# Analyzing Participants' Engagement during Online Meetings Using Unsupervised Remote Photoplethysmography with Behavioral Features

Alexander Vedernikov[1], Zhaodong Sun[1], Virpi-Liisa Kykyri[2],
Mikko Pohjola[2], Miriam Nokia[2], and Xiaobai Li[3,1,*]

[1]Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[2]Department of Psychology, University of Jyväskylä, Finland
[3]State Key Laboratory of Blockchain and Data Security, Zhejiang University, China

{aleksandr.vedernikov, zhaodong.sun}@oulu.fi, xiaobai.li@zju.edu.cn
{virpi-liisa.kykyri, mikko.j.pohjola, miriam.nokia}@jyu.fi

## Abstract

*Engagement measurement finds application in healthcare, education, services. The use of physiological and behavioral features is viable, but the impracticality of traditional physiological measurement arises due to the need for contact sensors. We demonstrate the feasibility of unsupervised remote photoplethysmography (rPPG) as an alternative for contact sensors in deriving heart rate variability (HRV) features, then fusing these with behavioral features to measure engagement in online group meetings. Firstly, a unique Engagement Dataset of online interactions among social workers is collected with granular engagement labels, offering insight into virtual meeting dynamics. Secondly, a pre-trained rPPG model is customized to reconstruct rPPG signals from video meetings in an unsupervised manner, enabling the calculation of HRV features. Thirdly, the feasibility of estimating engagement from HRV features using short observation windows, with a notable enhancement when using longer observation windows of two to four minutes, is demonstrated. Fourthly, the effectiveness of behavioral cues is evaluated when fused with physiological data, which further enhances engagement estimation performance. An accuracy of 94% is achieved when only HRV features are used, eliminating the need for contact sensors or ground truth signals; use of behavioral cues raises the accuracy to 96%. Facial analysis offers precise engagement measurement, beneficial for future applications.*

## 1. Introduction

The rapid transition to online communication underlines the importance of engagement analysis in virtual meetings.

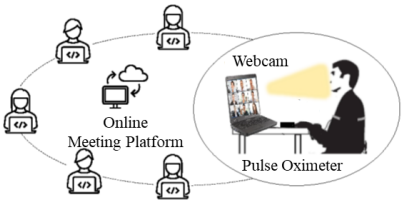



Figure 1. A large *Engagement Dataset* of realistic online meetings. Facial videos and cPPG were recorded.

With the absence of physical cues and direct interaction between individuals, assessing engagement becomes more challenging. Nevertheless, estimating engagement during these interactions offers key insights into participant behavior, group dynamics, and individual input, helping meeting organizers promote effective collaboration.

Engagement analysis in virtual meetings often relies on facial and body language recognition [11, 23], although they can't gauge direct physiological responses like heart rate variability (HRV), which is difficult for individuals to fake [45]. Electrocardiography (ECG) would be ideal but is limited by its need for direct contact. In contrast, remote photoplethysmography (rPPG), a computer vision-based technique, assesses cardiac activity through facial color changes. This makes it suitable for online engagement estimation as it obviates the need for direct contact and specialized equipment [37]. Using unsupervised deep learning for rPPG signal extraction eliminates the need for ground truth signals, labeled datasets, and expert annotations, enhancing flexibility in engagement estimation.

This article pioneers the application of unsupervised rPPG measurement technology in estimating engagement during online meetings. It also proposes an enhancement in engagement estimation by integrating behavioral cues such

*Corresponding author.

as facial expression and motion. This study discusses potential impediments affecting the performance of rPPG in engagement analysis. A significant influence on engagement estimation performance arises from the size of the observation window employed for calculating HRV features from the obtained rPPG signal. Moreover, a *Engagement Dataset* of online group meetings (captured in real-world scenarios) among social workers is collected (Fig. 1), the first to explore engagement levels in group interactions, especially with many participants. It contains 1.5-hour videos for in-depth engagement analysis, unveiling evolving interactions and gradual shifts in group dynamics, and incorporates heart rate (HR) data from contact PPG (cPPG), a feature absent in public datasets. By using a continuous engagement range of -10 to +10, the *Engagement Dataset* captures nuances that tend to be neglected in other datasets [19, 24, 27]. The analysis reveals that participant engagement is linked to work effectiveness and mitigates job stress, underlining the study's practical value.

## 2. Related work

Automated engagement analysis research, initiated by Whitehill *et al*. [47] in 2014, showcased machine learning's capacity to estimate engagement with human-like precision. Subsequent studies explored a wide range of applications [26], including education [39], social media [16], news [20], human-robot [38] and human-human interaction [10], consumer engagement [16], healthcare [43], games [12], and film viewing analysis [8]. These studies explored behavioral features like facial expressions [18], eye gaze [13], and body gestures [25]. There were studies that utilized physiological features, *e.g*., [41] used features extracted from electroencephalogram (EEG) signals. In addition, studies also analyzed modalities reliant solely on text [5], reaction time [35], and response accuracy [35]. Rather than using a single modality, scientists have investigated multi-modality fusion methods merging facial expression-related features with speech/audio [22], head and body pose/gestures/motions [36], physiological signals (such as electroencephalographic - EEG activity [3], thermal signals [17], and electrodermal activity - EDA [15]), game events [36], mouse behavior [50], and contextual information [4]. Past research on engagement estimation is robust, but fails to leverage the improvements rPPG data can provide. Notably, rPPG technology has been applied in affective computing, showing promise in assessing emotional states such as depression [9], stress [37], and embarrassment [48], indicating unexploited potential for engagement analysis.

Over the past decade, only one study in 2016 by Monkaresi *et al*. [34] explored HR for engagement estimation, but it faced major constraints. Firstly, they relied on contact ECG sensors for HR, impractical in real use. Our approach, however, uses a non-intrusive method, eliminating contact sensors. Secondly, the video methods of that time [21] confined their research to lab data, suffering in real HR detection scenarios. Conversely, our method is tested and effective in real-life conditions. Moreover, using only seven basic HR statistical features, they failed to fully exploit HR signals' potential, leading to a high clinical error rate, as the authors acknowledged. Their experiments showed that HR signals were less effective than facial expressions in estimating engagement, likely due to unreliable cardiac information. HRV features, strongly linked to mental states [30], offer potential as efficient engagement indicators demanding more research, driving this work's innovative approach to physiological signals.

As engagement estimation methods advanced, multiple datasets were created. Real-world e-learning engagement was first examined by the *DAiSEE* dataset [19] in 2016. The horizons of student engagement analysis in educational games were broadened by the *Multimodal Affective State Recognition Dataset* [36]. Using the *MHHRI* dataset [10], engagement in dyadic human and triadic human-human-robot contexts was explored. In 2017, *UE-HRI* dataset [7] further expanded the scope of human-robot interaction and centered around interactions with the robot Pepper. In the domain of e-learning, the *EngageWild* dataset [24] and *VRESEE* datasets [40] delved deep into students' engagement patterns. YouTube gaming videos and facial engagement modalities were uniquely integrated in the *FaceEngage* dataset [12]. A predictive direction in the field was signified by the *PAFE* dataset [27] in 2022 and the EngageNet dataset [42] in 2023, both utilizing different contexts. Primarily, all mentioned datasets offer only visual modality for engagement analysis, with just two exceptions: the *MHHRI* dataset [10] includes data from audio, video, depth, electrodermal activity (EDA), temperature, and 3-axis wrist acceleration, while the *UE-HRI* dataset [7] delivers information from a microphone array, cameras, depth sensors, sonars, lasers, and user feedback captured through the robot's touchscreen. The lack of cPPG data in public resources for rPPG engagement methods underlines the necessity for new datasets and unsupervised approaches.

This paper introduces the application of unsupervised deep learning for calculating rPPG signals, an unexplored modality in previous automated engagement analyses, emphasizing its practical, non-contact, and non-intrusive nature. Furthermore, various behavioral feature sets are also evaluated and fused with physiological features to further boost the performance. The work is established on a novel self-collected *Engagement Dataset*, capturing real-world online video meetings of social workers with consultant therapists for the purpose of reducing work-related stress. The *Engagement Dataset* construction, the proposed method framework, and experimental results are explained in the following sections.

## 3. Dataset construction

### 3.1. Data collection

The *Engagement Dataset*, which contains facial videos, cPPG data, and engagement annotations, was collected (Tab. 1). The *Engagement Dataset* comprises recordings from group online video meetings. Serving as a reflection of participants' everyday work, each online video meeting was organized using the Zoom platform and involved the participation of seven to nine individuals with one consultant. These participants were social services employees working with individuals facing mental health challenges and substance abuse issues. Prior to each session, participants were given detailed instructions on how to set up their recording environment. To stream and record participants' facial videos during online meetings, webcams were installed, and OBS Studio was used. Throughout the sessions, a research assistant provided online guidance using Zoom Breakout Rooms. Participants were encouraged to freely express themselves and move, as long as their faces remained visible in the videos. The cPPG signals were captured using the Beurer 80 pulse oximeter. Following each online video meeting, the facial videos and cPPG signals were synchronized based on their timestamps.

### 3.2. Data annotation

Psychology students worked as research assistants to perform engagement annotation based on the observed behavior of the target subject from face video. DARMA, a continuous measurement system, was used for engagement annotation. It synchronizes media playback and continuous recording of observational measurement conducted with a computer joystick at a sampling rate of 20Hz. The DARMA software provided a continuous coding time series consisting of two values per second [28], with 10 ("High Engagement") and -10 ("Low Engagement"). Research assistants underwent training to familiarize themselves with distinct engagement levels before starting the coding. Inter-rater reliability in control sessions was good (correlation coefficient over 0.8).

The scale of -10 to +10 facilitated a detailed understanding of engagement, allowing for robust and nuanced annotations. This range was chosen over a binary or three-point scale because human emotional states and levels of engagement are continuous and nuanced. The chosen range captures these nuances by providing a more granular scale. It is broad enough to encompass the extreme ends of the engagement spectrum while being fine-grained enough to account for slight variations within these extremes. Additionally, this continuous scale can better account for individual differences in engagement levels, allowing for the representation of each subject's unique engagement signature.

| | |
|---|---|
| **Engagement labels** | [-10, 10], 2 labels/second |
| **Engagement classes** | 3 (Low / Middle / High) |
| **Participants (male/female)** | 25 (3/22) |
| **Consultants (male/female)** | 2 (1/1) |
| **Video recordings** | 109 |
| **cPPG data** | 106 |
| **Average duration** | 1 hr 23 min 58 sec |
| **Total length** | 153 hours |

Table 1. *Engagement Dataset* statistics and properties.

### 3.3. Data statistics and properties

The *Engagement Dataset* comprises 24 recorded group online video meetings, each lasting approximately 1.5 hours. The *Engagement Dataset* contains two modalities, namely facial video recordings and cPPG signals. Each online video meeting involved between seven to nine participants and one consultant. The resolution of 1920 × 1080 and the frame rate of 60 fps were recommended to optimize video quality, although variations were observed due to different cameras and recording environments. The *Engagement Dataset* contains several unique properties as opposed to previous datasets studying engagement. 1) **Duration of data.** With 1.5-hour videos compared to shorter public clips, the *Engagement Dataset* enables in-depth engagement analysis, uncovering evolving interactions, gradual shifts in interactions, and group dynamics. 2) **Group dynamics.** Unlike earlier datasets, the *Engagement Dataset* involves more participants per session, amplifying group interaction complexities for better engagement analysis and enhancing group dynamic insights. 3) **Recorded cPPG signals.** The presence of cPPG signals in the *Engagement Dataset*, not presented in other datasets, underscores its unique importance in advancing engagement estimation. 4) **Granular annotation.** The *Engagement Dataset* uniquely adopts a -10 to +10 range of engagement, emphasizing the fluidity of human responses. Such a spectrum effectively captures engagement nuances often overlooked in other datasets. 5) **Real-world setting.** Rooted in real-world scenarios, the *Engagement Dataset* showcases authentic online meetings of social service employees. Unlike other datasets in simulated and well-controlled environments, the *Engagement Dataset* promises true engagement data pertinent to real-world situations.

## 4. Method

### 4.1. rPPG curves reconstruction from facial videos

The proposed method is illustrated in Fig. 2. The original videos are pre-processed to extract frames and acquire facial landmarks using the OpenFace library [6]. This tool effectively addresses issues related to head motion, providing precise tracking of facial landmarks. The detected facial landmarks are employed to determine regions of inter-

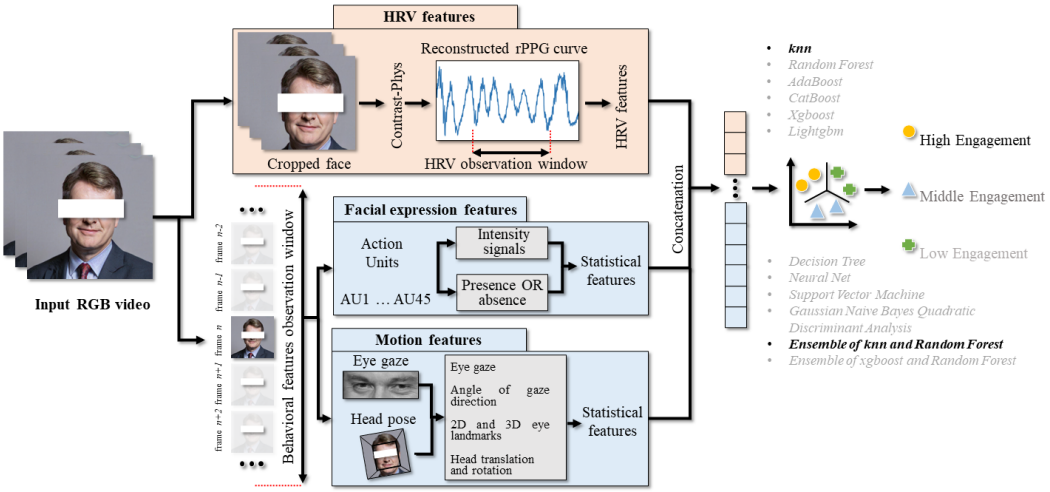



Figure 2. The diagram depicts engagement estimation based on heart rate variability (HRV) and behavioral features. HRV features are calculated from reconstructed rPPG signals, and behavioral features are computed using the OpenFace library [6]. Feature-level multimodality fusion is employed, followed by a classifier.

est (ROI) on exposed skin areas. The faces are cropped and then resized to dimensions of 128 × 128, preparing them for input into the unsupervised Contrast-Phys model [44], a computer vision-based technique, for the subsequent rPPG curves reconstruction. The Contrast-Phys approach is based on the principle where the utilization of a 3DCNN model allows the derivation of multiple rPPG signals from distinct spatiotemporal locations within each video. Two videos, randomly selected from the *Engagement Dataset*, constitute the input of Contrast-Phys. One video yields spatiotemporal rPPG (ST-rPPG) block $P$, rPPG samples $[p_1, \ldots, p_N]$, and associated power spectrum densities (PSDs) $[f_1, \ldots, f_N]$. The other video provides ST-rPPG block $P'$, rPPG samples $[p'_1, \ldots, p'_N]$, and corresponding PSDs $[f'_1, \ldots, f'_N]$, following the same procedure. The contrastive loss pulls together PSDs from the same video while pushing apart PSDs from distinct ones. Contrast-Phys implies that using rPPG spatiotemporal similarity, the PSDs from the same ST-rPPG block should resemble each other as follows $\text{PSD}\{P(t_1 \rightarrow t_1 + \Delta t, h_1, w_1)\} \approx \text{PSD}\{P(t_2 \rightarrow t_2 + \Delta t, h_2, w_2)\} \implies f_i \approx f_j, i \neq j$ and $\text{PSD}\{P'(t_1 \rightarrow t_1 + \Delta t, h_1, w_1)\} \approx \text{PSD}\{P'(t_2 \rightarrow t_2 + \Delta t, h_2, w_2)\} \implies f'_i \approx f'_j, i \neq j$. Subsequently, to bring together PSDs (positive pairs) from the same video, the mean squared error is suggested to be utilized as the loss function. When normalized based on the total count of positive pairs, the positive loss term is:

$$L_p = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left( \| f_i - f_j \|^2 + \| f'_i - f'_j \|^2 \right) / \left(2N(N-1)\right) \quad (1)$$

On the other hand, the cross-video rPPG dissimilarity suggests that the PSDs resulting from spatiotemporal sampling of two separate ST-rPPG blocks will be distinct. This attribute for the two input videos is described as $\text{PSD}\{P(t_1 \rightarrow t_1 + \Delta t, h_1, w_1)\} \neq \text{PSD}\{P'(t_2 \rightarrow t_2 + \Delta t, h_2, w_2)\} \implies f_i \neq f'_j$. Next, the task of distancing PSDs (negative pairs) from two different videos is achieved when the negative mean squared error is used as the loss function. Then, the overall quantity of negative pairs is employed to normalize the negative loss term:

$$L_n = -\sum_{i=1}^{N} \sum_{j=1}^{N} \| f_i - f'_j \|^2 / N^2 \quad (2)$$

Finally, the overall loss function combines both positive and negative loss terms: $L = L_p + L_n$.

The model is pre-trained on the Oulu Bio-Face database [29]. This approach enables the reconstruction of rPPG signals in any recorded facial video, eliminating the need for ground truth data in the future.

### 4.2. Definition of observation window for HRV and behavioral features

The -10 to +10 engagement scale is divided into three classes - Low, Medium, and High engagement - through a process designed to simplify the complexity of engagement analysis while retaining a significant level of detail. This triadic classification approach provides a balance between the -10 to +10 scale's granularity and a binary scale's simplicity, serving as a practical and efficient method for multi-classifying engagement levels. In High engagement (Score: 10), the participant is either speaking, attempting to take a turn, speaking over someone else, or, as a listener, actively showing engagement through minimal vocalizations (such as 'mmm'), nods, and/or facial expres-

sions. In Medium engagement (Score: 0), the participant either gazes at the speaker and appears to be listening, or gazes away while still seeming attentive. Finally, in the Low engagement (Score: -10), the participant gazes away or has their eyes closed, appearing not to follow the conversation, possibly engaging in side activities like turning away or opening emails. The process of converting continuous engagement labels into three classes and definition of observation windows for HRV and behavioral features (BF) is detailed below.

A 5-seconds time period was chosen for detecting intervals of stable engagement. Reliability risks arise from short, context-lacking clips. Assessing longer clips is more complex due to mixed engagement levels. Therefore, video recordings were partitioned into 5-seconds intervals, and the standard deviations of engagement were computed for each interval. The median value of calculated standard deviations across the collected *Engagement Dataset* was subsequently estimated. This median served as a threshold for filtering out 5-seconds intervals with higher engagement standard deviations. This process excluded unstable intervals with significant engagement fluctuations. The engagement values obtained from the filtered intervals were classified as follows: -10 to 0 as 'Low engagement', 0 to 5 as 'Medium engagement', and 5 to 10 as 'High engagement'. Example of engagement data of a specific participant in a particular video meeting after filtering process is shown in Fig. 6 (see supplementary material). After the filtering process, the distribution of the samples was as follows: 4001 samples were categorized as 'Low Engagement', 23069 as 'Medium Engagement', and 8320 as 'High Engagement'. Next, the midpoint of each 5-seconds filtered interval becomes the center of an HRV and behavioral observation window (Fig. 3). Subsequently, HRV features are computed with a 60-seconds observation window that requires a 30-seconds step in both directions, while BF is derived using a 2-seconds observation window following similar logic.

### 4.3. Feature extraction

**HRV features.** Reconstructed rPPG signals were processed using the Neurokit2 library [32] to identify systolic peaks. This allowed for inter-beat intervals (IBIs) calculation, yielding HRV features. These included three Poincaré plot features, 16 time-domain features, and five frequency-domain features, making a total of 24 HRV features. The impact of HRV observation window size on engagement estimation performance was investigated by extracting HRV features from the reconstructed rPPG signals using various observation window sizes (60, 90, 120, 150, 180, 210, and 240 seconds).

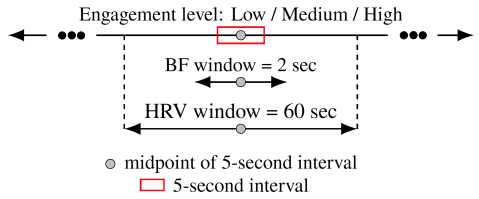



Figure 3. Observation windows concept for extracting BF and HRV features in engagement analysis.

**Facial expression features.** Action Units (AUs) were used for the computation of facial expression features [46]. 17 unique AUs were detected and tracked using the OpenFace library [6] (see supplementary material Tab. 6). AUs were represented in two ways: as intensity signals (on a scale from 0 to 5) and as binary classifications for their presence or absence. Hence, 34 features were obtained, providing a comprehensive insight into engagement-related facial movements.

**Motion features.** Four distinct sets of motion features were extracted using the OpenFace library [6]. These included: six gaze tracking features illustrating the 3D coordinates of each eye's gaze direction; two features representing each eye's gaze direction; 280 features outlining 2D and 3D landmarks around each eye; and six features pertaining to the translation and rotation of the head in 3D space. Collectively, these 294 features provided a comprehensive examination of facial and eye movements associated with engagement.

In the following, 'behavioral features' (BF) denotes motion and facial expressions. These BF were computed for each frame extracted from the collected video recordings. To consider the temporal dynamics of these features, a 2-seconds observation window was adopted, wherein the average value of each feature was computed.

### 4.4. Engagement classification

Three-class classification of engagement has been conducted in two stages. In the first stage, classification was done using only 24 HRV features with various classifiers. These included knn, Random Forest, AdaBoost, CatBoost, xgboost, lightgbm, Support Vector Machine (with 'poly', 'rbf', and 'sigmoid' kernels), Decision Tree, Gaussian Naive Bayes, Quadratic Discriminant Analysis, Neural Net, an ensemble of knn and Random Forest, and an ensemble of xgboost and Random Forest. Out of these, the knn and the ensemble of knn and Random Forest (knn+RF) proved to be the most effective. In the second stage, these two most effective classifiers were employed. A feature-level fusion technique was used to combine 24 HRV and 328 BF features into a single feature vector for a specific 5-seconds engagement interval, which was then fed into the classifier.

## 5. Results

### 5.1. Experimental protocol

The *Engagement Dataset* was split into 80% training and 20% test sets. To ensure model robustness and generalization, a 5-fold subject dependent cross-validation was applied to the training data. The trained model predicts outcomes for the testing set. This protocol reliably measures model performance while maintaining test set independence [14]. Accuracy and the Receiver Operating Characteristic Area Under the Curve (ROC AUC) metrics were initially used to evaluate models. For the best performing models, the F1 score and confusion matrices were additionally calculated to enhance the comprehensiveness of the assessment.

### 5.2. HR measurement accuracy

For the *Engagement Dataset*, the HR from the reconstructed rPPG signals was compared to that of ground truth cPPG using mean absolute error (MAE) and root mean square error (RMSE) as evaluation metrics. An MAE of 5.15 bpm and an RMSE of 7.81 bpm were obtained, indicating promising results given the uncontrolled conditions of the video recordings. The SOTA performance [37] was marked by MAE values of 3.55, 5.99, and 9.26 bpm in controlled lab settings.

### 5.3. Short observation window HRV features for engagement estimation

Utilizing a short observation window for HRV features calculation, the proposed approach's performance on the *DAiSEE* [19] dataset (focused on students' engagement) and *Engagement Dataset* is shown in Tab. 2. For the *DAiSEE* dataset with 10-seconds snippets, the highest accuracy of 54.49% was achieved by the Random Forest model. Meanwhile, a 49.40% accuracy was yielded by the knn model on *Engagement Dataset* using a 10-seconds observation window. The performance of the proposed method is constrained by the 10-seconds snippets of the *DAiSEE* dataset. Reliable HRV features are not offered by such a short observation window, a fact also reflected in *Engagement Dataset* performance using a 10-seconds observation window. Longer videos are required for robust HRV and engagement estimation. This study's video recordings allow the proposed method's capability to be fully realized using long HRV observation windows. In the following subsection, the influence of HRV observation window size is analyzed in detail.

### 5.4. Effects of HRV observation window size on the performance of engagement estimation

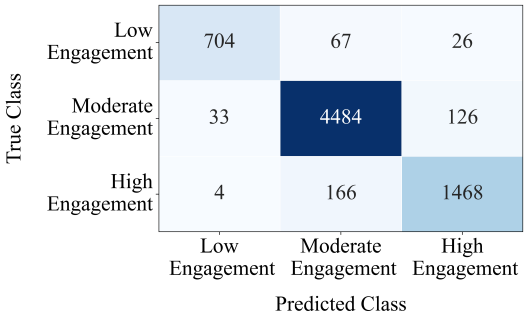
With the constraints of a 10-seconds observation window addressed in the preceding section, the effects of HRV observation windows of extended lengths are analyzed. The experiments revealed that knn and knn and Random Forest ensemble (knn+RF) were the most effective machine learning classifiers. Models' performance was assessed across different HRV observation window values (Tab. 3). As the HRV observation window size increased from 60 seconds to 240 seconds, significant improvements were observed in the accuracy and ROC AUC scores for both models. At an HRV observation window size of 60 seconds, the knn model achieved an accuracy of 0.816 and a ROC AUC score of 0.870, while the knn+RF model showed slightly better results with an accuracy of 0.816 and a ROC AUC score of 0.906. At the optimal HRV observation window size of 240 seconds, accuracy of the knn model reached 0.937 and its ROC AUC score peaked at 0.960, while the knn+RF model outperformed with an accuracy of 0.940 and a ROC AUC of 0.983. The HRV observation window size has a significant

| Method | Accuracy [%] |
|---|---|
| *DAiSEE* [19] | |
| InceptionNet Video Level [19] | 46.40 |
| InceptionNet Frame Level [19] | 47.10 |
| C3D Training [19] | 48.60 |
| Inflated 3D ConvNet [49] | 52.35 |
| ResNet + TCN (weighted loss) [2] | 53.70 |
| **HRV + Random Forest** | **54.49** |
| C3D FineTune [19] | 56.10 |
| C3D LRCN [19] | 57.90 |
| DFSTN [31] | 58.84 |
| C3D + TCN [2] | 59.97 |
| ResNet + LSTM [2] | 61.15 |
| ResNet + TCN [2] | 63.90 |
| EfficientNet B7 + TCN [40] | 64.67 |
| EfficientNet B7 + Bi-LSTM [40] | 66.39 |
| Ordinal TCN [1] | 67.40 |
| EfficientNet B7 + LSTM [40] | 67.48 |
| *Engagement Dataset* | |
| **HRV + knn** | **49.40** |

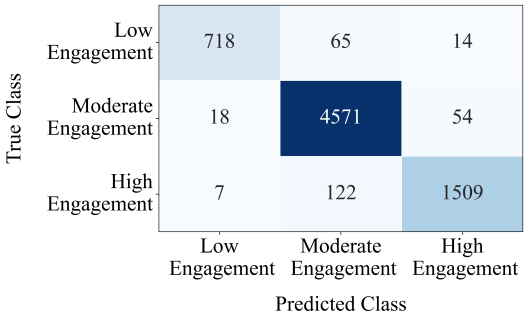
Table 2. Proposed method's performance using short observation window HRV features for engagement estimation based on two datasets: (1) *DAiSEE* [19]; (2) *Engagement Dataset*.

| HRV window [sec] | Accuracy [-] knn | Accuracy [-] knn+RF | ROC AUC [-] knn | ROC AUC [-] knn+RF |
|---|---|---|---|---|
| 60 | 0.816 | 0.816 | 0.870 | 0.906 |
| 90 | 0.882 | 0.880 | 0.923 | 0.947 |
| 120 | 0.910 | 0.911 | 0.946 | 0.967 |
| 150 | 0.924 | 0.926 | 0.953 | 0.975 |
| 180 | 0.931 | 0.936 | 0.955 | 0.977 |
| 210 | 0.937 | 0.938 | 0.961 | 0.983 |
| 240 | 0.937 | **0.940** | 0.960 | **0.983** |

Table 3. Evaluation metrics of the knn model and ensemble of knn and Random Forest model for engagement estimation based on HRV features calculated at different values of HRV observation window.




(a) Utilization of HRV features.




(b) Utilization of HRV and Behavioral features.

Figure 4. Ensemble of knn and Random Forest model's confusion matrices for engagement estimation (HRV observation window of 240 seconds) based on (a) HRV features; (b) HRV and Behavioral features.

| HRV window [sec] | HRV | HRV + BF (1) | HRV + BF (2) | HRV + BF (3) | HRV + BF (4) | HRV + BF (5) | HRV + BF (1+2) | HRV + BF (4+5) | HRV + BF (all) |
|---|---|---|---|---|---|---|---|---|---|
| 60 | 0.816 | 0.864 | 0.855 | 0.861 | 0.896 | 0.919 | 0.856 | 0.928 | 0.930 |
| 90 | 0.880 | 0.903 | 0.902 | 0.870 | 0.936 | 0.931 | 0.891 | 0.942 | 0.929 |
| 120 | 0.911 | 0.929 | 0.929 | 0.881 | 0.952 | 0.944 | 0.919 | 0.952 | 0.934 |
| 150 | 0.926 | 0.934 | 0.936 | 0.884 | 0.953 | 0.944 | 0.927 | 0.954 | 0.940 |
| 180 | 0.936 | 0.937 | 0.944 | 0.892 | 0.954 | 0.947 | 0.932 | 0.955 | 0.938 |
| 210 | 0.938 | 0.939 | 0.944 | 0.896 | 0.955 | 0.951 | 0.933 | 0.958 | 0.941 |
| 240 | 0.940 | 0.943 | 0.949 | 0.899 | 0.956 | 0.956 | 0.934 | **0.960** | 0.940 |

Table 4. Ensemble of knn and Random Forest model's accuracy for engagement estimation based on HRV and various sets of BF at different values of HRV observation window. (1) stands for gaze tracking, (2) for angle of gaze direction, (3) for 2D and 3D landmarks of specific points around each eye, (4) for translation and rotation of the head in 3D space, and (5) for facial Action Units.

impact on the performance of engagement estimation models and larger observation window sizes yield better results. With longer observation windows, finer and more complex HRV patterns in the rPPG signals can be discerned, thus enabling more accurate and robust predictions based on the extracted HRV features. Moreover, noise and short-term variability in the signals, which may confuse or degrade the performance of the classifiers, are better averaged out over longer observation periods. Table 3 demonstrates that a 120-seconds HRV observation window is already sufficient for the proposed method to showcase its capability. The confusion matrix of the model with the highest performance is shown in Fig. 4a. A high differentiation is observed among the three classes, slightly better at detecting Moderate engagement than Low and High engagement. The F1 scores for Low, Moderate, and High levels of engagement are 0.92, 0.96, and 0.90, respectively. To improve performance in Low and High engagement, the model requires more balanced data or better features for class distinction. All in all, the *DAiSEE* dataset's 10-seconds clips prevent an evaluation of the effects of HRV observation window length on engagement estimation. Yet, from the *Engagement Dataset*, it's evident that HRV features can robustly measure engagement when given a sufficient observation window size, such as 2 to 4 minutes.

### 5.5. Effectiveness of BF selection when fusing with HRV for engagement estimation

While HRV features have established a strong foundation, relying solely on physiological signals limits prediction accuracy. Adding BF is crucial for improved metrics as human engagement involves not only physiological but also subtle behavioral cues. Thus, a model fusing both physiological and behavioral aspects could enhance prediction.

In Tab. 4, the ensemble of knn and Random Forest model's accuracy using both HRV and various BF combinations, including (1) gaze tracking, (2) angle of gaze direction, (3) 2D and 3D landmarks of specific points around each eye, (4) translation and rotation of the head in 3D space, and (5) facial Action Units, is shown. The ROC AUC metric, mirroring the accuracy trend, was omitted. The model's performance is significantly influenced by the BF selection. Set (3) is found to be less efficient with longer HRV observation windows; however, sets (4) and (5) are shown to boost prediction accuracy. The highest performance was observed when HRV was paired with (4), indicating "translation and rotation of the head in 3D space", and (5), pointing to "facial Action Units". Significant metric enhancement is observed when BFs are incorporated as the model's accuracy increases from 0.816 to 0.930 for a 60-seconds observation window. Enhanced performance

due to BF inclusion is consistently observed across all data. Yet, the boost is more pronounced at smaller HRV observation window sizes. In Fig. 4b, the confusion matrix of the knn+RF model, incorporating HRV and both (4) and (5) BF, is presented, as opposed to the HRV-only feature confusion matrix depicted in Fig. 4a. It outperforms the singular HRV model, underscoring that the inclusion of BF yields superior engagement estimation. After fusing with BF, the F1 scores for Low, Moderate, and High levels of engagement improved to 0.93, 0.97, and 0.94, respectively. Further refinement through enhanced feature selection or the exploration of alternate models might offer improved differentiation between Low and High Engagement.

### 5.6. Combining HRV observation window size and BF set for engagement estimation

The effect of combining BF (4+5) set with different HRV observation window sizes on the model's performance is shown in Fig. 5. As the HRV observation window size increases, the performance metrics of the model tends to improve, up to a certain point. However, the performance improvement isn't strictly linear and plateaus after a certain HRV observation window size. In the knn+RF model with BF (4+5), accuracy improves from 0.928 (60 seconds) to 0.960 (240 seconds), but the difference between 210 seconds and 240 seconds is marginal. Thus, while the integration of BF improves the model, gains are notably larger for smaller HRV observation windows. Nevertheless, with the addition of BF, the best HRV-only model's performance was enhanced by 2%. Table 5 demonstrates the comparison of the proposed method with the method proposed by Mohamad *et al*. [33] which is publicly available and aimed for engagement estimation. The *Engagement Dataset* was processed using their method, following the same protocol and data split as described in Sec. 5.1.

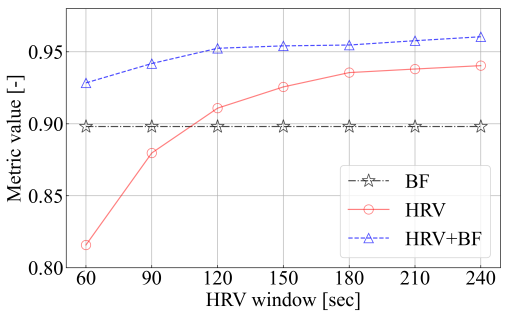



Figure 5. Accuracy of the ensemble of knn and Random Forest model for engagement estimation based on HRV and both (4) and (5) BF calculated at different values of HRV observation window. The accuracy of the model, based solely on BF, remains constant due to its independence from the value of the HRV window.

| Method | Accuracy [-] | ROC AUC [-] |
|---|---|---|
| Mohamad *et al*. [33] | 0.633 | 0.500 |
| HRV | 0.940 | 0.983 |
| HRV + BF (4+5) | **0.960** | **0.991** |

Table 5. Comparison of methods' performances for engagement estimation based on *Engagement Dataset*. Ensemble of knn and Random Forest model was used for "HRV" and "HRV + BF (4+5)" methods.

## 6. Conclusion

The remote measurement of heart rate variability, along with facial and body language recognition, is used for engagement analysis in virtual meetings, which have become increasingly popular over the recent years. This article pioneers the application of unsupervised rPPG measurement technology in estimating engagement during online meetings. It first introduces the *Engagement Dataset* centered on social workers' online video meetings. The *Engagement Dataset* is accompanied by granular engagement labels that capture the essence of virtual meeting dynamics. Subsequently, the effect of HRV observation window size on engagement estimation performance was assessed using both collected and public datasets. Short HRV observation windows proved to be unreliable for HRV feature calculation, while longer observation windows (e.g., 2-4 minutes) provided a robust foundation for engagement estimation. Further, the significance of selecting the right behavioral features set was evaluated. Given the crucial importance of the right HRV window choice, a performance increase from 49.40% through 81.60% to 94% was witnessed when the HRV observation window size was adjusted from 10 seconds to 60 seconds and then to 240 seconds. Moreover, when the correct BF set was used, performance was further boosted by up to 2%.

In future, the proposed method is to be validated using more datasets. With the formulated method, engagement analysis at the clip level was carried out, setting the stage for future entire video recording level analysis. Additionally, exploration of engagement fluctuation on a group level, e.g., the analysis of synchrony or interactions of the participants, is planned to better analyze the online event. Moreover, this method is intended to be advanced into a deep learning model, trained end-to-end for engagement estimation.

## Acknowledgments

This work was supported by the Research Council of Finland (former Academy of Finland) Academy Professor project EmotionAI (grants 336116, 345122), and the Finnish Work Environment Fund (Project 200414). The authors also acknowledge CSC-IT Center for Science, Finland, for providing computational resources.

## References


[1] Ali Abedi and Shehroz Khan. Affect-driven ordinal engagement measurement from video. *arXiv preprint arXiv:2106.10882*, 2021. 6

[2] Ali Abedi and Shehroz S. Khan. Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. In *Proceedings - 2021 18th Conference on Robots and Vision, CRV 2021*, page 151 – 157, 2021. Cited by: 15; All Open Access, Green Open Access. 6

[3] Ioannis Arapakis, Miguel Barreda-Angeles, and Alexandre Pereda-Baños. Interest as a proxy of engagement in news reading: Spectral and entropy analyses of eeg activity patterns. *IEEE Transactions on Affective Computing*, 10(1): 100–114, 2017. 2

[4] Sinem Aslan, Nese Alyuz, Cagri Tanriover, Sinem E Mete, Eda Okur, Sidney K D'Mello, and Asli Arslan Esme. Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019. 2

[5] Thushari Atapattu, Menasha Thilakaratne, Rebecca Vivian, and Katrina Falkner. Detecting cognitive engagement using word embeddings within an online teacher professional development community. *Computers & Education*, 140: 103594, 2019. 2

[6] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018. 3, 4, 5

[7] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. Ue-hri: a new dataset for the study of user engagement in spontaneous human-robot interactions. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 464–472, 2017. 2

[8] Sergio Benini, Mattia Savardi, Katalin Balint, Andras Balint Kovacs, and Alberto Signoroni. On the influence of shot scale on film mood and narrative engagement in film viewers. *IEEE Transactions on Affective Computing*, 13(2):592–603, 2019. 2

[9] Constantino Álvarez Casado, Manuel Lage Cañellas, and Miguel Bordallo López. Depression recognition using remote photoplethysmography from facial videos. *IEEE Transactions on Affective Computing*, 14(4):3305–3316, 2023. 2

[10] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, 10(4):484–497, 2017. 2

[11] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision*, 131(6):1346 – 1366, 2023. Cited by: 13; All Open Access, Green Open Access, Hybrid Gold Open Access. 1

[12] Xu Chen, Li Niu, Ashok Veeraraghavan, and Ashutosh Sabharwal. Faceengage: robust estimation of gameplay engagement from user-contributed (youtube) videos. *IEEE Transactions on Affective Computing*, 13(2):651–665, 2019. 2

[13] Youjin Choi, JooYeong Kim, and Jin-Hyuk Hong. Immersion measurement in watching videos using eye-tracking data. *IEEE Transactions on Affective Computing*, 13(4): 1759–1770, 2022. 2

[14] David Cournapeau, Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, and Vincent Michel. scikit-learn: Machine learning in python. https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation, 2023. Accessed: 2023-08-29. 6

[15] Ilana Dubovi. Cognitive and emotional engagement while learning with vr: The perspective of multimodal methodology. *Computers & Education*, 183:104495, 2022. 2

[16] Seyed Pouyan Eslami, Maryam Ghasemaghaei, and Khaled Hassanein. Understanding consumer engagement in social media: The role of product lifecycle. *Decision Support Systems*, 162:113707, 2022. 2

[17] Chiara Filippini, Edoardo Spadolini, Daniela Cardone, Domenico Bianchi, Maurizio Preziuso, Christian Sciarretta, Valentina del Cimmuto, Davide Lisciani, and Arcangelo Merla. Facilitating the child–robot interaction by endowing the robot with the capability of understanding the child engagement: The case of mio amico robot. *International Journal of Social Robotics*, 13:677–689, 2021. 2

[18] Goren Gordon, Samuel Spaulding, Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. Affective personalization of a social robot tutor for children's second language skills. In *Proceedings of the AAAI conference on artificial intelligence*, 2016. 2

[19] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*, 2016. 2, 6

[20] Misbah Ul Hoque, Kisung Lee, Jessica L Beyer, Sara R Curran, Katie S Gonser, Nina SN Lam, Volodymyr V Mihunov, and Kejin Wang. Analyzing tweeting patterns and public engagement on twitter during the recognition period of the covid-19 pandemic: A study of two us states. *IEEE Access*, 10:72879–72894, 2022. 2

[21] YungChien Hsu, Yen-Liang Lin, and Winston Hsu. Learning-based heart rate detection from remote photoplethysmography features. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4433–4437, 2014. 2

[22] Yuyun Huang, Emer Gilmartin, and Nick Campbell. Conversational engagement recognition using auditory and visual cues. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, page 590 – 594, 2016. Cited by: 11. 2

[23] Aditya Kamath, Aradhya Biswas, and Vineeth Balasubramanian. A crowdsourced approach to student engagement recognition in e-learning environments. In *2016 IEEE Win-*

*ter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016. 1

[24] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2018. 2

[25] Shoroog Khenkar and Salma Kammoun Jarraya. Engagement detection based on analyzing micro body gestures using 3d cnn. *Computers, Materials & Continua*, 70(2), 2022. 2

[26] Puneet Kumar, Alexander Vedernikov, and Xiaobai Li. Measuring non-typical emotions for mental health: A survey of computational approaches. *arXiv preprint arXiv:5456274*, 2024. 2

[27] Taeckyung Lee, Dain Kim, Sooyoung Park, Dongwhi Kim, and Sung-Ju Lee. Predicting mind-wandering with facial videos in online lectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2104–2113, 2022. 2

[28] Iolanda Leite, Marissa McCoy, Daniel Ullman, Nicole Salomons, and Brian Scassellati. Comparing models of disengagement in individual and group interactions. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, page 99–105, New York, NY, USA, 2015. Association for Computing Machinery. 3

[29] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, page 242 – 249, 2018. Cited by: 47; All Open Access, Green Open Access. 4

[30] Wen Chieh Liang, John Yuan, Deh Chuan Sun, and Ming Han Lin. Changes in physiological parameters induced by indoor simulated driving: Effect of lower body exercise at mid-term break. *Sensors*, 9(9):6913–6933, 2009. 2

[31] Jiacheng Liao, Yan Liang, and Jiahui Pan. Deep facial spatiotemporal network for engagement prediction in online learning. *Applied Intelligence*, 51:6609–6621, 2021. 6

[32] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689 – 1696, 2021. Cited by: 205; All Open Access, Bronze Open Access, Green Open Access. 5

[33] Omid Mohamad Nezami, Mark Dras, Len Hamey, Deborah Richards, Stephen Wan, and Cécile Paris. Automatic recognition of student engagement using deep learning and facial expression. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 273–289. Springer, 2020. 8

[34] Hamed Monkaresi, Nigel Bosch, Rafael A. Calvo, and Sidney K. D'Mello. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, 8(1):15–28, 2017. 2

[35] Teresa M Ober, Corinne J Brenner, Alvaro Olsen, Bruce D Homer, and Jan L Plass. Detecting patterns of engagement in a digital cognitive skills training game. *Computers & Education*, 165:104144, 2021. 2

[36] Athanasios Psaltis, Konstantinos C Apostolakis, Kosmas Dimitropoulos, and Petros Daras. Multimodal student engagement recognition in prosocial games. *IEEE Transactions on Games*, 10(3):292–303, 2017. 2

[37] Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappé, and Fan Yang. Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 14(1):622–636, 2023. 1, 2, 6

[38] Hanan Salam, Oya Celiktutan, Hatice Gunes, and Mohamed Chetouani. Automatic context-driven inference of engagement in hmi: A survey. *arXiv preprint arXiv:2209.15370*, 2022. 2

[39] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13 (4):2132–2143, 2022. 2

[40] Tasneem Selim, Islam Elkabani, and Mohamed A Abdou. Students engagement level detection in online e-learning using hybrid efficientnetb7 together with tcn, lstm, and bi-lstm. *IEEE Access*, 10:99573–99583, 2022. 2, 6

[41] Rabi Shaw, Chinmay Mohanty, Bidyut Kr Patra, and Animesh Pradhan. 1d multi-point local ternary pattern: A novel feature extraction method for analyzing cognitive engagement of students in flipped learning pedagogy. *Cognitive Computation*, pages 1–14, 2022. 2

[42] Monisha Singh, Ximi Hoque, Donghuo Zeng, Yanan Wang, Kazushi Ikeda, and Abhinav Dhall. Do i have your attention: A large scale engagement prediction dataset and baselines, 2023. 2

[43] Lars Steinert, Felix Putze, Dennis Küster, and Tanja Schultz. Audio-visual recognition of emotional engagement of people with dementia. In *Interspeech*, pages 1024–1028, 2021. 2

[44] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13672 LNCS:492 – 510, 2022. Cited by: 2; All Open Access, Green Open Access. 4

[45] Zhaodong Sun, Alexander Vedernikov, Virpi-Liisa Kykyri, Mikko Pohjola, Miriam Nokia, and Xiaobai Li. 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*, page 216–220, New York, NY, USA, 2023. Association for Computing Machinery. 1

[46] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2): 97–115, 2001. 5

[47] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 5(1): 86 – 98, 2014. Cited by: 379. 2

[48] Yi-Chiao Wu, Li-Wen Chiu, Chun-Chih Lai, Bing-Fei Wu, and Sunny S. J. Lin. Recognizing, fast and slow: Complex emotion recognition with facial expression detection and remote physiological measurement. *IEEE Transactions on Affective Computing*, 14(4):3177–3190, 2023. 2

[49] Hao Zhang, Xiaofan Xiao, Tao Huang, Sanya Liu, Yu Xia, and Jia Li. An novel end-to-end network for automatic student engagement recognition. In *ICEIEC 2019 - Proceedings of 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication*, page 342 – 345, 2019. Cited by: 15. 6

[50] Zhaoli Zhang, Zhenhua Li, Hai Liu, Taihe Cao, and Sannyuya Liu. Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology. *Journal of Educational Computing Research*, 58(1):63–86, 2020. 2