Multimodal Context-Based Continuous Authentication

Saandeep Aathreya, Meghna Chaudhary, Tempestt Neal, Shaun Canavan
University of South Florida, Florida, United States
{saandeepaath,meghna1,tjneal,scanavan}@usf.edu

Abstract

We present a new multimodal, context-based dataset for continuous authentication. The dataset contains 27 subjects, with an age range of $[8, 72]$*, where data has been collected across multiple sessions while the subjects are watching videos meant to elicit an emotional response. Collected data includes accelerometer data, heart rate, electrodermal activity, skin temperature, and face videos. We also propose a baseline approach for fair comparisons when using the proposed dataset. The approach uses a combination of a pretrained backbone network with supervised contrastive loss for face. Time-series features are also extracted, from the physiological signals, which are used for classification. This approach, on the proposed dataset, results in an average accuracy, precision, and recall of 76.59%, 88.90, and 53.25, respectively, on electrical signals, and 90.39%, 98.77, and 75.71, respectively on face videos.*

1. Introduction

Owing to advancement of technology, biometric systems have become more applicable in preserving information on high security applications such as healthcare, banking, and e-commerce. Due to the inefficacy of current systems [49], the need for an increased level of security from continuous authentication (CA) systems is essential. Current solutions for continuous authentication systems largely depend on the user's behavioral information, such as keystroke and mouse dynamics [53, 2, 13], gaze patterns [18, 56], and face recognition [17]. These solutions, however, have certain limitations. More specifically, they require the users to engage with the system by performing a predefined task in an uninterrupted manner [14]. For example, keystroke dynamics-based CA requires the user to continue typing on the keyboard to provide appropriate data to the system. Gaze patterns require certain amount of synergy between the user and system to produce adequate results [14].

Due to the vulnerability of the prior systems, we propose to use sensor-based signals such as accelerometer, photoplethysmogram (PPG), and electrodermal activity (EDA) sensors that can potentially overcome the above mentioned issues. More specifically, we use 4 different signals, namely- *Accelerometer (ACC)*, *Heart Rate (HR)*, *EDA*, and *Temperature (TEMP)*. Firstly, the aforementioned biosignals do not prompt the user for a specific input since they collect data in an unobtrusive manner. In addition to this, they are appropriate for systems which do not impose any restrictions on movement of the user. The PPG signals were traditionally intended for healthcare practices because of the signals' innate property of being distinctive towards human liveliness [1, 48, 15]. This makes them an unlikely target of spoof attacks and forging since this would require falsifying physiological signals involuntarily from the subject [39].

While PPG signals contain valuable information, and can be difficult to spoof, they are subjective and often change over time. Because of this, it is important to evaluate CA systems across multiple time periods. The existing literature largely investigates the effect of physiological signals for CA over a single session (i.e., continuous signal measurement during the same time period) [28, 7, 36]. Practically, the enrollment and authentication phase of a CA systems occurs across different sessions. Furthermore, emotions play a critical role in actuating a physiological response from the user (and vice-versa) [48, 54, 19]. For example, an increase in stress is often indicated by spikes in heart rate and EDA signals [34]. Considering this, we collected a new dataset from multiple sessions and emotion-based contexts. The dataset contains various physiological signals, as well as face video information. The participants watch several videos that are meant to elicit different emotional responses such as *Sadness*, *Content*, *Disgust*, and *Happiness*. This is done to vary the physiological activity within the participants [35]. This occurs for a total of three sessions with an interval ranging from a few days to weeks between each session. As an added value, we augment our physiological dataset with facial and body pose videos of the participants reacting to these videos. Face information is a valuable modality [13], and we show that it can be used to assist the physiological signals. The contributions of this work are 3-fold and can be summarized as follows:

1. A new dataset is presented that contains physiologi-

cal signals such as heart rate, EDA, skin temperature, motion activity with accelerometer data (collectively called electrical signals), and face videos from two different angles. The data is collected across multiple sessions, and contexts within each session. Each context comprises of participant watching a specific video to elicit an emotional response. A total of 27 subjects from diverse background and age-range (young, adults, and older adults) participated in the study. The dataset is available to the community and can be obtained upon request.

2. A subject-wise baseline continuous authentication approach on both electrical signals and face images is proposed. For face images, we propose a combination of a pretrained backbone network with a supervised contrastive loss [32]. For electrical signals, we extract a collection of features on time-series data and use glassbox models [44] for classification. We utilize this model to observe which feature-sets contribute the most across different experiments

3. The proposed solution across different sessions and context and report different metrics on each, including accuracy, precision, recall, FRR, and FAR.

2. Related Work

2.1. Continuous Authentication

Crouse et. al. [13] combined face-based features with Inertial Measurement Unit (IMU) data to improve the face recognition accuracy and demonstrate the effectiveness of the new system for unobtrusive and continuous authentication. To do this, the authors utilize the smartphone devices' accelerometer, gyroscope, and magnetometer (collectively called IMU) data to correct camera sensor orientation and face image. They combined face data with the IMU data for continuous authentication. The frequency of CA happens every t_{delay} seconds wherein a threshold score, t_{login}, incrementally reduces every time there is a verification failure. Gopal et. al. [22] proposed a low interval, robust CA system using only 3-axis accelerometer data. This involved extracting 52 features from the raw accelerometer data by dividing them into overlapping chunks and creating features per chunk. This was followed by feature ranking and selection using correlation-based algorithms [21]. The authors then construct baseline and temporal models using Random Forest [6] and Neural Networks.Martinho et. al. [38] proposed a multi-biometric system using Electrocardiogram (ECG) and Blood Volume Pulse (BVP). The participants performed guided writing, touch pad usage, and free-form writing on a laptop. The ECG data was collected using chest and forearm sensors and BVP data through wrist sensors. Feature denoising and segmentation was performed on both modalities and features were extracted from a fixed window waveform. Decision level fusion between KNN [3] and Naive-Bayes[52] were performed on both uni and multi-modal data.

2.2. PPG Dataset for Continuous Authentication

The public dataset Biosec3 [27] includes PPG data using a fingertip device from 170 participants over multiple sessions. In each session, the participant underwent a period of relaxation for 3 minutes followed by an exercise for a duration of 1.5 minutes. The PPG-ACC dataset [5] includes 7 subjects with an age range of $[20, 52]$ including four males and seven females. 15 PPG signals for each subject along with their accelerometer reading were collected. The data was collected only for a single session wherein the participants performed two exercises - squats and stepper followed by a resting period. Schmidt et. al. [46] introduced an emotion based PPG dataset that utilized chest and wrist worn sensors for data collection. More specifically, ECG, EDA, Temperature and Electromyography (EMG) data were recorded for 15 participants with 12 males and three females. During the session, the participants were subjected to four different conditions - Baseline, wherein the participant sat for 20 minutes without performing any task, Amusement, where the subjects watched 11 funny video clips, Stress, where the subjects delivered a five minute speech on their personal traits and finally, a guided meditation session. Koelstra et. al. [33] presented a multimodal dataset called DEAP for the analysis of human affect states. 32 participants watched a 1-minute long excerpts of music videos. Each participant watched 40 such videos and self-reported the affect states in terms of arousal and valence. The dataset included EEG signals of all the participants and video information of 22 of 32 participants.

The proposed dataset extends these works in multiple ways. First, similar to the DEAP dataset [33], the proposed dataset uses similar tasks to elicit emotions. Unlike DEAP, we extend the proposed dataset to multiple sessions. Second, the Biosec3 dataset [27] used multiple sessions, however, only one context was used. In the proposed dataset, the subject data is collected across multiple emotion-based stimuli (i.e., context). Third, the DEAP dataset provides an arousal and valence value either positive $(+1)$, negative (-1) or neutral (0) value. This is extended as the proposed dataset also includes the compound emotional state of the participant. Finally, the proposed dataset contains a larger age range, including children and elderly subjects, compared to many current datasets. For example, the age range of PPG-ACC is $[20, 52]$, whereas the proposed dataset has an age range of $[8, 72]$.

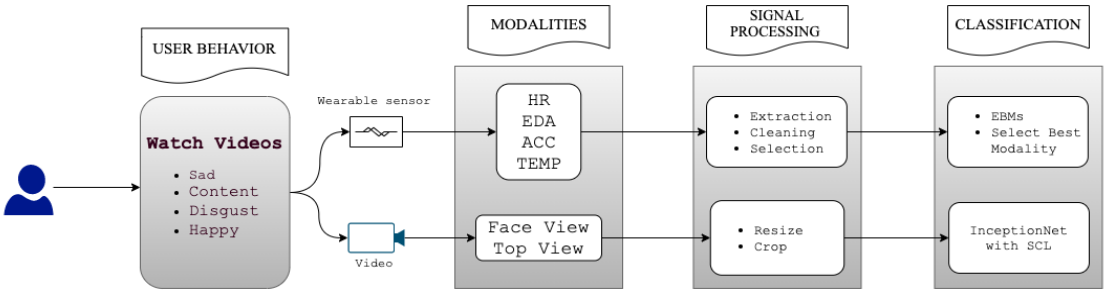


Figure 1. Overview of the data collection pipeline. Per session, each participant watches four context videos. Camcorders and wearable sensors collect the respective modalities such as videos and electrical signals (Section 3.2). Then signal preprocessing for each modality is performed (Section 4.1). Finally both the modalities are trained independently using separate classifier (Section 4.2).

Table 1. Demographic information of dataset subjects.

Type	Category	Number of cases	% of cases
Sex	Male	17	63
	Female	10	37
Race	Asian	11	41
	White	9	33
	Middle Eastern/North African	4	15
	African American	2	7
	Hispanic/Latino	1	4
Age	8 – 18	4	15
	19 – 35	18	67
	36 – 72	5	18

3. Context-based CA Dataset

3.1. Data Collection

Figure 1 shows the entire pipeline of the data collection process through classification. More specifically, each participant performs a task (e.g., watch a video), which is the *user behavior* module. Next, wearable sensors and a camcorder record the respective data across various *modalities*. The raw signal then undergoes preprocessing, feature extraction and data selection process under the *signal processing* module. Finally, we use the extracted features within a continuous authentication (CA) system by using all modalities in the *classification* module. Altogether, 32 subjects participated in the data collection process. In this study, the exclusion criteria for our dataset include participants who have not completed all three sessions. Five subjects did not complete all three sessions and accordingly, we evaluate our dataset and CA systems on 27 participants (Table 1).

A major goal of the study is to investigate continuous authentication where subjects are eliciting different affective states across multiple sessions. This is motivated by previous works that have shown expression can impact identifying faces [29]. To facilitate this, we select four short clips per session, each to bring forth a certain target emotion from the participant, namely *Sad*, *Content*, *Disgust*, and *Happy*. It is important to note that the intended/target affective state and the self reported state can be different [55]. Prior to the study, the participants completed an initial virtual meeting, and read and signed a consent form. The participants then filled out a demographic form which contained information as summarized in Table 1. For children below 18 years of age, the demographic information were filled by the accompanied guardian. The participant is then equipped with a wearable sensor and given instructions on its use. Then, the participant begins watching the context videos. The average length of each session is $\simeq$ 30 minutes. The study was approved by XXXX (removed for double-blind review).

Baseline Condition. Once the participant is ready, they begin the session by performing a 30 second breathing exercise. This is a baseline setup for the participant with the aim of inducing a neutral affective state [41] and nullify any affect priming [42] originating out of external factors before beginning the session.

Affect Elicitation. We focus on collecting spontaneous facial expressions and reactions from the participants. For recording this spontaneous affect behaviour each task was regulated and guided by a research assistant in order for the participant to familiarize themselves with the environment. Moreover, movie clips and videos have shown effectiveness in evoking an emotional response from the participants [23, 11]. To this end, video clips eliciting a specific affect is played for the participants to watch.

Table 2 describes the content of each video per session played for both children and adults, making a total of 12 videos that a participant watches throughout the data collection process across three sessions. We play the same set of context videos for adults and older adults per session, but maintain a separate set of videos more applicable to children. The average length of the video clips is $\simeq$ 60 seconds. Altogether, the participant watches four videos during the session, each eliciting a target emotion in the following order - *Sadness*, *Content*, *Disgust*, and *Happiness*. As outlined in Figure 2, each video is succeeded by the breathing

Table 2. Stimulus videos played for the participant

Target Emotion	Age-Group	Session 1	Session 2	Session 3
Sad	Adult	911 crash	911 call by a toddler	93 year man in court
	Children	Lion King	Toy Story Ending	Scene from COCO
Content	Adult	Puppies playing	Man with his cat	Rescued puppies
	Children	Puppies playing	Kid with puppies	Scene from Despicable Me
Disgust/Stressed	Adult	Eating Caterpillar	Iguana chased by snakes	Fear factor lying with worms
	Children	Eating Caterpillar	Scene from Pirates	Timon and Pumba eat worms
Happy	Adult	Prank video	Man falling	YouTube fails
	Children	Prank video	scene from UP	YouTube fails

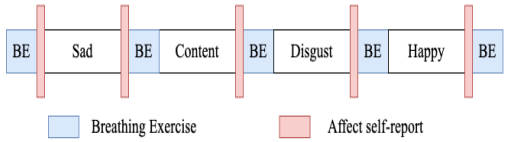
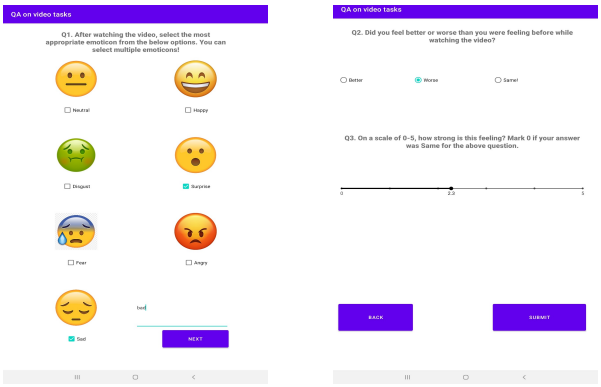
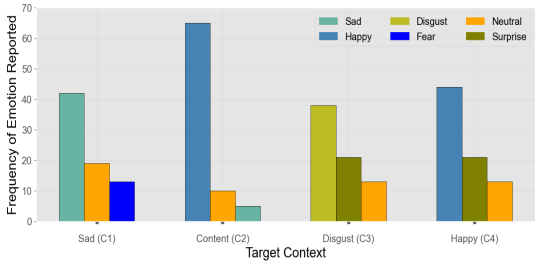


Figure 2. Order of videos shown along with subject self-reporting and breathing exercises (BE).

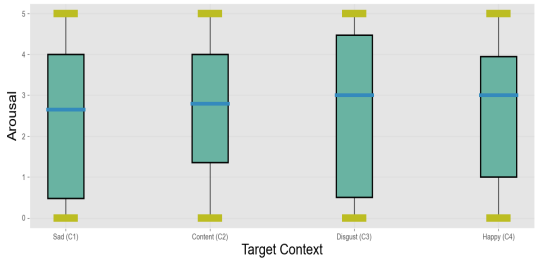


(a) (b)

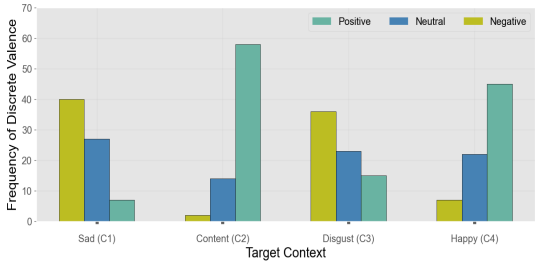
Figure 3. User interface of our Android self-report app where the user logs affect response after watching the stimulus video



(a) Top three emotions reported per task by the all participants



(b) Box plot of distribution of arousal levels reported.



(c) Discrete valance distribution for positive (+1), negative (-1), and neutral (0) levels.

Figure 4. Distributions of all three reported affect labels. Each plot considers all the participants across all three sessions.

exercise (Section 3.1) with a similar purpose of removing affect priming originating from the preceding video.

Affect Self-Report. To extend the usability of the dataset, we additionally provide the quantified affect response of the participant after watching the video ($Q1$). Here, the user is prompted to enter different information such as emotion felt while watching the video, the overall effect of the video on the user (positive, negative, or neutral), and the scale to which the user felt the positive or negative effect, if any. To do this, we built a simple Android application that presents a three-part questionnaire ($Q1, Q2, Q3$) to the participant once they finish watching a video. The layout of the app is shown in Figure 3.

After each affect stimulus, the participants self-report their emotions, which are listed as *Neutral*, *Happy*, *Disgust*, *Surprise*, *Fear*, *Angry*, and *Sad* (Figure 3a). In addition to this, we allow a custom text field where the user can enter any emotion that is not covered in $Q1$. As established in the literature [16], the participants can experience multiple emotions under a single context video. Based on this, we allow the participants to select multiple emotion categories. The layout of questionnaires $Q2$ and $Q3$ is shown in Figure 3b. The focus of $Q2$ is to collect the discrete subjective experience of the participant after watching the video. Therefore, we discretize the continuous valence levels into *Positive* $(+1)$, *Negative* (-1) and *Neutral* (0).

Figure 4a compares the highest reported emotions (top three) and the target emotion for each context video for all the subjects across all the sessions. It can be observed that the two mostly align with each other. For example, after viewing the video specifically designed to evoke sadness (C1), participants self reported feeling *Sad* a collective of 42 times, across all three sessions, and similarly, participants self reported feeling *Happy* 65 times while watching the content video (C2). This suggests that the context videos effectively succeeded in eliciting the targeted emotion. $Q3$ focuses on extracting the arousal values for each report. This measures the intensity with which the participant felt the reported emotion. Figure 4b shows the box plot for the continuous values reported for all subjects and sessions per context. The Interquartile Range of the plots indicate a rela-

(a) Frontal face view. (b) Top angle face View.

Figure 5. Different views from the collected video.

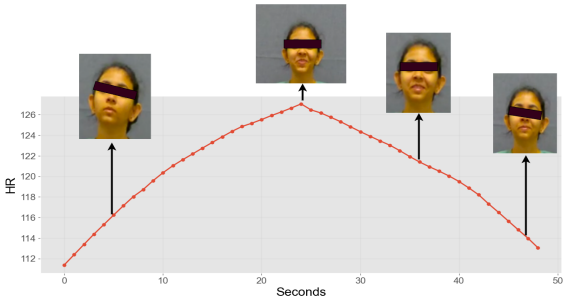


Figure 6. Sample subject showing facial images and heart rate over time. As can be seen here, the subject has a more intense expression when the HR is at it's peak. Conversely, the expression is lower intensity when the HR is lower (e.g., at 0 and 50 seconds).

tively high variance between the arousal values reported by the participants, with the least variance shown for the Content context (C2). Additionally, the left-skewed boxes indicate a higher frequency towards low-valued scores. Figure 4c shows the frequency of each discrete valence category per context video for all the subjects and sessions. Once again, this is in accordance to Figure 4a, where we show the agreement between target emotion and self-reported emotion. For example, the *Disgust* context ($C3$) shows 36 participants felt worse than before after watching the video. Similarly, for Happy ($C4$), 45 participants felt better than before post context session.

3.2. Sensor and Camera Setup

The acquisition system consists of two cameras for capturing face and body movement and an Empatica E4 [1] wearable sensor to capture the electrical signals, namely accelerometer data (ACC) *(32 Hz)*, skin temperature (TEMP) *(4 Hz)*, heart rate (HR) *(1 Hz)*, and electrodermal activity (EDA) *(4 Hz)*. The subjects wore the watch on their non-dominant hand. For video capture, we used two DVC 4K camcorders, for two different angles - *face* view and *top* view, which records at 60fps. Figure 5 shows both the camera views obtained to capture face and body movement. We chose two angles, as it has shown that pan camera angles can affect the overall result in affect related tasks [12]. For the electrical signals, we use the signals stored from the watch directly. As part of the CA system design and the dataset, we only use the segment of video and signals, where the participant is watching one of the four context videos. Another motivation for using physiological signals along with face is it has been shown that there is a correlation between intense expressions and corresponding physiological signals [25]. Our data collection supports this as shown in Figure 6.

4. Authentication System Design

The raw data captured must be first segmented into the respective contexts. This means the electrical signals and the videos must contain only those fragments that corresponds to one of the four context videos. To do this, we automate the process using the audio of the session and the audio of the four context videos for that session and apply signal correlation using Fast Fourier Transform [30] which demonstrates higher efficacy in acquiring the exact timestamps of the context video. Using this information, we segment both the video and electrical signals.

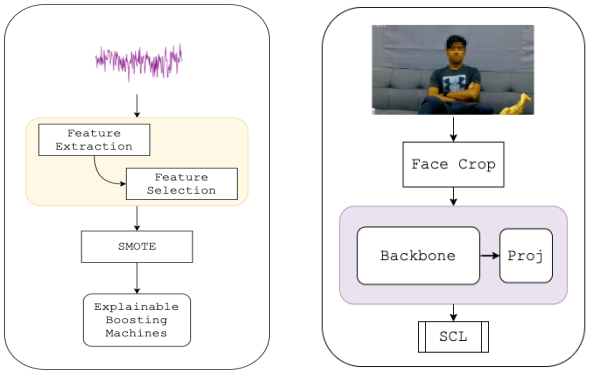
4.1. Feature Extraction and Selection

Electrical Signals. Feature extraction of the electrical signals was done using a sliding window technique which has been employed previously in literature [45, 22]. Primarily, we start with ACC data which consists of tri-axial coordinate information and calculate the magnitude per time step, which is given by $m = \sqrt{x^2 + y^2 + z^2}$, where x, y, and z are the coordinates. Similar to the works of Gopal et. al. [22], we discard the first and the last 100 samples as part of the data cleaning process. Next, we generate samples using a window size of 500 samples without any overlapping criteria (i.e., $[500, 0]$). We follow the same process for all the other signals with different window size and overlapping criteria. For EDA, TEMP and HR, the values are $[100, 75]$, $[100, 75]$, and $[20, 15]$, respectively. This indicates that the verification frequency on these signals can have a time interval as low as $\simeq 1.5$ seconds (EDA).

Since our CA system provides authentication decisions at the feature level, we extract several time and frequency domain features from each of the electrical signals. We use *tsfresh* [9], a tool which specializes in feature extraction from time series data and provides an extensive list of features from each column of the data[2]. This results in a feature set of $\simeq 900$ dimensions per signal. Subsequently, feature selection is performed using *tsfresh* which applies a significance test on uni-variate features and evaluates the corresponding p values using a Benjamini Hochberg test [10].

[1] http://www.empatica.com/research/e4/

[2] https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html



(a) Electrical Signals (b) Images

Figure 7. Training/enrollment pipeline for the two modalities. (a) Discrete valance distribution for positive (+1), negative (-1), and neutral (0) levels. (b) Face detection and cropping occurs in real-time from MediaPipe [37]. Then training occurs via combination of deep networks and Supervised Contrastive Loss (SCL).

Images. We first randomly select frames from each video to remove redundant information from consecutive frames. Face cropping is then performed for each video and the background is masked out to remove any artifacts. We employ an open source tool called MediaPipe [37], which is a light-weight face detection and cropping tool. This enables us to provide authentication decisions at frame level.

4.2. Classification

Typically, a CA system is composed of regular one-shot authentication, which determines the legitimacy of the current user at every shot. This is inline with a one-vs-all classification problem where the model treats the genuine user as *one* class, and all the other imposter users under the *all* class. To this end, we create a one-vs-all classifier (binary classifiers) for each subject using both electrical signals and images independently. The following subsection details the respective algorithms.

Electrical Signals. Given the training samples for each electrical signal, which are of varying dimensions, we observe a high class imbalance in the dataset. To facilitate enhanced training, we choose to balance the classes using the Synthetic Minority Oversampling Technique (SMOTE) [20], where synthetic samples for the minority class can be generated. It works according to the principle of nearest neighbor where it interpolates new data between the target feature and the neighboring feature of the same class. More precisely, we only apply SMOTE to the training samples.

We use the augmented training samples with a glass-box classifier called Explainable Boosting Machines (EBM) [43]. EBMs fall under the realm of explainable models that provide the optimal trade-off between the expressiveness of black-box models and high interpretability of linear models. Figure 7a details the overview the proposed training approach on electrical signals.

Images. Figure 7b shows the training pipeline for image based continuous authentication. To extend the explainability to images, we employ a representation learning method [4] to our algorithm. Instead of directly using a discriminative deep learning classifier, we first input the cropped images, $\mathbf{x} \in \mathcal{R}^D$ to a CNN backbone network, in this case, an Inception Net V1 [47] network pretrained on VGG Face dataset [8] which maps the input image $\mathbf{x}$ to a representation vector $\mathbf{r} = Backbone(x) \in \mathcal{R}^{D_e}$. Next, we feed the latent features $\mathbf{r}$ to a single layer MLP network called the projection network, which maps $\mathbf{r}$ to vector $\mathbf{z} = Proj(r) \in \mathcal{R}^{D_z}$. The D_z dimensional feature vector is optimized using a Supervised Contrastive Loss [32] which aims at maximizing the distance between the features of different classes and minimize the distance between the features of similar classes. Mathematically, the loss function is given by

$$\mathcal{L}_{con} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in N} \exp(z_i \cdot z_a / \tau)} \quad (1)$$

Where P is the set of all positive data points with the same class as i, except i, and N is the set of all data points (positives and negatives) except i. As we show in Section 5, this aids in minimizing the class imbalance problems for images and simultaneously, provide embeddings which show evidence of class separability in latent space. During test time, we freeze the backbone network to get the latent space and run it through a single MLP classifier to provide the authentication decision.

5. CA Performance Assessment

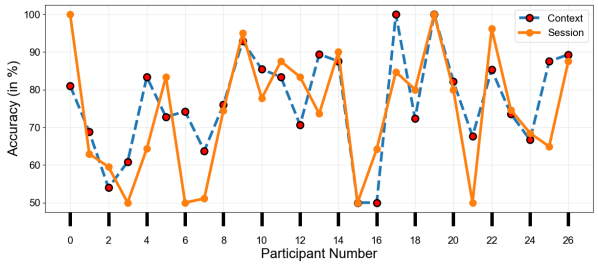
To evaluate the performance of both modalities (electrical signals and images), and assess the effectiveness of the proposed system, we perform two preliminary experiments based on train and validation data split. As mentioned in Section 4.2, the experiments are performed on the two modalities independently. For each of the experiments, to be consistent with literature [50, 22, 13], we report the following metrics - Accuracy, Precision, Recall, False Acceptance Rate (FAR), and False Reject Rate (FRR).

5.1. Train and Test Split

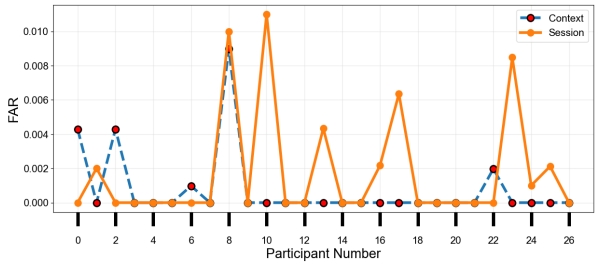
For the first experiment, we split our dataset in a *session-wise* manner. Since we construct a subject-specific model (i.e., model for each subject), if $\mathcal{S}$ is the list of all subjects in the dataset, all the data of a target subject $s \in \mathcal{S}$ from $\{session\ 1, session\ 2\}$ are enrolled as samples of *one* class in the training set and $\{session\ 3\}$ as the validation set. Similarly, to get a diverse representation of *all* class, we use $\{session\ 1, session\ 2\}$ data from the rest of

Table 3. Average accuracy, precision and recall scores list for each modality for session and context-based splits

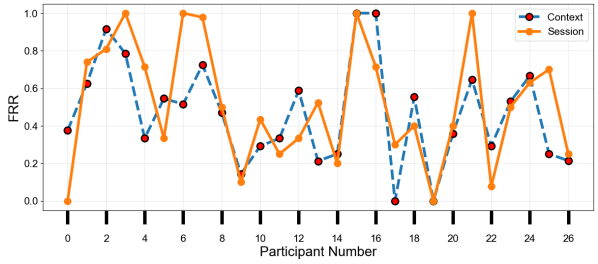
Modality	Split	Accuracy (SD) ↑	Precision ↑	Recall ↑
Physiology	Session	74.19 (15.08)	77.39	48.56
	Context	**76.59 (13.44)**	**88.90**	**53.25**
Image	Session	86.86 (16.66)	80.19	61.67
	Context	**90.39 (10.87)**	**98.77**	**75.71**



(a) Accuracy (in %)



(b) False Acceptance Rate (FAR)



(c) False Reject Rate (FRR)

Figure 8. Subject-wise accuracy, FAR, and FRR scores for electrical signals for both type of splits.

the subjects, $\overline{s} \in \mathcal{S} \setminus s$ as training set and $\{session\ 3\}$ as the validation set. Session-wise split analyzes the model's effectiveness in a longitudinal manner, especially in electrical signals, since the CA systems must show robustness against the long-term persistence of these signals. We use *session 3* as the validation set since by session 3, we hypothesize that the participant is most comfortable compared to session 1 and session 2. This reflects the real-world setting where the user is more likely to require CA on devices they are most accustomed to.

For the second experiment, we perform a *context-based* split on the dataset. This implies that, for a given subject, all the data belonging to 3 of the 4 contexts (Sad, Content, Disgust, Happy) from the three sessions are used as the training set and the fourth context is used as the validation set. In this work, we utilize $\{Sad, Content, Happy\}$ as the training set and $\{Disgust\}$ as the validation set. Similar to session-based split, the genuine users (s) are enrolled as part of *one* class, and all the other subjects ($\overline{s}$) are enrolled as part of *all* class. We designate *Disgust* as the validation context since it showed highest variance in terms of self-reported affect intensity (Figure 4b), and also displayed a good ratio of positive, negative and neutral affect compared to other contexts (Figure 4c).

Table 4. Top-5 features for electrical signals (session-based split).

Feature	Parameters	Importance Score
agg_linear_trend	**'attr': 'intercept', 'chunk_len': 50, 'f_agg': 'max'**	**0.49**
	'attr': 'intercept', 'chunk_len': 50, 'f_agg': 'mean'	0.47
change_quantiles	'f_agg': 'var', 'isabs': True, 'qh': 0.6, 'ql': 0.2	0.47
cwt_coefficients	'coeff': 11, 'w': 10, 'widths': (2, 5, 10, 20)	0.46
	'coeff': 12, 'w': 10, 'widths': (2, 5, 10, 20)	0.44
	'coeff': 3, 'w': 20, 'widths': (2, 5, 10, 20)	0.43
fft_aggregated	'aggtype': 'skew'	0.39
number_crossing_m	'm': 1	0.36

Table 5. Top-3 features for electrical signals (context-based split).

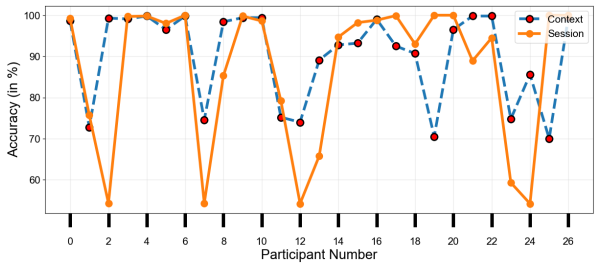
Feature	Parameters	Importance Score
agg_linear_trend	**'attr': 'intercept', 'chunk_len': 5, 'f_agg': 'mean'**	**0.74**
fft_coefficient	'attr': 'angle', 'coeff': 56	0.33
	'attr': 'angle', 'coeff': 65	0.30
	'attr': 'real', 'coeff': 24	0.29
ratio_beyond_r_sigma	'r': 5	0.28

5.2. Electrical Signals

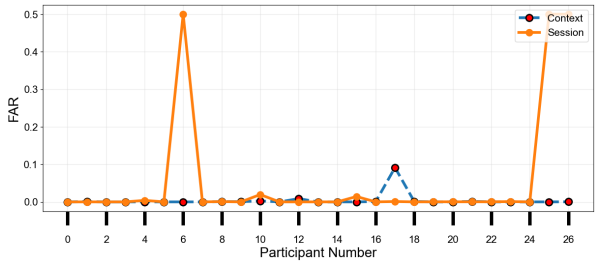
It is important to note that for electrical signals, we perform a search for the best modality for subject-wise models. Therefore, for each subject, we save the modality with the best overall performance and assign the modality as the primary feature set for the subsequent experiment (i.e., context-based split). This is to ensure the features are consistent thereby providing reliable results.

Table 3 (top row) reports the average accuracy (with standard deviation), precision, and recall scores for both session and context-based split. Context-based split performs better than session-wise split, especially in terms of precision. This can be explained, in part, by the same session data being available in both training and validation. This result implies that the model has a low false positive rate thereby limiting incorrect access to imposter users. The same can be observed in Figure 8 which plots the subject-wise accuracy, FAR, and FRR. Intuitively, the high number of false positives can be observed in Figures 8a and 8b which tends to show higher frequency of troughs and peaks, respectively, for session-based split vs. context-based split.

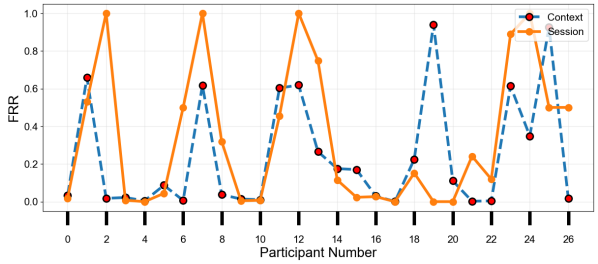
Due to the large amount of features, we use the transparency of the EBMs to provide the most important features used for authentication. EBMs are a family of additive models [24] and therefore, supports a global explanation metric called mean absolute score, which is an average absolute



(a) Accuracy (in %)



(b) False Acceptance Rate (FAR)



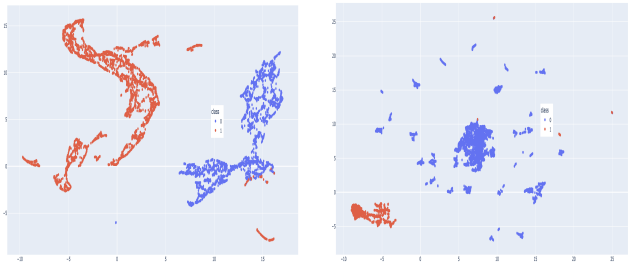
(c) False Reject Rate (FRR)

Figure 9. Subject-wise accuracy, FAR, and FRR scores for images for both type of splits

contribution a feature makes on each sample in the training set. Tables 4 and 5 report the top-5 and top-3 features for session and context-based split, respectively. We also present the specific parameter combination and the corresponding scores for each. In can be observed that in both cases, aggregate linear trend [51] accounts for the most contribution. Additionally, we see the frequency features (continuous wavelet transform and fast Fourier coefficients) add value to the model. A deeper analysis into the feature sets is out of the scope of this paper and is left for future work.

5.3. Images

Table 3 (bottom row) reports the average accuracy, precision, and recall for both the experiments. Additionally, Figure 9 shows the subject-wise results for accuracy, FAR, and FRR. Similar to electrical signals, the metrics suggest an improved performance on context-based split compared to session-based split. Moreover, the performance of image-based CA is superior to that of electrical signals across all splits. This can be explained, in part, by images allowing extraction of semantic information from the data compared to electrical signals, which consists of higher volume of artifacts. There is, however, a trade-off between performance



(a) Session-based split (b) Context-based split

Figure 10. Sample UMAP embeddings of deep features, $r \in \mathcal{R}^{D_e}$.

and latency of authentication decisions between the two. The proposed method on electrical signals allows for feature extraction every $\simeq 1.5$ seconds, which is relatively efficient compared to the computationally heavy face detection and cropping, followed by a deep learning inference on images.

In order to showcase the ability of the deep learning model to distinuish between classes, we visualize the UMAP embeddings [40] of the latent output, $r \in \mathcal{R}^{D_e}$. Figures 10a and 10b shows the visualization of a sample subject's validation set for both the splits. We can observe a distinct segregation between the feature points of genuine and imposter users for both splits.

6. Conclusion

We proposed a new multimodal, continuous authentication dataset centered around emotions as context and across multiple sessions. We collected videos of 27 participants under four different affect contexts (Sad, Content, Disgust, and Happy) per session and the corresponding electrical signals (accelerometer, electrodermal activity, temperature, and heart rate). We also collected self-report information of the participant in the form of compound emotional labels, discretized valence levels, and continuous arousal values. We performed continuous authentication experiments on both modalities based on session, and context-based data splits. We also highlighted the feature contribution using the mean absolute scores offered by EBMs. While these results are encouraging, there are some limitations. First, only one context (disgust) was used for validation in the context-based data split. We will validate other contexts in future work. We also validated the modalities separately. a fusion-based approach may boost performance, as it has been effective in past works [26, 31]. We will also collect a larger more uniform dataset across age, gender, and ethnicity.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants No. 2039373 and 2238389. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Ecg analysis: a new approach in human identification. *IEEE transactions on instrumentation and measurement*, 50(3):808–812, 2001. 1

[2] M. Agrawal, P. Mehrotra, R. Kumar, and R. R. Shah. Defending touch-based continuous authentication systems from active adversaries using generative adversarial networks. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 1

[3] H. Aidos and A. L. N. Fred. k-nearest neighbor classification using dissimilarity increments. In *International Conf. on Image Analysis and Recognition*, volume 7324, pages 27–33, June 2012. 2

[4] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 6

[5] G. Biagetti, P. Crippa, L. Falaschetti, L. Saraceni, A. Tiranti, and C. Turchetti. Dataset from ppg wireless sensor for activity monitoring. *Data in brief*, 29:105044, 2020. 2

[6] L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001. 2

[7] M. Brunet, T. Van Gelder, A. Åsberg, V. Haufroid, D. A. Hesselink, L. Langman, F. Lemaitre, P. Marquet, C. Seger, M. Shipkova, et al. Therapeutic drug monitoring of tacrolimus-personalized therapy: second consensus report. *Therapeutic drug monitoring*, 41(3):261–307, 2019. 1

[8] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 6

[9] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307:72–77, 2018. 5

[10] M. Christ, A. W. Kempa-Liehr, and M. Feindt. Distributed and parallel time series feature extraction for industrial big data applications. *arXiv preprint arXiv:1610.07717*, 2016. 5

[11] D. M. Clark. On the induction of depressed mood in the laboratory: Evaluation and comparison of the velten and musical procedures. *Advances in Behaviour Research and Therapy*, 5(1):27–49, 1983. 3

[12] S. Cosentino, E. I. Randria, J.-Y. Lin, T. Pellegrini, S. Sessa, and A. Takanishi. Group emotion recognition strategies for entertainment robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 813–818. IEEE, 2018. 5

[13] D. Crouse, H. Han, D. Chandra, B. Barbello, and A. K. Jain. Continuous authentication of mobile user: Fusion of face image and inertial measurement unit data. In *2015 International Conference on Biometrics (ICB)*, pages 135–142. IEEE, 2015. 1, 2, 6

[14] G. Dahia, L. Jesus, and M. Pamplona Segundo. Continuous authentication using biometrics: An advanced review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(4):e1365, 2020. 1

[15] S. Dargan and M. Kumar. A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. *Expert Systems with Applications*, 143:113114, 2020. 1

[16] R. J. Davidson, P. Ekman, C. D. Saron, J. A. Senulis, and W. V. Friesen. Approach-withdrawal and cerebral asymmetry: emotional expression and brain physiology: I. *Journal of personality and social psychology*, 58(2):330, 1990. 4

[17] E. Derman and A. A. Salah. Continuous real-time vehicle driver authentication using convolutional neural network based face recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 577–584. IEEE, 2018. 1

[18] S. Eberz, K. B. Rasmussen, V. Lenders, and I. Martinovic. Evaluating behavioral biometrics for continuous authentication: Challenges and metrics. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 386–399, 2017. 1

[19] D. Fabiano and S. Canavan. Emotion recognition using fused physiological signals. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 42–48. IEEE, 2019. 1

[20] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018. 6

[21] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern recognition letters*, 31(14):2225–2236, 2010. 2

[22] S. R. K. Gopal and D. Shukla. A temporal memory-based continuous authentication system. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2021. 2, 5, 6

[23] J. J. Gross and R. W. Levenson. Emotion elicitation using films. *Cognition & emotion*, 9(1):87–108, 1995. 3

[24] T. J. Hastie. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge, 2017. 7

[25] S. Hinduja, S. Canavan, and G. Kaur. Multimodal fusion of physiological signals and facial action units for pain recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 577–581. IEEE, 2020. 5

[26] Y. Huang, J. Yang, P. Liao, and J. Pan. Fusion of facial expressions and eeg for multimodal emotion recognition. *Computational intelligence and neuroscience*, 2017, 2017. 8

[27] D. Y. Hwang, B. Taha, and D. Hatzinakos. Variation-stable fusion for ppg-based biometric system. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8042–8046. IEEE, 2021. 2

[28] D. Y. Hwang, B. Taha, D. S. Lee, and D. Hatzinakos. Evaluation of the time stability and uniqueness in ppg-based biometric system. *IEEE Transactions on Information Forensics and Security*, 16:116–130, 2020. 1

[29] S. R. Jannat, D. Fabiano, S. Canavan, and T. Neal. Subject identification across large expression variations using 3d facial landmarks. In *Pattern Recognition. ICPR International*

Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I, pages 5–13. Springer, 2021. 3

[30] A. Kaso. Computation of the normalized cross-correlation by fast fourier transform. *PloS one*, 13(9):e0203434, 2018. 5

[31] T. Keshari and S. Palaniswamy. Emotion recognition using feature-level fusion of facial expressions and body gestures. In *2019 international conference on communication and electronics systems (ICCES)*, pages 1184–1189. IEEE, 2019. 8

[32] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 2, 6

[33] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011. 2

[34] K. Kyriakou, B. Resch, G. Sagl, A. Petutschnig, C. Werner, D. Niederseer, M. Liedlgruber, F. Wilhelm, T. Osborne, and J. Pykett. Detecting moments of stress from measurements of wearable physiological sensors. *Sensors*, 19(17):3805, 2019. 1

[35] T.-M. Lee, P.-L. Lee, I.-H. Lee, W.-K. Lee, T.-Y. Wu, H.-T. Hsu, C.-L. Yeh, P.-J. Lin, and K.-K. Shyu. Study of heart-rate variability in a video task using holo-hilbert spectral analysis. *Biomedical Signal Processing and Control*, 71:103229, 2022. 1

[36] G. Lovisotto, H. Turner, S. Eberz, and I. Martinovic. Seeing red: Ppg biometrics using smartphone cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 818–819, 2020. 1

[37] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 6

[38] M. Martinho, A. Fred, and H. Silva. Towards continuous user recognition by exploring physiological multimodality: An electrocardiogram (ecg) and blood volume pulse (bvp) approach. In *2018 International Symposium in Sensing and Instrumentation in IoT Era (ISSI)*, pages 1–6. IEEE, 2018. 2

[39] R. Matta, J. K. Lau, F. Agrafioti, and D. Hatzinakos. Real-time continuous identification system using ecg signals. In *2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 001313–001316. IEEE, 2011. 1

[40] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 8

[41] H. Mori, H. Yamamoto, M. Kuwashima, S. Saito, H. Ukai, K. Hirao, M. Yamauchi, and S. Umemura. How does deep breathing affect office blood pressure and pulse rate? *Hypertension research*, 28(6):499–504, 2005. 3

[42] S. T. Murphy and R. B. Zajonc. Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. *Journal of personality and social psychology*, 64(5):723, 1993. 3

[43] H. Nori, S. Jenkins, P. Koch, and R. Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019. 6

[44] A. Rai. Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48:137–141, 2020. 2

[45] S. Rasnayaka and T. Sim. Action invariant imu-gait for continuous authentication. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2022. 5

[46] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018. 2

[47] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 6

[48] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang. A review of emotion recognition using physiological signals. *Sensors*, 18(7):2074, 2018. 1

[49] I. Stylios, S. Kokolakis, O. Thanou, and S. Chatzis. Key factors driving the adoption of behavioral biometrics and continuous authentication technology: an empirical research. *Information & Computer Security*, 2022. 1

[50] H. C. Volaka, G. Alptekin, O. E. Basar, M. Isbilen, and O. D. Incel. Towards continuous authentication on mobile phones using deep learning models. *Procedia Computer Science*, 155:177–184, 2019. 6

[51] G. S. Watson. Linear least squares regression. *The Annals of Mathematical Statistics*, pages 1679–1699, 1967. 8

[52] G. I. Webb, E. Keogh, and R. Miikkulainen. Naïve bayes. *Encyclopedia of machine learning*, 15:713–714, 2010. 2

[53] L. Xiaofeng, Z. Shengfei, and Y. Shengwei. Continuous authentication by free-text keystroke based on cnn plus rnn. *Procedia computer science*, 147:314–318, 2019. 1

[54] J. Zhang, Z. Yin, P. Chen, and S. Nichele. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, 59:103–126, 2020. 1

[55] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 3

[56] Y. Zhang, W. Hu, W. Xu, C. T. Chou, and J. Hu. Continuous authentication using eye movement response of implicit visual stimuli. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–22, 2018. 1