head pose, body posture, facial action units, and eye gaze—to detect momentary anxiety. Our results show that specific combinations of these features yield higher predictive performance compared to individual modalities. Furthermore, deep learning models outperformed traditional methods in classifying participants as SAD or non-SAD. These findings highlight the potential of low-cost, non-invasive video-based approaches to improve the scalability and accessibility of social anxiety detection.

## References

- [1] Nasser A Alomari, Sohaila K Bedaiwi, Abdulrahman M Ghasib, Ammar J Kabbarah, Saad A Alnefaie, Nada Hariri, Munirah A Altammar, Abdulaziz M Fadhel, Fai M Altowairqi, Abdulrahman Ghasib, et al. 2022. Social anxiety disorder: Associated conditions and therapeutic approaches. *Cureus* 14, 12 (2022).
- [2] Alaa Althubaiti. 2016. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare* (2016), 211–217.
- [3] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 59–66. doi:10.1109/FG.2018.00019
- [4] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. Deep neural networks and tabular data: A survey. *IEEE transactions on neural networks and learning systems* 35, 6 (2022), 7499–7519.
- [5] Nigel TM Chen and Patrick JF Clarke. 2017. Gaze-based assessments of vigilance and avoidance in social anxiety: A review. *Current psychiatry reports* 19 (2017), 1–9.
- [6] Jin Hyun Cheong, Eshin Jolly, Tiankang Xie, Sophie Byrne, Matthew Kenney, and Luke J Chang. 2023. Py-feat: Python facial expression analysis toolbox. *Affective Science* 4, 4 (2023), 781–796.
- [7] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margarita Tejera Padilla, Feng Zhou, and Fernando De la Torre. 2009. Detecting depression from facial actions and vocal prosody. In *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, 1–7.
- [8] Wheidima Carneiro de Melo, Eric Granger, and Abdenour Hadid. 2019. Combining Global and Local Convolutional 3D Networks for Detecting Depression from Facial Expressions. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. 1–8. doi:10.1109/FG.2019.8756568
- [9] Thomas G. Dietterich. 2000. Ensemble Methods in Machine Learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems (MCS '00)*. Springer-Verlag, Berlin, Heidelberg, 1–15.
- [10] Maria J Garcia-Rubio, Laura Espin, Vanesa Hidalgo, Alicia Salvador, and Jesus Gomez-Amor. 2017. Autonomic markers associated with generalized social phobia symptoms: heart rate variability and salivary alpha-amylase. *Stress* 20, 1 (2017), 61–68.
- [11] Mihai Gavrilescu and Nicolae Vizireanu. 2019. Predicting depression, anxiety, and stress levels from videos using the facial action coding system. *Sensors* 19, 17 (2019), 3693.
- [12] Giorgos Giannakakis, Matthew Padiaditis, Dimitris Manousos, Eleni Kazantzaki, Franco Chiarugi, Panagiotis G Simos, Kostas Marias, and Manolis Tsiknakis. 2017. Stress and anxiety detection using facial cues from videos. *Biomedical Signal Processing and Control* 31 (2017), 89–101.
- [13] Eva Gilboa-Schechtman and Iris Shachar-Lavie. 2013. More than a face: a unified theoretical perspective on nonverbal social cue processing in social anxiety. *Frontiers in human neuroscience* 7 (2013), 904.

- [14] A Harrewijn, MJW Van der Molen, B Verkuil, SW Sweijen, JJ Houwing-Duistermaat, and PM Westenberg. 2018. Heart rate variability as candidate endophenotype of social anxiety: A two-generation family study. *Journal of Affective Disorders* 237 (2018), 47–55.
- [15] Hecht-Nielsen. 1989. Theory of the backpropagation neural network. In *International 1989 Joint Conference on Neural Networks*. 593–605 vol.1. doi:10.1109/IJCNN.1989.118638
- [16] Toshiro Horigome, Brian Sumali, Momoko Kitazawa, Michitaka Yoshimura, Kuoching Liang, Yuki Tazawa, Takanori Fujita, Masaru Mimura, and Taishiro Kishimoto. 2020. Evaluating the severity of depressive symptoms using upper body motion captured by RGB-depth sensors and machine learning in a clinical interview setting: a preliminary study. *Comprehensive psychiatry* 98 (2020), 152169.
- [17] Haena Kim, Jung Eun Shin, Yeon-Ju Hong, Yu-Bin Shin, Young Seok Shin, Kiwan Han, Jae-Jin Kim, and Soo-Hee Choi. 2018. Aversive eye gaze during a speech in virtual environment in patients with social anxiety disorder. *Australian & New Zealand Journal of Psychiatry* 52, 3 (2018), 279–285.
- [18] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 2021. 1D convolutional neural networks and applications: A survey. *Mechanical systems and signal processing* 151 (2021), 107398.
- [19] T.-Y. Kwok and D.-Y. Yeung. 1993. Experimental analysis of input weight freezing in constructive neural networks. In *IEEE International Conference on Neural Networks*. 511–516 vol.1. doi:10.1109/ICNN.1993.298610
- [20] Benjamin Lindemann, Timo Müller, Hannes Vietz, Nasser Jazdi, and Michael Weyrich. 2021. A survey on long short-term memory networks for time series prediction. *Procedia Cirp* 99 (2021), 650–655.
- [21] Dongdong Liu, Bowen Liu, Tao Lin, Guangya Liu, Guoyu Yang, Dezhen Qi, Ye Qiu, Yuer Lu, Qinmei Yuan, Stella C Shuai, et al. 2022. Measuring depression severity based on facial expression and body movement using deep convolutional neural network. *Frontiers in Psychiatry* 13 (2022), 1017064.
- [22] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).
- [23] Haroon Rashid, Sanjana Mendu, Katharine E Daniel, Miranda L Beltzer, Bethany A Teachman, Mehdi Boukhechba, and Laura E Barnes. 2020. Predicting subjective measures of social anxiety from sparsely collected mobile sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–24.
- [24] Robert Rosenman, Vidhura Tennekoon, and Laura G Hill. 2011. Measuring bias in self-reported data. *International Journal of Behavioural and Healthcare Research* 2, 4 (2011), 320–332.
- [25] Nader Salari, Pegah Heidarian, Masoud Hassanabadi, Fateme Babajani, Nasrin Abdoli, Maliheh Aminian, and Masoud Mohammadi. 2024. Global Prevalence of Social Anxiety Disorder in Children, Adolescents and Youth: A Systematic Review and Meta-analysis. *Journal of Prevention* (2024), 1–19.
- [26] Sara Shafique, Iftikhar Ahmed Khan, Sajid Shah, Waqas Jadoon, Rab Nawaz Jadoon, and Mohammed ElAffendi. 2022. Towards automatic detection of social anxiety disorder via gaze interaction. *Applied Sciences* 12, 23 (2022), 12298.
- [27] Zhaodong Sun, Alexander Vedernikov, Virpi-Liisa Kykyri, Mikko Pohjola, Miriam Nokia, and Xiaobai Li. 2022. Estimating Stress in Online Meetings by Remote Physiological Signal and Behavioral Features. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*. 216–220.
- [28] Justin W Weeks, Ashley N Howell, Akanksha Srivastav, and Philippe R Goldin. 2019. “Fear guides the eyes of the beholder”: Assessing gaze avoidance in social anxiety disorder via covert eye tracking of dynamic social stimuli. *Journal of anxiety disorders* 65 (2019), 56–63.
- [29] Chieh-Chen Wu, Wen-Chun Yeh, Wen-Ding Hsu, Md Mohaimenul Islam, Phung Anh Alex Nguyen, Tahmina Nasrin Poly, Yao-Chin Wang, Hsuan-Chia Yang, and Yu-Chuan Jack Li. 2019. Prediction of fatty liver disease using machine learning algorithms. *Computer methods and programs in biomedicine* 170 (2019), 23–29.
- [30] Le Yang, Dongmei Jiang, and Hichem Sahli. 2021. Integrating Deep and Shallow Models for Multi-Modal Depression Analysis—Hybrid Architectures. *IEEE Transactions on Affective Computing* 12, 1 (2021), 239–253. doi:10.1109/TAFFC.2018.2870398
- [31] Huijun Zhang, Ling Feng, Ningyun Li, Zhanyu Jin, and Lei Cao. 2020. Video-based stress detection through deep learning. *Sensors* 20, 19 (2020), 5552.
- [32] Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. 2020. Visually Interpretable Representation Learning for Depression Recognition from Facial Images. *IEEE Transactions on Affective Computing* 11, 3 (2020), 542–552. doi:10.1109/TAFFC.2018.2828819
- [33] Yu Zhu, Yuanyuan Shang, Zhuhong Shao, and Guodong Guo. 2018. Automated Depression Diagnosis Based on Deep Networks to Encode Facial Appearance and Dynamics. *IEEE Transactions on Affective Computing* 9, 4 (2018), 578–584. doi:10.1109/TAFFC.2017.2650899