EDOSE: Emotion Datasets from Open Source EEG with a Real-Time Bracelet Sensor

Payongkit Lakhan, Nannapas Banluesombatkul, Vongsagon Changniam, Ratwade Dhithijaiyratn, Irawadee Thawornbut, Ekkarat Boonchieng and Theerawit Wilaiprasitporn, *Member, IEEE*

***Abstract*— This is the first concrete investigation of emotion recognition capability or affective computing using a low-cost, open source electroencephalography (EEG) amplifier called OpenBCI. The most important aspect for this kind of study is effective emotional elicitation. According to state-of-the-art related works, movie clips are widely used for visual-audio emotional elicitation. Here, two-hundred healthy people of various age ranges participated in an experiment for effective clip selection. Typical self-assessment, affective tags and unsupervised learning (k-mean clustering method) were incorporated to select the top 60 effective clips from 120 candidates. A further 43 participants gathered for an emotional EEG using OpenBCI and peripheral physiological signals using a real-time bracelet sensor while watching the selected clips. To evaluate the performance of OpenBCI toward emotion-related applications, the data on binary classification tasks was analyzed to predict whether elicited EEG has a high or low level of valence/arousal. As in the previous study on emotional EEG datasets, power spectral densities were extracted as the input features for a basic machine learning classifier; the support vector machine. The experimental results for the proposed datasets or EDOSE outperformed those from the state-of-the-art EEG datasets in a study of affective computing, namely DEAP, MAHNOB-HCI, DECAF and DREAMER. Finally, the EDOSE dataset can assist the researcher (upon request) in various fields such as affective computing, human neuroscience, neuromarketing, mental health, etc.**

***Index Terms*— Emotion EEG, Emotion recognition, Affective computing, Emotion datasets, OpenBCI**

I. Introduction

THE twenty-first century is experiencing a dramatic increase in wearable devices for physiological data acquisition. Sensors, storage and computational resources are inexpensive with acceptable accuracy. Such devices are rapidly becoming game changers in human behavioral studies. The low-cost, open source electroencephalography (EEG) amplifier from OpenBCI Inc. is one great example of a game changing device [1]. Researchers and individual investigators can now study human brain activities on a budget of USD1,000–2,000. Emotion recognition from EEG is a challenging concept in human behavioral research which tries to predict or classify human mental responses from an external stimulus. Several feasible applications of emotion recognition are used to measure the quality of experience (QoE) from multimedia perception [2], mental health assessment such as for depression [3] and human computer interaction.

In 1980, Russel proposed valence-arousal (VA) scores as a standard qualitative measurement for affective or emotion-related studies [4]. The Self-Assessment Manikin (SAM) is advantageous for VA scores in the evaluation of emotion responses [5]. The valence score measures feelings ranging from negativity (e.g. unpleasantness, sadness, stress) to positivity (e.g. pleasantness, happiness, elation). The arousal score measures feelings ranging from inactivity (e.g. boredom, disinterest) to activity (e.g. excitement, alertness) [6], [7]. DEAP [6], MAHNOB-HCI [8], DREAMER [7] and SEED [9] are well-known emotion datasets which mainly include EEG (brain wave), electrodermal activity (EDA) and heart rate variability (HRV) responses from emotion-elicited VDO clips. EDA and HRV are peripheral physiological signals to EEG which are strongly related to the sympathetic nervous system. The sympathetic system is usually activated by emotional induction.

According to the existing emotional datasets, the uniqueness of each can be summarized into two major perspective views. Firstly, the methods of video clip based stimulus selection are shown in Table I. All datasets use the affective related qualitative measures. However, in this study, the selected clips containing in the proposed datasets, namely *EDOSE*, came from the largest number of participants (the largest population) which was about four times greater than the other datasets. Using affective tags and qualitative scoring with the k-mean algorithm, 60 potential clips were finally selected from 120 clip candidates for EDOSE. Secondly, the practicality of emotional EEG based-applications is mainly considered as presented in Table II. For EDOSE, the OpenBCI-EEG amplifier was used since it is portable, open source, low cost, and has a small number of electrodes (more user-friendly) [1]. On the other hand, EEG datasets from DEAP, MAHNOB-HCI and SEED were based on high-end amplifiers. Although DREAMER is a low-cost device, it is not open source and has

This work was supported by The Thailand Research Fund and Office of the Higher Education Commission under Grant MRG6180028 and Junior Science Talent Project, NSTDA, Thailand.

P. Lakhan, N. Banluesombatkul and T. Wilaiprasitporn are with Bio-inspired Robotics and Neural Engineering Lab, School of Information Science and Technology, Vidyasirimedhi Institute of Science & Technology, Rayong, Thailand theerawit.w at vistec.ac.th

V. Changniam is with Department of Tool and Materials Engineering, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

R. Dhithijaiyratn is with Department of Electrical Engineering, Chulalongkorn University, Bangkok, Thailand

I. Thawornbut is with British International School, Phuket, Thailand

E. Boonchieng is with Center of Excellence in Community Health Informatics, Chiang Mai University, Chiang Mai, Thailand

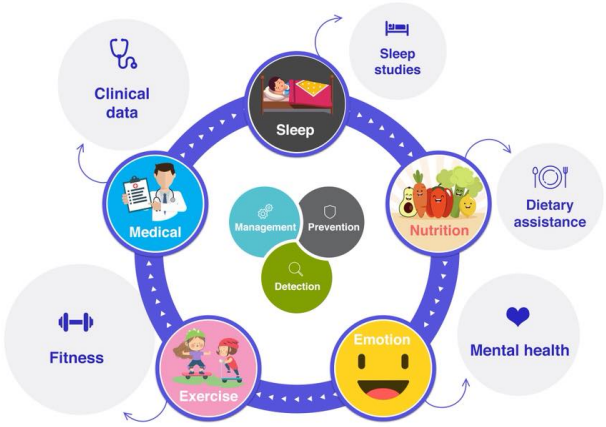


Fig. 1: Interested data in on-going research. EDOSE is located in the emotion/mental health category.

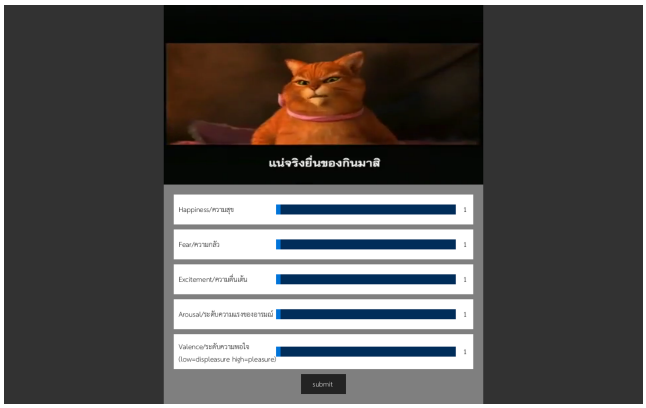


Fig. 2: Experimental screen for clip selection.

a higher number of electrodes. Moreover, DREAMER datasets were gathered from approximately only half the number of participants compared to EDOSE.

Even if the EEG data in EDOSE was obtained from low sampling frequency and a small number of electrodes, it outperformed DEAP, MAHNOB-HCI and DREAMER in conventional classification tasks in valence and arousal using a similar algorithm. The algorithm involves simple feature extraction, such as power spectral density (PSD), with a classifier constructed from a support vector machine (SVM). In contrast to DREAMER, EDOSE not only contains EEG data but also sympathetic data from a real-time bracelet sensor named Empatica4 or E4 [10]. E4 data contains EDA, skin temperature and blood volume pulse (BVP) including their derivatives (heart rate, inter-beat interval, etc.), all of which are indirect measures of sympathetic nervousness. In this paper, EDOSE is proposed as the first open access emotional-induced EEG database from a low-cost, open source EEG amplifier: OpenBCI. EDOSE can facilitate the researcher (upon request) in various fields such as affective computing, human neuroscience, neuromarketing, mental health, etc. We are currently working on a smart living research studio for predictive health care and EDOSE (emotion/mental health) data is also part of the study, see Figure 1.

The remainder of this paper includes the methodology used in two experiments in Section II. Section III presents the data analysis on emotion recognition, with the results and discussion reported and explained in sections IV and V, respectively. Finally, section VI summarizes the paper.

II. Methodology

This section begins with an experiment (Experiment I) to select the effective film trailers for emotional elicitation. The selected trailers are then used to elicit human moods. Finally, EEG and peripheral physiological signals are recorded and analyzed as described in Experiment II. The experiments follow the Helsinki Declaration of 1975 (as revised in 2000), approved by the internal review board of the Chiangmai University Thailand.

A. Experiment I: Elicited Video Selection

Film trailers are among the affective tags labeled by the Internet Movie Database (IMDb), in which drama (happiness), comedy, romance, action, sci-fi, horror and mystery are main categories. These are simply defined in this paper as three genres: Happiness, Excitement and Fear, respectively. Five participants randomly picked 40 famous film trailers per genre, resulting in 120 trailer clips in total (sound tracks with subtitles). Afterward, the 120 clips were randomly split into four groups of 30 clips each. Then, 200 participants ($n = 200$), of almost equal numbers of males and females, ranging from 15–22 years old, watched at least one out of the four groups. To deal with the different trailer durations, the last minute of each clip was selected for the experiment. Finally, each participant evaluated his/her emotions through the qualitative measure, of scoring for Valence (V), Arousal (A), Happiness (H), Fear (F) and Excitement (E) (qualitative measures on a continuous scale of 1–9) after each clip had finished playing (Figure 2).

In an analysis of the qualitative data evaluated by a group of sample participants, the k-means clustering method was used [12], with H, F and E scores as the features to find three extremely different emotional elicitations (no. of clusters or k = 3). As a consequence of the clustering, 20 clips were selected from each cluster closest to the centroids for further use in Experiment II. Thus, 60 emotional elicited clips were selected in total.

Due to the same clips being scored by different participants, the k-mean outputs from the same clips may not be classified into the same clusters. This presented a problem in the calculation of average euclidean distances from the outputs to the centroids of the clusters. To overcome this issue, a majority cluster (mode value) of each clip was obtained and the k-mean outputs then filtered out from the same clips which did not fall into the majority cluster. Finally, the average euclidean distances were calculated from the rest outputs of each clip to the centroid of its cluster. In this way, the first 20 clips closest to the centroids of each cluster could be ranked for the following experiment.

TABLE I: Video clip based stimulus selection from the proposed and existing datasets

Database	No. participant	No. video clip	No. selected video clip	Selection method
DEAP [6]	14	1,084	40	manual selection, affective tags, maxmize ratings
MAHNOB-HCI [8]	50	155 (from 21 movies)	20	manual selection, affective tags
DREAMER [7]	-	-	18	all clips from previous study [11]
SEED [9]	44	168	72	affective tags, mean ratings distribution
EDOSE	200	120	60	affective tags, scorings, k-mean algorithm

TABLE II: Comparison of EEG amplifiers, number of participants from the proposed and existing datasets

Database	EEG devices	No. electrodes	Sampling rate [Hz]	Low cost	Open source	Portable	No. participant
DEAP [6]	Biosemi Active II	32	512	No	No	No	32
MAHNOB-HCI [8]	Biosemi Active II	32	512	No	No	No	30
DREAMER [7]	Emotiv Epoc	14	256	Yes	No	Yes	23
SEED [9]	ESI NeuroScan	6(temporal)	1000	No	No	No	15 (3 days/each)
EDOSE	OpenBCI	8	250	Yes	Yes	Yes	43

Fig. 3: Experimental setup.

B. Experiment II: Gathering Emotional EEG and Peripheral Physiological Signals

1) Participants and Sensors: The participants of this study consisted of 43 healthy people aged between 16 and 34 ($n = 43$, male = 21, female = 22). They were asked to wear only two devices, OpenBCI (EEG) and Empatica4 (E4). Figure 3 shows the experimental setups. To cover brain wave information across the scalp, an eight-channel EEG electrode montage was selected (Fp1, Fp2, Fz, Cz, T3, T4, Pz and Oz) with reference and ground on both mastoids. This montage is a subset of the international standard 10–20 system for EEG electrode placement. For a bracelet sensor or E4, multiple sensors were equipped, including electrodermal activity (EDA), skin temperature (Temp) and blood volume pulse (BVP) with their derivatives (heart rate, inter-beat interval, etc.). The participants wore E4 like a watch, although there were extension cables with gel-electrodes attached (the Kendall Foam Electrodes) to measure EDA at the middle knuckle of the index and middle fingers. Both devices recorded simultaneously. During the experiment, participants were asked to place their chin on the designated chin rest, and move as little as possible, in order to prevent artifact interference.

2) Experimental Protocol: We developed software to present the emotion stimulus clips and record EEG with peripheral physiological signals from the wearable devices as shown in Figure 4. The first page of the program consists of a survey to collect relevant information from the participants (age, gender, hours of sleep, level of tiredness and favorite movie genre). As in Experiment I, the video clips for the participants were randomly selected into forty three groups with fifteen clips each. Each group consisted of nine identical clips including three per cluster (the top three clips from each cluster as explained in Experiment I). The remaining six clips were selected randomly; two per cluster. Following each clip, participants completed a survey (scroll bars) to evaluate their emotions, including: V, A, H, F and E (on a continuous scale of 1–9).

III. Data Analysis

In this section, the data is analyzed according to Valence (V), Arousal (A), Happiness (H), Fear (F) and Excitement (E) classification (binary tasks, low-high) using data from OpenBCI and E4. Furthermore, the related EEG related factors (electrode channels, frequency bands) are studied to assist in future applications such as the development of a practical system for emotion recognition.

A. Feature Extraction

In order to obtain stability and specific emotional responses from the EEG and peripheral physiological signals, the program started recording signals after the stimulus clip had been played for two seconds. In order to avoid filtering artifacts, two seconds of recorded signals were removed before the end of the clip. In total, there were 56 seconds of signals for each elicited clip. Typical EEG-preprocessing was incorporated, including common average reference (CAR) to find the reference signal for all electrode channels, and independent component analysis (ICA) used to remove artifact components. Even if there was no electrooculography (EOG) in the experiments, ICA components were removed upon inspection by the expert. Both CAR and ICA were implemented using the MNE-python package [13]. Conventional feature extraction, as shown Table III, was incorporated into the recorded data, including EEG, EDA, BVP and Temp. These features were based on the previous works [6]–[8], as the baseline for comparison with the results of this study.

B. Classification

1) binary classification tasks with labeling by simple thresholding: The recorded signals were taken from a participant

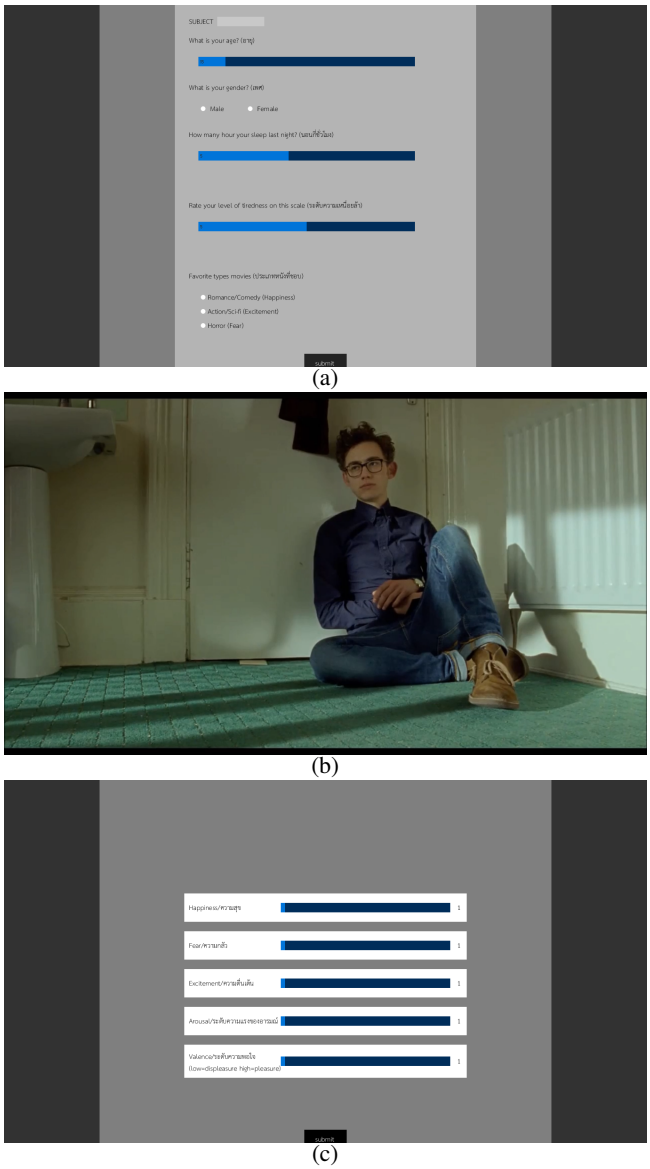


Fig. 4: (a) represents the beginning of the experimental program used to acquire demographics, hours of sleep, level of tiredness and favorite movie genre. (b) and (c) represent the clip player in full screen mode and the self-emotional assessment form, respectively.

with one clip as a single sample. Hence, there were 645 (43 participants $\times$ 15 clips) samples in total. We labeled each sample using self-emotional assessment scores (V, A, H, F and E). Due to binary classification purpose, we manually set the threshold of each scores to be 5 (from scores range 1-9). For example, samples with H scores lower than 5 were labeled as low H and high H vice versa. Here, we had six binary classification tasks which were low-high of V, A, H, F and E. Table III shows the set of input features for the SVM with a linear kernel classifier in all tasks. We incorporated *gridsearchcv* to achieve an optimal set of SVM parameters. A leave-one-clip-out cross-validation was performed on these experiments. Since only nine out of 15 clips were watched by all participants, nine-fold cross-validation was conducted. In each fold, one common clip was selected as a test sample and the others were training samples.

2) binary classification tasks with labeling by k-mean clustering: Using the same population as in the previous subsection, k-mean clustering was conducted using their V and A scores. According to four combinations of V and A models in a conventional emotion-related study (low (L) or high (H) V and A scores), the number of clusters or k-mean was set to four (LVLA, LVHA, HVLA, LVHA). Figure 6 shows the results of k-mean clustering. The samples in each cluster were then labeled following their own clusters, either low or high V and A. The classification tasks were conducted as in the previous subsection except we interested only V and A classification tasks. Here, the same clip may not have an identical label, so nine-fold cross-validation was performed based on samples instead of leaving-one-clip-out.

EEG electrode channels: In this study, the electrode setups were explored for the future hardware design of a user-friendly EEG device with the capability for emotion recognition. Here, three or four out of eight electrode channels were strategically selected as mentioned in Section II, B. The process began by studying the middle line of the human scalp (F_z, C_z, P_z O_z). Another eight sets of channels were then created for the study (FP_1, FP_2, F_z), (FP_1, FP_2, C_z), (FP_1, FP_2, C_z), (FP_1, FP_2, P_z), (FP_1, FP_2, O_z), (T_3, T_4, F_z), (T_3, T_4, C_z), (T_3, T_4, P_z) and (T_3, T_4, O_z). These were taken from a combination of either the temporal lobe or frontal lobe and one channel from the middle line. Binary classification tasks (low-high V and A) were subsequently performed as in the previous experiments and the results compared among the sets of channels.

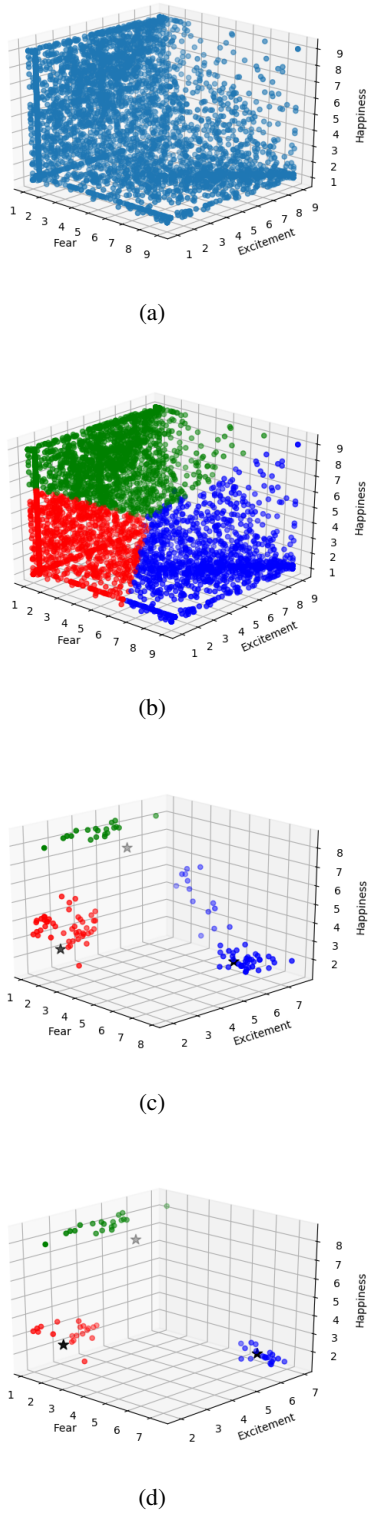
EEG frequency bands: The aim of this study is to find important frequency bands for EEG signals. From four basic frequency band signals, θ (3–7 [Hz]), α (8–13 [Hz]), β (14–29 [Hz]) and γ (30–47 [Hz]), the following combinations of interested frequencies were created: (θ, α), (θ, β), (θ, γ), (α, β), (α, γ), (β, γ), (θ, α, β), (θ, α, γ), (θ, β, γ), (α, β, γ). Finally, binary task classification (low-high V and A) was performed as in the previous experiments and the results subsequently compared among the sets of frequency bands.

IV. Results

Here, the results are sequentially organized according to the experiments. K-mean clustering results for the stimulus clip selection were performed step-by-step to gather the most effective clips for EEG and peripheral physiological signals in the experiments. Finally, the gathered datasets were analyzed using a simple machine learning algorithm and the results compared to the previous related datasets for applications of affective computing.

A. Elicited Video Selection

Qualitative results, Happiness (H), Fear (F) and Excitement (E) scores from all participants on every candidate clip, were scattered on three dimensional spaces (H, F, E), as shown in Figure 5 (a). In Figure 5 (b), all scattered points (samples) were labeled (colorized) according to three output clusters using the k-mean method. As explained in Section II: Experiment I, the



(a)

(b)

(c)

(d)

Fig. 5: (a) was the scatter plot of all qualitative samples from Experiment I. K-mean clustering on (a) provided the output cluster as in (b). After removing the qualitative samples not belonging to the same majority cluster of the samples from each clip, the remaining candidates were performed in (c). Eventually, in (d), the 20 points per cluster were retained to give the nearest distances to the centroids for emotional stimulation in Experiment II.

TABLE III: Features extracted from EEG and empatica signals in each clip

Signal	Extracted features
EEG(32)	θ (3–7 [Hz]), α (8–13 [Hz]), β (14–29 [Hz]) and γ (30–47 [Hz]) power spectral density for each channel.
EDA(21)	average skin resistance, average of derivative, average of derivative for negative values only, proportion of negative samples in the derivative vs all samples, number of local minima in the EDA signal, average rising time of the EDA signal, 14 spectral power in the 0–2.4 [Hz] bands, zero crossing rate of Skin conductance slow response 0–0.2 [Hz], zero crossing rate of Skin conductance very slow response 0–0.08 [Hz].
BVP(13)	average and standard deviation of HR, HRV, and inter beat intervals, energy ratio between the frequency bands 0.04–0.15 [Hz] and 0.15–0.5 [Hz], spectral power in the bands 0.1–0.2 [Hz], 0.2–0.3 [Hz], 0.3–0.4 [Hz], low frequency 0.01–0.08 [Hz], medium frequency 0.08–0.15 [Hz] and high frequency 0.15–0.5 [Hz] components of HRV power spectrum.
Temp.(4)	average, average of its derivative, spectral power in the band 0–0.1 [Hz], 0.1–0.2 [Hz].

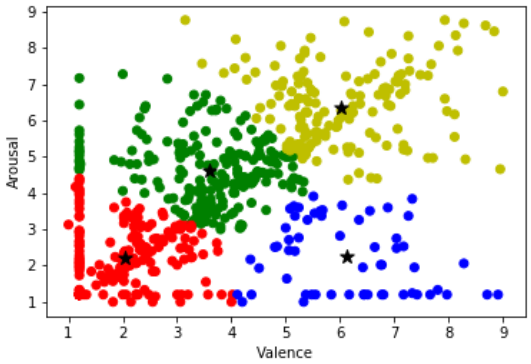


Fig. 6: K-mean clustering using valence and arousal scores labeled

qualitative results or scattered points were removed from a clip not labeled the same as the majority of labels in that clip. Thus, the scatter points of each cluster were subsequently obtained as shown in Figure 5 (c). Finally, only 20 points in each cluster (60 clips in total) were retained to give the nearest distances to the centroids for emotional stimulation in Experiment II, Figure 5 (d). Here, we hypothesized that the selected clips (Table IV) using the proposed unsupervised learning were the most effective for emotional elicitation.

B. Binary Classification Tasks with Labeling by Simple Thresholding

In these tasks, ground truth labels were created using self-emotional assessment scores (V, A, H, F and E) empirically setting the threshold at five. Scores higher than the threshold were assigned to the high level class and vice versa for scores lower than the threshold. The input features extracted from

TABLE IV: Selected clips used to elicit emotion during the recording of EEG and peripheral physiological signals in Experiment II. The mean and standard deviations of qualitative results from each clip were also calculated and inserted into this table. The samples (Samp.) denote the number of participants who performed emotional self-assessment using qualitative measures after watching the clip.

ID	Movie clip trailers	Affective Tags (IMDb)	Valence	Arousal	Happiness	Fear	Excitement	Samp.
1	The Help	Drama	4.28 ± 2.08	3.34 ± 1.96	4.41 ± 2.12	1.26 ± 0.37	2.56 ± 1.66	82
2	Star Wars: The Last Jedi	Action, Adventure, Fantasy, Sci-Fi	4.98 ± 2.15	4.36 ± 2.08	4.06 ± 2.16	1.79 ± 1.00	5.09 ± 2.12	81
3	Suicide Squad	Action, Adventure, Fantasy, Sci-Fi	4.73 ± 2.02	4.12 ±1.90	4.36 ± 2.18	1.73 ± 1.41	4.37 ± 2.14	81
4	Pacific Rim	Action, Adventure, Sci-Fi	4.10 ± 2.03	3.81 ± 2.16	3.93 ± 2.04	1.31 ± 0.52	4.25 ± 2.12	44
5	War for the Planet of the Apes	Action, Adventure, Drama, Sci-Fi, Thriller	4.73 ± 2.06	4.04 ± 2.20	2.57 ± 1.90	2.67 ± 1.76	4.83 ± 2.08	41
6	God Help the Girl	Drama, Musical, Romance	4.38 ± 2.45	3.09 ± 2.02	5.00 ± 2.45	1.27 ± 0.47	3.01 ± 1.85	43
7	Rogue One: A Star Wars Story	Action, Adventure, Sci-Fi	5.09 ± 1.78	4.65 ± 2.15	3.34 ± 2.17	2.04 ± 1.23	5.28 ± 2.27	42
8	Blade Runner 2049	Drama, Mystery, Sci-Fi	4.44 ± 2.47	4.34 ± 2.38	3.15 ± 2.02	2.39 ± 1.73	4.81 ± 2.29	44
9	Hope Springs	Comedy, Drama, Romance	4.78 ± 2.46	3.56 ± 2.12	5.09 ± 2.34	1.19 ± 0.17	2.84 ± 1.89	45
10	Ghost in the Shell	Action, Drama, Sci-Fi	4.77 ± 2.10	4.28 ± 2.27	3.41 ± 2.13	2.07 ± 1.57	5.01 ± 2.27	47
11	Point Break	Action, Crime, Sport	4.65 ± 2.33	4.49 ± 2.40	3.31 ± 2.33	2.08 ± 1.41	5.12 ± 2.56	40
12	The Hunger Games	Adventure, Sci-Fi, Thriller	5.42 ± 2.26	4.76 ± 2.18	3.66 ± 2.09	2.06 ± 1.34	5.38 ± 2.05	42
13	Crazy, Stupid, Love.	Comedy, Drama, Romance	4.82 ± 2.54	4.01 ± 2.52	5.12 ± 2.64	1.21 ± 0.25	3.06 ± 2.35	43
14	Arrival	Drama, Mystery, Sci-F	4.84 ± 2.22	4.82 ± 2.28	3.28 ± 2.03	2.64 ± 1.90	5.66 ± 2.40	44
15	Mr. Hurt	Comedy, Romance	4.50 ± 2.21	3.59 ± 2.20	4.66 ± 2.19	1.23 ± 0.21	2.41 ± 1.43	42
16	American Assassin	Action, Thriller	4.19 ± 2.18	4.80 ± 2.33	2.90 ± 1.88	2.48 ± 1.92	5.03 ± 2.36	43
17	G.I. Joe: Retaliation	Action, Adventure, Sci-Fi	4.69 ± 2.39	4.11 ± 2.31	3.46 ± 2.09	1.55 ± 0.93	5.02 ± 2.54	48
18	Beginners	Comedy, Drama, Romance	4.42 ± 2.45	3.08 ± 2.18	4.97 ± 2.20	1.23 ± 0.27	2.42 ± 1.48	44
19	Open Grave	Horror, Mystery, Thriller	3.70 ± 1.90	4.70 ± 2.30	1.90 ± 1.14	4.03 ± 2.40	4.55 ± 2.46	43
20	Flipped	Comedy, Drama, Romance	5.05 ± 2.69	3.75 ± 2.40	5.43 ± 2.48	1.16 ± 0.08	2.88 ± 2.12	44
21	The Choice	Drama, Romance	4.58 ± 2.11	4.12 ± 1.98	4.63 ± 2.16	1.43 ± 0.68	3.24. ± 1.71	82
22	Danny Collins	Biography, Comedy, Drama	4.54 ± 2.20	3.55 ± 2.07	5.01 ± 2.09	1.17 ± 0.09	2.76 ± 1.68	82
23	The Big Sick	Comedy, Drama, Romance	4.44 ± 2.10	3.21 ± 1.80	4.56 ± 2.15	1.28 ± 0.35	2.93 ± 1.90	86
24	Monsters University	Animation, Adventure, Comedy	5.55 ± 2.07	4.01 ± 2.12	6.15 ± 2.12	1.24 ± 0.23	4.44 ± 2.14	47
25	Kung Fu Panda 3	Animation, Action, Adventure	6.11 ± 2.17	4.46 ± 2.31	6.44 ± 2.37	1.25 ± 0.24	4.14 ± 2.11	47
26	Baby Driver	Action, Crime, Drama	5.29 ± 2.10	4.65 ± 2.27	5.19 ± 2.32	1.53 ± 0.88	5.30 ± 2.28	44
27	The Good Dinosaur	Animation, Adventure, Comedy	6.43 ± 2.15	4.63 ± 2.33	6.47 ± 2.18	1.19 ± 0.41	4.22 ± 2.07	43
28	About Time	Comedy, Drama, Fantasy	5.25 ± 2.60	4.29 ± 2.68	5.97 ± 2.32	1.22 ± 0.34	4.25 ± 2.24	43
29	Ordinary World	Comedy, Drama, Music	4.88 ± 1.72	3.93 ± 1.72	5.47 ± 1.68	1.22 ± 0.20	3.45 ± 1.98	49
30	Lion	Biography, Drama	5.36 ± 2.30	4.62 ± 2.65	5.01 ± 2.49	1.79 ± 1.25	3.93 ± 2.35	48
31	Shrek Forever After	Animation, Adventure, Comedy	5.87 ± 2.12	3.85 ± 2.37	6.29 ± 2.29	1.26 ± 0.29	4.35 ± 2.49	44
32	Chappie	Action, Crime, Drama	5.47 ± 2.26	4.43 ± 2.20	4.54 ± 2.45	1.69 ± 0.73	5.31 ± 2.31	47
33	Guardians of the Galaxy Vol. 2	Action, Adventure, Sci-Fi	6.15 ± 2.40	4.61 ± 2.34	5.85 ± 2.40	1.44 ± 1.01	5.56 ± 2.50	48
34	The Intern	Comedy, Drama	6.34 ± 2.02	5.06 ± 2.13	6.31 ± 1.98	1.23 ± 0.39	3.74 ± 2.34	42
35	La La Land	Comedy, Drama, Music	5.44 ± 2.24	4.09 ± 2.37	5.55 ± 2.29	1.49 ± 1.28	3.15 ± 1.99	47
36	Ice Age: Collision Course	Animation, Adventure, Comedy	6.38 ± 2.36	4.96 ± 2.40	6.92 ± 1.94	1.21 ± 0.26	4.87 ± 2.36	43
37	Frozen	Animation, Adventure, Comedy	5.88 ± 2.40	4.17 ± 2.38	6.31 ± 2.41	1.24 ± 0.40	4.35 ± 2.36	47
38	Transformers: The Last Knight	Action, Adventure, Sci-Fi	4.57 ± 2.06	4.18 ± 1.96	4.10 ± 2.33	1.91 ± 1.22	4.87 ± 1.97	42
39	Divergent	Adventure, Mystery, Sci-Fi	5.87 ± 1.81	4.87 ± 2.00	4.75 ± 2.27	1.95 ± 1.34	5.84 ± 2.05	49
40	Why Him?	Comedy	5.85 ± 2.24	4.60 ± 2.40	6.03 ± 2.30	1.25 ± 0.39	4.06 ± 2.44	43
41	The Boy	Horror, Mystery, Thriller	3.85 ± 2.09	4.92 ± 1.97	1.78 ± 1.06	5.21 ± 2.23	5.01 ± 2.16	82
42	Jigsaw	Crime, Horror, Mystery	4.04 ± 2.14	4.68 ± 2.15	2.07 ± 1.37	4.65 ± 2.18	4.91 ± 2.18	86
43	Shutter	Horror, Mystery, Thriller	3.53 ± 2.20	4.71 ± 2.25	1.68 ± 0.97	5.23 ± 2.34	4.63 ± 2.29	81
44	Ladda Land	Horror	4.61 ± 2.20	4.81 ± 2.23	1.95 ± 1.47	5.62 ± 2.13	4.88 ± 2.07	42
45	No One Lives	Horror, Thriller	4.20 ± 2.04	4.84 ± 2.28	1.94 ± 1.35	4.97 ± 2.56	5.07 ± 2.33	47
46	Tales from the Crypt	Horror	3.83 ± 2.22	4.41 ± 2.21	2.21 ± 1.88	4.67 ± 2.36	4.68 ± 2.41	44
47	Orphan	Horror, Mystery, Thriller	4.07 ± 2.35	5.18 ± 2.11	1.85 ± 1.40	5.11 ± 2.15	4.68 ± 2.27	40
48	Unfriended	Drama, Horror, Mystery	4.34 ± 2.57	5.40 ± 2.57	1.98 ± 1.93	5.34 ± 2.53	5.37 ± 2.55	43
49	Poltergeist	Horror, Thriller	4.13 ± 2.44	5.28 ± 2.55	1.91 ± 1.70	5.85 ± 2.36	5.20 ± 2.60	43
50	Jeruzalem	Horror	4.20 ± 2.12	4.82 ± 2.14	2.02 ± 1.42	5.00 ± 2.23	4.90 ± 2.27	49
51	Leatherface	Crime, Horror, Thriller	3.92 ± 2.11	4.77 ± 2.39	1.89 ± 1.27	4.93 ± 2.53	4.90 ± 2.50	47
52	The Babadook	Drama, Horror	3.62 ± 2.06	4.84 ± 2.19	1.67 ± 1.05	5.34 ± 2.11	4.74 ± 2.17	43
53	Oculus	Horror, Mystery	4.12 ± 2.11	5.81 ± 1.90	1.76 ± 1.29	6.07 ± 1.81	5.36 ± 2.28	42
54	The Witch	Horror, Mystery	3.69 ± 2.17	4.69 ± 2.44	1.89 ± 1.38	5.12 ± 2.40	4.54 ± 2.37	42
55	Trick 'r Treat	Comedy, Horror, Thriller	4.50 ± 2.27	5.59 ± 2.30	1.93 ± 1.42	5.68 ± 2.26	5.55 ± 2.34	42
56	The Woman in Black	Drama, Fantasy, Horror	4.23 ± 2.29	4.67 ± 2.27	1.94 ± 1.54	5.21 ± 2.33	4.79 ± 2.16	42
57	The Possession	Horror, Thriller	4.83 ± 2.47	5.32 ± 2.29	1.82 ± 1.44	5.90 ± 2.42	5.66 ± 2.38	42
58	Crimson Peak	Drama, Fantasy, Horror	3.93 ± 2.15	4.79 ± 2.33	2.33 ± 1.71	4.81 ± 2.43	5.10 ± 2.16	43
59	Program na winyan akat	Horror, Thriller	3.95 ± 2.11	4.98 ± 2.48	1.72 ± 1.20	5.52 ± 2.45	5.00 ± 2.26	43
60	The Pact	Horror, Mystery, Thriller	4.37 ± 2.00	5.56 ± 2.40	1.86 ± 1.42	6.23 ± 2.05	5.62 ± 2.50	44

TABLE V: Accuracy and F1 score for low-high classification

	Accuracy						F1 Score					
	Valence [1.79 : 1]	Arousal [2.03 : 1]	Happiness [2.41 : 1]	Fear [5.20 : 1]	Excitement [0.66 : 1]	Reward [0.60 : 1]	Valence	Arousal	Happiness	Fear	Excitement	Reward
EEG	0.6718	0.7286	**0.7158**	**0.8450**	**0.7054**	**0.6615**	0.5900	0.6523	**0.5760**	**0.6172**	**0.5848**	**0.5904**
E4	0.6563	0.6950	**0.7158**	**0.8450**	0.6848	0.6253	0.3922	0.4765	**0.5760**	**0.6172**	0.3974	0.3808
Fusion	**0.6821**	**0.7441**	0.7054	0.8424	0.6873	0.6512	**0.5990**	**0.6723**	0.5359	0.5610	0.5488	0.5847

Numbers inside brackets are class ratios [low : high].
Bold indicates the best results for the proposed datasets.

TABLE VI: Comparison accuracy and F1 scores for low-high classification with the other datasets.

Modality	Accuracy		F1 Score	
	Valence	Arousal	Valence	Arousal
EDOSE (EEG)	0.6718	0.7286	0.5900	0.6523
EDOSE (Empartica)	0.6563	0.6950	0.3922	0.4765
EDOSE (Fusion)	**0.6821**	**0.7441**	**0.5990**	**0.6723**
class ratio	1.79:1	2.03:1	1.79:1	2.03:1
EDOSE (EEG) :k-mean	0.6460	0.6977	0.6107	0.6193
EDOSE (Empartica) :k-mean	0.5943	0.6693	0.3770	0.4563
EDOSE (Fusion) :k-mean	**0.6718**	**0.7054**	**0.6369**	**0.6403**
class ratio	0.68:1	1.87:1	0.68:1	1.87:1
DREAMER (EEG) [7]	0.6249	0.6217	0.5184	0.5767
DREAMER (ECG) [7]	0.6237	0.6237	0.5305	0.5798
DREAMER (Fusion) [7]	0.6184	0.6232	0.5213	0.5750
DEAP (EEG) [6]	0.5760	0.6200	0.5630	0.5830
DEAP (Peripheral) [6]	0.6270	0.5700	0.6080	0.5330
MAHNOB-HCI (EEG) [8]	0.5700	0.5240	0.5600	0.4200
MAHNOB-HCI (Peripheral) [8]	0.4550	0.4620	0.3900	0.3800
DECAF (MEG) [14]	0.5900	0.6200	0.5500	0.5800
DECAF (Peripheral) [14]	0.6000	0.5500	0.5900	0.5400

Numbers at class ratio row are class ratios [low : high].
Bold indicates the best results for all datasets.

TABLE VII: Accuracy and F1 score form channel selection

Channel	Accuracy		F1 Score	
	Valence	Arousal	Valence	Arousal
FP_1, FP_2, F_z	0.6162	0.6767	0.3782	0.3989
FP_1, FP_2, C_z	0.6162	0.6558	0.3782	0.4681
FP_1, FP_2, P_z	0.6046	0.6674	0.5148	0.5087
FP_1, FP_2, O_z	0.6162	0.6767	0.3782	0.3989
T_3, T_4, F_z	**0.6813**	0.6767	**0.6421**	0.3989
T_3, T_4, C_z	0.6511	0.6767	0.6067	0.3989
T_3, T_4, P_z	0.6209	0.6767	0.5687	0.3989
T_3, T_4, O_z	0.6325	0.6767	0.5869	0.3989
F_z, C_z, P_z, O_z	0.6139	**0.6883**	0.5394	**0.5524**

Bold indicates the best results for our proposed datasets.

EEG, E4 and Fusion (EEG and E4) were used to train and test the model. Table V presents the mean accuracy and mean F1 scores of nine-fold. The condition with Fusion of EEG and E4 as the input features reached 68.21% and 74.41% of mean accuracy for V and A, respectively. However, input features consisting of either EEG or E4 provided better results than Fusion for H, F and E. In terms of F1 score, the results were consistent with the mean accuracy. Furthermore, EEG, E4 and Fusion as input features were compared for binary classification tasks. One-way repeated measures ANOVA revealed significant differences in the classification of V ($F(2, 16) = 48.726$, $p < 0.05$) and A ($F(1.196, 9.570) = 61.931$, $p < 0.05$) (low-high) levels. Pairwise comparison of E4 (A: 0.477 $\pm$ 0.032, V: 0.392$\pm$ 0.018) was significantly worse than both EEG (A: 0.652 $\pm$ 0.016, V: 0.590 $\pm$ 0.024) and Fusion (A: 0.672$\pm$ 0.013, V: 0.559 $\pm$ 0.025) in terms of F1 Score, $p < 0.05$.

C. Binary Classification Tasks with Labeling by k-mean Clustering

As shown in Table IV, V and A scores were widely distributed as their large standard deviation values. If the population is larger, its standard deviation value might be higher. Hence, the fixed threshold is not suitable. To solve this problem, k-mean clustering was proposed for labeling low-high levels of V and A, and binary classification tasks subsequently performed. Figure 6 shows the results following completion of k-mean clustering. We empirically labeled according to the VA model [4], with the Blue and Red group for low-A (LA), and the Green and Yellow group for high-A (HA). In terms of valence, the Red and Green group was selected for low-V (LV), and the Blue and Yellow group for high H (HV).

TABLE VIII: Accuracy and F1 score from varying EEG frequency bands

Frequency band	Accuracy		F1 Score	
	Valence	Arousal	Valence	Arousal
θ	0.5943	0.6563	0.3704	0.3919
α	0.5943	0.6563	0.3704	0.3919
β	0.5943	0.6563	0.3704	0.3919
γ	0.5659	0.6563	0.4510	0.3919
θ, α	0.5943	0.6563	0.3704	0.3919
θ, β	0.5943	0.6563	0.3704	0.3919
θ, γ	0.5659	0.6563	0.4537	0.3919
α, β	0.5943	0.6486	0.3704	0.4156
α, γ	0.6305	0.6486	0.5895	0.4962
β, γ	0.6227	0.6408	0.5757	0.4208
θ, α, β	0.5943	0.6486	0.3704	0.4156
θ, α, γ	0.6253	0.6460	0.5823	0.4943
θ, β, γ	0.6098	0.6408	0.5636	0.4208
α, β, γ	**0.6202**	**0.6718**	**0.5875**	**0.5693**

Bold indicates the best results for our proposed datasets.

Table VI presents the mean accuracy and mean nine-fold F1 scores. The condition with Fusion of EEG and E4 as the input features reached mean accuracy at 67.18% and at 70.54% for V and A, respectively. In terms of F1 score, Fusion features provided the best results for V and A (0.6369 and 0.6403, respectively). One-way repeated measures ANOVA revealed significant differences in the classification of V ($F(2, 16) = 26.871$, $p < 0.05$) and A ($F(2, 16) = 104.430$, $p < 0.05$) (low-high) levels. Pairwise comparison was performed, indicating that E4 (A: 0.377 $\pm$ 0.017, V: 0.456$\pm$ 0.026) was significantly worse than both EEG (A: 0.611 $\pm$ 0.029, V: 0.619 $\pm$ 0.024) and Fusion (A: 0.637$\pm$ 0.026, V: 0.640 $\pm$ 0.015) in terms of F1 Score, $p < 0.05$. Table VI also reports the V and A classification results from DREAMER [7], DEAP [6], MAHNOB-HCI [8] and DECAF [14] datasets using the same feature extraction and classification methods.

The results from channel selection are presented in Table VII for the mean accuracy and mean F1 nine-fold scores. In the low-high classification task for V, the condition with EEG channels from T_3, T_4, F_z as the input features reached mean accuracy at 68.13% and the mean F1-score at 0.6421. However, for A, the mean accuracy reached 68.83% and the mean F1-score reached 0.5524 using F_z, C_z, P_z, O_z. Finally, the results from varying EEG frequency bands as shown in Table VIII, present the mean accuracy and mean nine-fold F1 scores. Classification accuracy for V and A reached 62.02% and 67.18% using EEG features from α, β and γ bands. In terms of F1 score, the accuracy results were similar. The features from α, β and γ bands provided the best results in V and A classification tasks (0.5875 and 0.5693, respectively).

V. Discussion

We aim to discuss all the academic merits of this work for scientific purposes. One of the most important aspects of gathering an affective EEG with peripheral physiological signals is the use of an elicited emotion method. State-of-the-art works and the datasets in this study are based on audio-visual stimulation or movie clips. The strong point in this paper compared to previous works is that it involves a conservative clip selection from 200 participants of various ages (15–22 year olds) as shown in Table I. K-mean clustering is proposed as the method for ranking candidate clips. Finally, a list of effective clips is constructed for affective computing studies as shown in Table IV. These are all open to the public via YouTube, making it easy to use them for extended study.

Another strong point is an evaluation of the binary classification task for low-high levels of valence and arousal. In previous research, the threshold is empirically set on the self-emotional assessment scores for ground truth labeling [6]–[8], [14]. However, the list of selected clips in Table IV presents high standard deviations in the scores from the self-assessment. In a high population, simple thresholding might not work well since low or high levels of valence and arousal from different people are likely to vary. Thus, a simple mathematical based algorithm; k-mean clustering, is proposed for labeling the ground truth instead. As reported in Table VI,no statistical difference was found between two methods of ground truth labeling. Hence, labeling by k-mean clustering might work well in future work with a larger population or number of participants. Moreover, the proposed datasets or EDOSE outperformed the state-of-the-art datasets (both in accuracy and F Score) as shown in Table VI, including one involving magnetoencephalography (MEG) which is a functional brain mapping method (much more expensive than EEG).

Table VII and Table VIII show the results of the study of EEG factors, involving electrode channel selection for the future development of user-friendly devices for emotion recognition and frequency selection to obtain more effective EEG frequency bands as classification features, respectively. According to the aforementioned tables, T_3, T_4, F_z achieved the best results for valence classification both in terms of accuracy and F1 score, while the middle line (F_z, C_z, P_z, O_z) was promising for further improvement of an emotion recognition algorithm in both valence and arousal for low-high classification. Taking both results together, either T_3, T_4, F_z and F_z, C_z, P_z, O_z with all EEG frequency bands as input features is the best path for developing practical, usable low-cost devices for mental state recognition, especially concerning valence and arousal.

Finally, EDOSE is the first affective or emotion dataset involving a low-cost, open source device (OpenBCI). In comparison to the high-end EEG amplifier with a greater number of electrodes and sampling frequencies, OpenBCI demonstrates the capability to perform emotion recognition tasks. A low-cost, open source device is a game changer for encouraging developers or researchers to improve emotion recognition or classification algorithms with EDOSE. Then they can step further for online applications because the device is inexpensive so many people can afford to have it. From the academic viewpoint, the deep learning algorithm can be developed to attack the classification issue using EDOSE for training and testing data. There are two related works on deep learning with EEG in our group which can be adapted and incorporated into EDOSE datasets for further development of related engineering applications [15], [16]. Moreover, emotion recognition using peripheral physiological data from a real-time bracelet sensor or E4 remains a challenge. The E4 is a simple wearable device and resembles a watch so is very useful as a long-term affective state or mental health monitoring system. E4 data is also included in the EDOSE dataset.

VI. Conclusion

In summary, the researchers introduced the EDOSE dataset, along with the related experimental procedures. The dataset includes EEG and peripheral physiological signals from 43 participants while watching the elicited trailer clips. Subsequently, the participants self-assessment emotional scores were measured according to the scales of valence, arousal, happiness, fear, and excitement. Here, the researchers aimed to propose capabilities using a low-cost, open source EEG amplifier with sparse electrode channels for emotion recognition. To evaluate EDOSE, binary classification tasks (low-high valence or arousal) were conducted. The k-mean clustering algorithm was also applied to create ground truth labels from low and high clusters for both valence and arousal. On the other hand, previous works performed labeling by empirical thresholding. The results outperformed state-of-the-art studies. After publication of this paper, EDOSE will be available (upon requests) to other researchers for further development and studies in various fields.

References

[1] "Open source brain-computer interfaces," http://openbci.com/.

[2] S. Moon and J. Lee, "Implicit analysis of perceptual multimedia experience based on physiological response: A review," *IEEE Transactions on Multimedia*, vol. 19, no. 2, pp. 340–353, Feb 2017.

[3] Y. Li, D. Cao, L. Wei, Y. Tang, and J. Wang, "Abnormal functional connectivity of eeg gamma band in patients with depression during emotional face processing," *Clinical Neurophysiology*, vol. 126, no. 11, pp. 2078 – 2089, 2015.

[4] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[5] J. D. Morris, "Observations : Sam : The self-assessment manikin an efficient cross-cultural measurement of emotional response," 2000.

[6] 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*, vol. 3, no. 1, pp. 18–31, 2012.

[7] S. Katsigiannis and N. Ramzan, "Dreamer: a database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2018.

[8] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, Jan 2012.

[9] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, pp. 1–13, 2018.

[10] "The e4 wristband is a wearable research device that offers real-time physiological data acquisition and software for in-depth analysis and visualization." https://www.empatica.com/en-eu/research/e4/.

[11] C. A. Gabert-Quillen, E. E. Bartolini, B. T. Abravanel, and C. A. Sanislow, "Ratings for emotion film clips," *Behavior Research Methods*, vol. 47, no. 3, pp. 773–787, Sep 2015. [Online]. Available: https://doi.org/10.3758/s13428-014-0500-0

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hmlinen, "Meg and eeg data analysis with mne-python," *Frontiers in Neuroscience*, vol. 7, p. 267, 2013. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2013.00267

[14] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "Decaf: Meg-based multimodal database for decoding affective physiological responses," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, 2015.

[15] T. Wilaiprasitporn, A. Ditthapron, K. Matchaparn, T. Tongbuasirilai, N. Banluesombatkul, and E. Chuangsuwanich, "Affective eeg-based person identification using the deep learning approach," *arXiv preprint arXiv:1807.03147*, 2018.

[16] A. Ditthapron, N. Banluesombatkul, S. Ketrat, E. Chuangsuwanich, and T. Wilaiprasitporn, "Universal joint feature extraction for p300 eeg classification using semi-supervised autoencoder," *arXiv preprint arXiv:1808.06541*, 2018.