An EEG Dataset for Stable Affective Feature Selection

Zirui Lan [12], Yisi Liu [2*], Olga Sourina [12], Lipo Wang [3], Reinhold Scherer [4], Gernot Müller-Putz [5]

[1] *Nanyang Technological University, Singapore*
[2] *Fraunhofer Singapore, Singapore*
[3] *School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore*
[4] *School of Computer Science and Electronic Engineering, University of Essex, UK*
[5] *Institute of Neural Engineering, Graz University of Technology, Graz, Austria*

ARTICLE INFO

Keywords:
EEG dataset
Emotion Recognition
Affective features
Stable feature selection
Long-term aBCI performance

ABSTRACT

Affective brain-computer interface (aBCI) is a direct communication pathway between human brain and computer, via which the computer tries to recognize the affective states of its user and respond accordingly. As aBCI introduces personal affective factors into human-computer interactions, it could potentially enrich the user's experience during the interaction with a computer. Successful emotion recognition plays a key role in such a system. The state-of-the-art aBCIs leverage machine learning techniques which consist in acquiring affective electroencephalogram (EEG) signals from the user and calibrating the classifier to the affective patterns of the user. Many studies have reported satisfactory recognition accuracy using this paradigm. However, affective neural patterns are volatile over time even within the same subject. The recognition accuracy cannot be maintained if the usage of aBCI prolongs without recalibration. Existing studies have overlooked the performance evaluation of aBCI during long-term use. In this paper, we propose a dataset which includes multiple recording sessions spanning across several days for each subject. Multiple sessions across different days were recorded so that the long-term recognition performance of aBCI can be evaluated. Based on this dataset, we demonstrate that the recognition accuracy of aBCIs deteriorates when re-calibration is ruled out during the long-term usage. Then, we propose a stable feature selection method to choose the most stable affective features, for mitigating the accuracy deterioration to a lesser extent and maximizing the aBCI performance in the long run. We invite other researchers to test the performance of their aBCI algorithms on this dataset, and especially to evaluate the long-term performance of their algorithms.

1 Introduction

Emotions are a crucial element in our everyday communication. Though intuitive to human, it remains a challenging task for a computer to perceive the emotions of its user. Affective computing, as an emerging research topic that seeks to develop emotion-aware systems to recognize, interpret and process human emotion, has received increasing attention in recent years. Early works have focused on analyzing the physiological response to recognize emotions, such as heart rate [1], skin conductance [2], etc. These physiological reactions are regulated by the autonomic nervous systems under the influence of emotions, hence the possibility to interpret emotions by measuring such response. More recent studies have targeted the brain's role in perceiving and regulating emotions [3], giving rise to the affective brain-computer interface (aBCI). An electroencephalogram (EEG)-based aBCI is a direct communication pathway between human brain and computer by means of spontaneous EEG signals, bypassing the conventional pathways of peripheral nerves and muscles. Such an affective interface could potentially enrich the user's experience during the interaction with a computer if the computer is empowered to feel and respond to human emotions. In applications, an aBCI operates in such a paradigm that forms a loop as diagrammed in Fig. 1. In this paradigm, there are notably three core parts: signal acquisition, signal classification, and feedback to the user. The user generates EEG signals, which are captured by the EEG device. The EEG signals are then analysed and classified, the output of which are fed into an application which executes subroutines according to the recognized emotions. Feedback is then given to the user. Successful emotion recognition plays a key role in aBCI as it highly affects the quality of such an interface. The state-of-the-art aBCI leverages machine learning techniques which consist in acquiring affective EEG signals from the user and calibrating the classifier to the affective patterns of the user. Many studies about aBCI have reported satisfactory recognition accuracy using this paradigm [4-14]. In these studies, affective EEG data were collected within a relatively short period, and *k*-fold cross-validations were carried out to evaluate the recognition accuracy. In a *k*-fold evaluation,

* Corresponding author

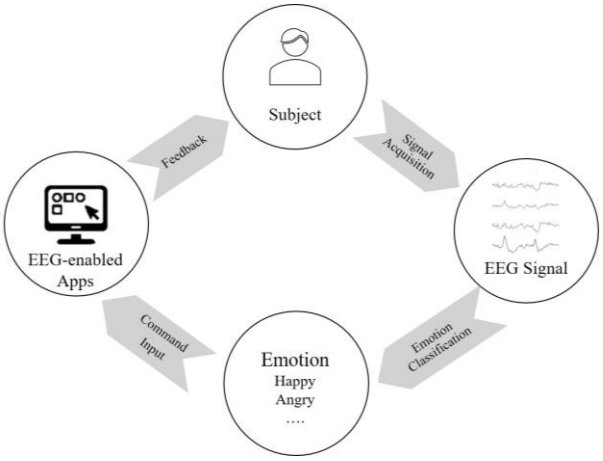

Fig. 1 A general affective brain-computer interface (aBCI) paradigm.

the EEG data are segmented into *k* nonoverlapping sections: *k*-1 folds are used to train the classifier, and the remaining fold is used to test the recognition accuracy. However, due to the volatility of affective neural pattern, the recognition accuracy cannot be maintained if the usage of aBCI prolongs without re-calibrating the classifier. The recognition accuracy assessed by cross-validating short-term EEG data is over-optimistic and can hardly represent the system performance in the long run. On the other hand, there is little study on the long-term recognition performance of aBCI, which may partly be due to the fact that few existing affective EEG datasets contain recordings over a long course of time.

We devote this paper to presenting an EEG dataset that contains multiple recordings on the same day and different days of the same subjects, and to the investigation of aBCI performance over a long course of time. As the (re-)calibration process may be time-consuming, tedious and laborious, we are motivated to mitigate the burden of frequent re-calibrations on the user of interest. Ideally, a stable affective EEG feature should give consistent measurement of the same emotion on the same subject over a long course of time. We presented a pilot study on the stability of affective EEG features in [15, 16], where we hypothesize that using stable EEG features may improve the long term recognition accuracy, while unstable features may worsen the recognition performance of the BCI in the long run. In [17], we propose a stable feature selection method to choose the optimal set of stable features that maximize the recognition accuracy of the system in the long run. In this paper, we aim at introducing the dataset used in our previous study [17], and make it available to the public[1]. We invite other researchers to test the performance of their aBCI algorithms on this dataset, and especially to evaluate the long-term performance of their algorithms.

This paper is organized as follows. Section 2 reviews the existing affective EEG datasets. Section 3 documents our data collection procedures. Section 4 introduces our proposed stable feature selection method. Section 5 elaborates on the simulations to evaluate the short-term and long-term performance of aBCI. Section 6 presents the results with discussions. Section 7 concludes this paper.

2 Review of existing affective EEG datasets

There are a few affective datasets available that contain EEG recordings. The enterface (2006, [18]) dataset includes the EEG and functional near infrared spectroscopy (fNIRS) recorded from 5 subjects. They adopted the pictorial affective stimuli from the International Affective Picture System (IAPS) to induce 3 emotions (calm, positive exiting, and negative exciting) on the subjects. The EEG signals were captured by a Biosemi Active II device with 54 effective EEG channels at a sampling rate of 1024 Hz. The MAHNOD HCI (2012, [19]) dataset provides the EEG recordings along with other physiological signals carried out on 27 subjects. Emotional video clips extracted from movies and online repositories were used as affective stimuli to elicit 6 emotions (disgust, amusement, joy, fear, sadness, and neutral). A 32-channel Biosemi Active II device was used to record the EEG signals. The DEAP (2012, [20]) dataset consists of the EEG and other peripheral physiological signals collected from 32 subjects using the Biosemi Active II device. Forty 1-minute long music videos were chosen as affective stimuli. After the exposure to each emotional stimulus, the subject was required to provide feedback on his/her truly felt emotion in the form of Self-Assessment Manikin (SAM) questionnaire [21]. The SAM feedback was regarded as the ground truth as to what emotion has been elicited on the subject. In these three datasets, the emotion elicitation experiment and EEG data collection were carried out on each subject within 1-2 hours in one day. No repeated elicitation experiment or EEG data collection is made on the same subject on different days. That is to say, the affective EEG data were collected within a relatively short period of time for each subject and therefore, these datasets are not suitable for the evaluation of the long-term classification performance on aBCIs. The SEED (2015, [22]) dataset is the first dataset that provides repeated affective EEG recordings on the same subject on different days. The SEED dataset comprises the EEG recordings from 15 subjects for 3 emotions (positive, neutral, and negative). Fifteen Chinese movie excerpts were selected as affective stimuli in the emotion induction experiment. The EEG signals were collected by an ESI NeuroScan system equipped with 64 channels. The emotion induction experiment and EEG data collection were repeated on each subject three times on three different days. Hence, this dataset makes possible the evaluation of long-term performance of aBCI.

Our dataset introduced in this paper complements the abovementioned existing datasets in two folds. Firstly, the existing datasets [18-20, 22] were collected using specialized, costly EEG devices such as Biosemi Active II (in [18-20]) and ESI NeuroScan (in [22]). Although these systems may provide better signal quality, they are bulky and not quite suitable for casual usage in everyday applications. In our dataset, we opt for a low-cost, portable consumer-grade EEG headset, which better simulates the application scenario an average user would encounter in everyday application. Secondly, the SEED dataset included 3 repeated measurements of the same induced affective states on 3 different days. In our dataset, we extend the repeated measurements to 16 times in a course of 8 days. We carry out two repetitions per day and thus, our dataset provides not only repeated recordings of the same induced affective states across

[1] www.ntu.edu.sg/home/lanz/download

different days, but also on the same days. In the next section, we elaborate on the experiment procedures for our data collection.

3 Data collection

3.1 Affective Stimuli Selection

The selection of affective stimuli plays a role in successful emotion elicitation. We select audio stimuli with known affective attributes from the International Affective Digitized Sounds (IADS, [23]) library. IADS is an established affective stimuli library that provides normative emotion stimuli for emotion induction experiment. IADS contains a collection of 167 sound clips, each lasting for 6 seconds. The affective attribute of each sound clip has been rated by and averaged over a pool of 100 subjects in terms of valence, arousal and dominance in accordance with Russel's 3D emotional model [24] on a scale of 1-9. By using Russel's 3D emotional model, emotions boil down to and are quantified by three orthogonal dimensions. The valence (V) dimension measures how pleasant an emotion is, ranging from unpleasant to pleasant. For example, both frightened and sad are unpleasant emotions and rated low in valence, whereas happy and surprised are pleasant emotions that score high in valence. Likewise, the arousal (A) dimension quantifies how intense an emotion is, ranging from inactive to active. For instance, sad is a lowly activated emotion whereas frightened is a highly activated emotion. The dominance (D) dimension reveals the dominating power associated with an emotion, ranging from submissive (lack of control) to dominating (in control of everything). When a person feels frightened, he/she lacks control of the surroundings and feels submissive. When a person feels angry, he/she stands in a dominating position, tends to aggress and is in high dominance level. If we consider each dimension to be binary – either high (H) or low(L) – then Russel's 3D emotional model identifies a total of 8 emotions: HVHAHD, HVHALD, HVLAHD, HVLALD, LVHAHD, LVHALD, LVLAHD, and LVLALD. Out of the eight emotions, we intend to induce the four emotions that are common in everyday life: HVLAHD (pleasant), HVHAHD (happy), LVHALD (frightened), and LVHAHD (angry).

To find stimuli that induce the four desired emotions in IADS, we consider rating equal to 5 as a threshold. Rating lower than 5 is considered low while that larger than 5 is considered high. We then select ten stimuli from IADS for each emotion class, as is tabulated in Table 1. For instance, the stimuli to induce pleasant emotion include those whose valence rating is larger than 5, arousal rating smaller than 5, and dominance rating larger than 5. Likewise, the same threshold applies to the other emotions except angry, where there are not enough ten stimuli with dominance rated higher than 5, and we marginally lower the threshold to allow dominance rated higher than 4 to be selected.

3.2 Data Collection Protocol

The data collection was carried out in a laboratory environment with controlled illumination. The EEG data were recorded with an Emotiv EPOC headset on the project PC. The Emotiv EPOC headset is a lightweight, portable and wireless EEG device. Specifically, the Emotiv EPOC was chosen because it is more likely to be used by the general consumers in a casual, everyday application than the costly, research-grade but bulky EEG device. Despite being affordable, the signal quality of EEG data recorded with Emotiv EPOC has been rigorously examined and compared to that of the NeuroScan device, a research-grade EEG system, leading to the conclusion that Emotiv EPOC compares well with NeuroScan for the reliable auditory ERPs (Event Related Potentials) [25, 26]. Other seminal studies validating the result quality produced by Emotiv EPOC can be found in [27-30].

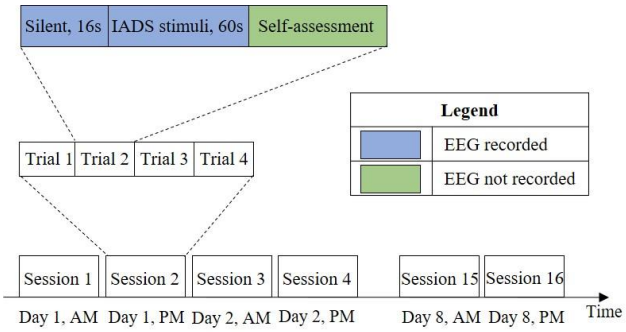

Fig. 2 Protocol of emotion induction experiment.

In existing datasets, e.g., enterface [18], MAHNOD HCI [19], and DEAP [20], EEG data were collected within a relatively short period in one single day for each subject. However, we stress that datasets with EEG recording limited to a relatively short time span are not enough for the evaluation of long-term aBCI performance. With this in mind, our data collection experiment was designed such that multiple EEG data recording sessions within the same day and across different days are carried out for each subject.

As shown in Fig. 2, for each subject, we carried out 16 recording sessions in a course of 8 days. Specifically, we conducted 2 recording sessions per day for each subject, one in the morning and the other in the afternoon. Each session consisted of four trials, each of which corresponded to one targeted induced emotion. The sequence of emotion induction was as such that trial 1 to 4 corresponded to pleasant, happy, frightened, and angry emotion, respectively. During each trial, the EEG recording started with a "tick" sound, following which a 16-second silent interval was given to the subject to get prepared for the stimuli exposure. After that, ten IADS stimuli were presented to the subject in the order shown in Table 1. The EEG recording of one trial lasted for 76 seconds. As soon as the stimuli presentation ended, the subject was required to fill out the self-assessment questionnaire, during which the EEG signals were not recorded. For the self-assessment, we adopted the modified Self-Assessment Manikin (SAM) questionnaire as was used in [20] for the DEAP dataset. Specifically, the subject needed to self-assess the emotional experience during stimuli exposure from these five dimensions on a scale of 1-9: the valence, arousal and dominance dimensions in line with Russel's 3D emotion model [24], plus the liking and familiarity dimensions. The valence scale ranges from unpleasant to pleasant. The arousal scale ranges from inactive to active. The dominance scale ranges from submissive to empowered. The liking scale ranges from disliking to liking, which is a personal

Table 1 Selected IADS stimuli for the emotion induction experiment. Valence: from unpleasant = 1 to pleasant = 9. Arousal: from inactive = 1 to active = 9. Dominance: from submissive = 1 to dominating = 9.

Targeted emotion	IADS Index	Stimulus description	Valence (mean ± std)	Arousal (mean ± std)	Dominance (mean ± std)
Pleasant	150	Seagull	6.95 ± 1.64	4.38 ± 1.64	5.91 ± 1.80
(HVLAHD)	151	Robin's chirping	7.12 ± 1.56	4.47 ± 1.56	5.73 ± 1.92
	171	Country night	5.59 ± 1.79	3.71 ± 1.79	5.52 ± 1.77
	172	Brook	6.62 ± 1.69	3.36 ± 1.69	6.21 ± 1.86
	377	Rain	5.84 ± 1.73	3.93 ± 1.73	5.70 ± 1.89
	809	Harp	7.44 ± 1.41	3.36 ± 1.41	6.29 ± 1.87
	810	Beethoven's music	7.51 ± 1.66	4.18 ± 1.66	6.07 ± 1.92
	812	Choir	6.90 ± 1.69	3.43 ± 1.69	5.69 ± 1.90
	206	Shower	6.20 ± 1.60	4.40 ± 1.60	5.62 ± 1.61
	270	Whistling	6.10 ± 1.83	4.23 ± 1.83	5.85 ± 1.93
		Mean	6.63 ± 1.66	3.95 ± 1.66	5.86 ± 1.85
Happy	109	Carousel	6.40 ± 2.13	5.64 ± 2.13	5.69 ± 1.93
(HVHAHD)	254	Video game	6.17 ± 1.65	5.58 ± 1.65	6.25 ± 2.05
	351	Applause	7.32 ± 1.62	5.55 ± 1.62	6.74 ± 1.71
	716	Slot machine	7.00 ± 2.17	6.44 ± 2.17	6.54 ± 2.03
	601	Colonial music	6.53 ± 1.66	5.84 ± 1.66	5.73 ± 1.58
	367	Casino 2	7.33 ± 1.74	6.72 ± 1.74	6.41 ± 1.98
	366	Casino 1	7.09 ± 1.73	6.26 ± 1.73	6.08 ± 2.19
	815	Rock & Roll music	7.90 ± 1.53	6.85 ± 1.53	6.86 ± 1.99
	817	Bongos	7.67 ± 1.46	7.15 ± 1.46	6.44 ± 1.73
	820	Funk music	6.94 ± 1.98	5.87 ± 1.98	5.97 ± 1.80
		Mean	7.04 ± 1.77	6.19 ± 1.77	6.27 ± 1.90
Frightened	275	Screaming	2.05 ± 1.62	8.16 ± 1.62	2.55 ± 2.01
(LVHALD)	276	Female screaming 2	1.93 ± 1.63	7.77 ± 1.63	2.69 ± 2.02
	277	Female screaming 3	1.63 ± 1.13	7.79 ± 1.13	2.32 ± 1.78
	279	Attack 1	1.68 ± 1.31	7.95 ± 1.31	2.30 ± 1.94
	284	Attack 3	2.01 ± 1.48	7.05 ± 1.48	2.99 ± 2.00
	285	Attack 2	1.80 ± 1.56	7.79 ± 1.56	2.41 ± 2.02
	286	Victim	1.68 ± 1.18	7.88 ± 1.18	2.31 ± 2.03
	290	Fight	1.65 ± 1.27	7.61 ± 1.27	2.89 ± 2.05
	292	Male screaming	1.99 ± 1.41	7.28 ± 1.41	2.82 ± 1.78
	422	Tire skids	2.22 ± 1.47	7.52 ± 1.47	2.62 ± 1.77
		Mean	1.86 ± 1.41	7.68 ± 1.41	2.59 ± 1.94
Angry	116	Buzzing	3.02 ± 1.65	6.51 ± 1.65	4.14 ± 2.11
(LVHAHD)	243	Couple sneeze	3.86 ± 1.70	5.19 ± 1.70	4.23 ± 1.90
	251	Nose blow	4.16 ± 2.02	5.14 ± 2.02	4.44 ± 1.89
	380	Jack hammer	3.70 ± 1.88	6.33 ± 1.88	4.18 ± 1.93
	410	Helicopter 2	4.86 ± 1.48	5.89 ± 1.48	4.59 ± 1.55
	423	Injury	3.31 ± 1.79	6.23 ± 1.79	4.22 ± 1.89
	702	Belch	4.45 ± 2.57	5.37 ± 2.57	5.23 ± 2.04
	706	War	4.16 ± 1.68	5.30 ± 1.68	4.55 ± 1.82
	729	Paper 2	4.30 ± 1.69	5.79 ± 1.69	5.33 ± 2.27
	910	Electricity	3.86 ± 1.83	6.18 ± 1.83	4.03 ± 1.84
		Mean	3.97 ± 1.83	5.79 ± 1.83	4.49 ± 1.92

preference of the subject and not to be confused with the valence dimension. The familiarity scale ranges from unfamiliar to familiar.

Six subjects participated in our data collection experiment (5 males and 1 female, aged 24-28). All subjects reported no history of mental diseases or head injuries. Prior to the commencement of the experiment, the procedure of the experiment, the use of self-assessment questionnaire and the meaning of each affective attribute (e.g., valence) have been well-explained to the subject both verbally and in writing. The experiment would proceed only if the subject expressed sufficient understanding of the affective attributes. Written consent was obtained from the subject before we proceed to data collection. During the experiment, the experimenter assisted the subject in setting up the EEG device. The start/stop recording was controlled by the experimenter. The subject was seated approximately 1 meter from the screen of the project PC and wearing a pair of earphones with volume properly adjusted. The subject was told to sit back and rest the arms on the armrests with minimum muscle movement to avoid contaminating the EEG signals. After each recording, the experimenter administered the digital questionnaire to the subject for the self-assessment. The subject completed the questionnaire on the same project PC, where the EEG recordings were saved together with the respective self-assessment responses.

3.3 Analysis of affective rating response

The self-assessment questionnaires collected from the subjects were analyzed to examine the effect of our emotion elicitation experiment. We first analyzed the variation of affective ratings across different sessions, where we computed the mean and standard deviation of the affective ratings collected from each subject across the sixteen sessions. As shown in Table 2, the standard deviations are mostly small (< 1) across ratings of different sessions. This suggests that the subjects have given consistent ratings in relation to each targeted emotion across

different sessions, which accounts for a low variation in feeling the same the emotion across different sessions. At first glance, the mean of ratings in Table 2 are trending similarly as the ground truth of the stimuli used in Table 1 for the respective emotion. We further validate this by computing the Pearson correlation coefficients between the subject's self-assessment ratings and the ground truth affective ratings in all sessions, as shown in Table 3. The results show significant positive correlation between subject's affective ratings and the ground truth for the respective affective attributes on all subjects ($p < 0.05/18$, where the significance level 0.05 is tightened and divided by the number of comparisons to compensate for multiple comparisons). The significant correlation between subject's self-assessment ratings and the ground truth suggests that the subject's feelings are largely in agreement with what the affective stimuli intend to elicit. It can be reasonably assumed that the subject felt the targeted emotions during the emotion elicitation experiment.

Additionally, we computed the pairwise Pearson correlation coefficients among valence, arousal, dominance, liking and familiarity of the subject's self-assessment ratings. The correlations are mostly insignificant between arousal and valence and between arousal and dominance. This suggests that the subjects are able to differentiate the two attributes well. It is worth mentioning that the correlation is significant between valence and liking and between dominance and liking on all subjects ($p < 0.05/60$), as presented in Table 4. Without implication of any causality, the direct correlation between valence and liking implies that a subject tends to like pleasant stimuli, and dislike unpleasant stimuli. Likewise, the direct correlation between dominance and liking implies that a subject tends to like the stimuli that make the subject feel dominating, and dislike the stimuli that make the subject feel submissive.

Table 2 Mean ± std of subject's self-assessment ratings across sixteen sessions. Valence: from unpleasant = 1 to pleasant = 9. Arousal: from inactive = 1 to active = 9. Dominance: from submissive = 1 to dominating = 9.

Subject	Targeted Emotion	Valence	Arousal	Dominance
1	Pleasant	7.81 ± 0.40	2.56 ± 0.51	6.75 ± 0.45
	Happy	7.63 ± 0.50	6.88 ± 1.02	6.38 ± 0.62
	Frightened	2.06 ± 0.25	6.75 ± 0.68	3.31 ± 0.48
	Angry	3.31 ± 0.48	5.81 ± 0.66	4.06 ± 0.44
2	Pleasant	7.69 ± 0.87	3.56 ± 1.03	6.81 ± 0.66
	Happy	8.56 ± 0.51	8.81 ± 0.40	7.38 ± 0.50
	Frightened	1.06 ± 0.25	7.00 ± 0.63	2.25 ± 0.45
	Angry	1.38 ± 0.62	3.44 ± 0.51	3.81 ± 0.40
3	Pleasant	6.44 ± 0.63	2.38 ± 0.81	7.00 ± 0.63
	Happy	6.19 ± 0.40	6.19 ± 0.40	6.25 ± 0.45
	Frightened	3.56 ± 0.63	6.25 ± 0.45	3.56 ± 0.73
	Angry	3.63 ± 0.50	6.19 ± 0.40	6.44 ± 0.51
4	Pleasant	5.44 ± 0.73	4.00 ± 1.10	6.25 ± 0.77
	Happy	6.88 ± 0.72	6.50 ± 1.21	7.13 ± 0.89
	Frightened	3.19 ± 0.40	6.81 ± 0.83	3.06 ± 0.44
	Angry	3.38 ± 0.50	6.38 ± 0.50	6.75 ± 0.45
5	Pleasant	7.25 ± 0.45	3.38 ± 1.15	6.38 ± 1.36
	Happy	7.75 ± 0.45	7.31 ± 0.48	7.56 ± 0.51
	Frightened	2.69 ± 0.48	6.69 ± 1.01	3.19 ± 1.22
	Angry	2.75 ± 1.34	7.06 ± 1.48	4.19 ± 1.42
6	Pleasant	5.63 ± 0.62	3.00 ± 1.15	7.00 ± 0.37
	Happy	6.63 ± 0.62	6.25 ± 0.58	7.00 ± 0.37
	Frightened	3.06 ± 0.44	6.88 ± 0.34	3.06 ± 0.25
	Angry	3.44 ± 0.51	6.38 ± 0.81	3.69 ± 1.01

Table 3 Pearson correlation coefficients between subject's self-assessment rating and IADS ground truth ratings.

Subject	Valence	Arousal	Dominance
1	0.9655	0.8288	0.8881
2	0.9324	0.6009	0.9498
3	0.8679	0.8162	0.8037
4	0.8559	0.7069	0.8206
5	0.8949	0.6494	0.7937
6	0.8935	0.8189	0.8870

Table 4 Pearson correlation coefficients between valence and liking and between dominance and liking of subject's self-assessment ratings.

Subject	Valence-Liking	Dominance-Liking
1	0.9681	0.9203
2	0.9635	0.9076
3	0.9076	0.5494
4	0.8426	0.4660
5	0.9477	0.7790
6	0.9446	0.8969

4 Proposed stable feature selection methods

In this section, we describe the approaches to our proposed feature selection algorithm. We firstly review EEG feature extraction methods in 4.1. Then, we introduce an ANOVA-based stability measurement model called Intra-class Correlation Coefficient (ICC) in 4.2. Our proposed feature selection algorithm is presented in 4.3.

4.1 Feature Extraction

4.1.1 Fractal Dimension

Let $\boldsymbol{x} \in \mathbb{R}^n$ denote a column vector of n EEG time series samples (raw signals) from one channel. Construct k new time series by re-sampling $\boldsymbol{x}$ as follows.

$$\boldsymbol{x}_k^m = \left[\boldsymbol{x}(m), \boldsymbol{x}(m+k), \dots, \boldsymbol{x}\left(m + \left\lfloor\frac{n-m}{k}\right\rfloor k\right)\right]^{\mathrm{T}}, m = 1, 2, \dots, k, \quad (1)$$

where $\lfloor\cdot\rfloor$ denotes the floor function, m the initial time series sample and k the interval. We compute the length of the curve for each new series as follows.

$$l_k^m = \frac{1}{k}\left\{\left(\sum_{i=1}^{\left\lfloor\frac{n-m}{k}\right\rfloor}|\boldsymbol{x}(m+ik) - \boldsymbol{x}(m+(i-1)k)|\right)\right\}\left(\frac{n-1}{\left\lfloor\frac{n-m}{k}\right\rfloor k}\right), \quad (2)$$

Let l_k denote the mean of l_k^m for $m = 1, 2, \dots k$, the fractal dimension of time series $\boldsymbol{x}$ is computed as [31]

$$FD = -\lim_{k\to\infty}\frac{\log(l_k)}{\log(k)}, \quad (3)$$

Apparently, in numerical evaluation, it is not possible for k to be infinite. It has proven [32, 33] that the computed fractal value approximates the true, theoretical fractal value reasonably well given a reasonably large k. Based on the study in [33], k = 32 yields a good balance between accuracy and computational resources required. In this study, we follow the same parameter setting.

4.1.2 Statistics

A set of six statistical features were adopted in [34] for EEG-based emotion recognition, which, in combination with the fractal dimension feature, have been demonstrated to improve the classification accuracy [34]. Six statistical features are computed as follows.

Mean of the raw signals:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}(i), \tag{4}$$

Standard deviation of the raw signals:

$$\sigma_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}(i) - \mu_x)^2}, \tag{5}$$

Mean of the absolute values of the first order difference of the raw signals:

$$\delta_x = \frac{1}{n-1}\sum_{i=1}^{n-1}|\boldsymbol{x}(i+1) - \boldsymbol{x}(i)|, \tag{6}$$

Mean of the absolute values of the first order difference of the normalized signals:

$$\tilde{\delta}_x = \frac{1}{n-1}\sum_{i=1}^{n-1}|\tilde{\boldsymbol{x}}(i+1) - \tilde{\boldsymbol{x}}(i)| = \frac{\delta_x}{\sigma_x}, \tag{7}$$

Mean of the absolute values of the second order difference of the raw signals:

$$\gamma_x = \frac{1}{n-2}\sum_{i=1}^{n-2}|\boldsymbol{x}(i+2) - \boldsymbol{x}(\mathrm{i})|, \tag{8}$$

Mean of the absolute values of the second order difference of the normalized signals:

$$\tilde{\gamma}_x = \frac{1}{n-2}\sum_{i=1}^{n-2}|\tilde{\boldsymbol{x}}(i+2) - \tilde{\boldsymbol{x}}(i)| = \frac{\gamma_x}{\sigma_x}. \tag{9}$$

In (4)–(9), $\tilde{\boldsymbol{x}}$ denotes the normalized (zero mean, unit variance) signals, i.e., $\tilde{\boldsymbol{x}} = (\boldsymbol{x} - \mu_x)/\sigma_x$.

4.1.3 Spectral Band Power

Spectral band power, or simply "power", is one of the most extensively used features in EEG-related research [4, 6, 10, 12, 14]. In EEG study, there is common agreement on partitioning the EEG power spectrum into several sub-bands (though the frequency range may slightly differ from case to case): alpha band, theta band, beta band etc. In our study, the EEG power features from theta band (4 – 8 Hz), alpha band (8 – 12 Hz), and beta band (12 – 30 Hz) are computed.

The power features are obtained by first computing the Fourier Transform on the EEG signals. The discrete Fourier Transform transforms a time-series $\boldsymbol{x} = [\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(N)]^{\mathsf{T}}$ to another series $\boldsymbol{s} = [\boldsymbol{s}(1), \boldsymbol{s}(2), \ldots, \boldsymbol{s}(N)]^{\mathsf{T}}$ in a frequency domain. $\boldsymbol{s}$ is computed as

$$\boldsymbol{s}(k) = \sum_{n=0}^{N-1} \boldsymbol{x}(n) e^{-\frac{j2\pi kn}{N}}, \tag{10}$$

where N is the number of sampling points. Then, the power spectrum density is computed as

$$\hat{\boldsymbol{s}}(k) = \frac{1}{N}|\boldsymbol{s}(k)|^2, \tag{11}$$

Lastly, the spectral band power features are computed by averaging the power spectrum density $\hat{\boldsymbol{s}}(k)$ over the targeted sub-band. E.g., the alpha band power is computed by averaging $\hat{\boldsymbol{s}}(k)$ over 8 – 12 Hz.

4.1.4 Higher Order Crossing

Higher Order Crossings (HOC) was proposed in [35] to capture the oscillatory pattern of EEG, and used in [34, 36-38] as features to recognize human emotion from EEG signals. The HOC is computed by first zero-meaning the time-series $\boldsymbol{x}$ as

$$\boldsymbol{z}(i) = \boldsymbol{x}(i) - \mu_x, \tag{12}$$

where $\boldsymbol{z}$ is the zero-meaned series of $\boldsymbol{x}$ and μ_x the mean of $\boldsymbol{x}$ computed as per (4). Then, a sequence of filter ∇ is successively applied to $\boldsymbol{z}$, where ∇ is the backward difference operator, $\nabla \equiv \boldsymbol{z}(i) - \boldsymbol{z}(i-1)$. Denote the kth-order filtered sequence of $\boldsymbol{z}$ as $\boldsymbol{\xi}_k(\boldsymbol{z})$, $\boldsymbol{\xi}_k(\boldsymbol{z})$ is obtained by iteratively applying ∇ on $\boldsymbol{z}$, as

$$\boldsymbol{\xi}_k(\boldsymbol{z}) = \nabla^{k-1}\boldsymbol{z}, \nabla^0\boldsymbol{z} = \boldsymbol{z}. \tag{13}$$

Then, as its name suggests, the feature consists in counting the number of zero-crossing, which is equivalent to the times of sign changes, in sequence $\boldsymbol{\xi}_k(\boldsymbol{z})$. We follow [34] and compute the HOC feature of order k = 1, 2, 3, …, 36.

4.1.5 Signal Energy

The signal energy is the sum of squared amplitude of the time-series signal [39], computed as

$$\varepsilon = \sum_i |\boldsymbol{x}(i)|^2. \tag{14}$$

4.1.6 Hjorth Feature

Hjorth [40] proposed three features of a time-series, which have been used as affective EEG features in [41, 42].

Activity:

$$a(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}(i) - \mu_x)^2, \tag{15}$$

where μ_x is the mean of $\boldsymbol{x}$ computed as per (4).

Mobility:

$$m(\boldsymbol{x}) = \sqrt{\frac{\mathrm{var}(\dot{\boldsymbol{x}})}{\mathrm{var}(\boldsymbol{x})}}, \tag{16}$$

where $\dot{\boldsymbol{x}}$ is the time derivative of the time-series $\boldsymbol{x}$, and var(·) is the variance operator.

Complexity:

$$c(\boldsymbol{x}) = \frac{m(\dot{\boldsymbol{x}})}{m(\boldsymbol{x})}, \tag{17}$$

which is the mobility of the time derivative of $\boldsymbol{x}$ over the mobility of $\boldsymbol{x}$.

4.2 Feature Stability Measurement

The stability of feature parameters was quantified by the Intraclass Correlation Coefficient (ICC). ICC allows for the assessment of similarity in grouped data. It describes how well the data from the same group resemble each other. ICC was often used in EEG stability study [43, 44]. ICC is derived from a one-way ANOVA model and defined as [45]

$$\mathrm{ICC} = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W}, \tag{18}$$

where MS_B, MS_W and k denote the mean square error between groups, the mean square error within group, and the number of samples in each group, respectively. A larger ICC value indicates higher similarity among group data. ICC tends to one

Table 5 The analysis of variance table

Treatment (emotion)	**Measurement**				**Total**	**Average**
1	x_{11}	x_{12}	$\cdots$	x_{1k}	$x_{1\cdot}$	$\bar{x}_{1\cdot}$
2	x_{21}	x_{22}	$\cdots$	x_{2k}	$x_{2\cdot}$	$\bar{x}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$x_{n\cdot}$	$\bar{x}_{n\cdot}$
					$x_{\cdot\cdot}$	$\bar{x}_{\cdot\cdot}$
Source of variance	***Sum of squares***				***Degree of freedom***	***Mean square***
Between treatment	$SS_B = k\sum_{i=1}^{n}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$				$n-1$	$MS_B = SS_B/(n-1)$
Within treatment	$SS_W = SS_T - SS_B$				$nk-n$	$MS_W = SS_E/(nk-n)$
Total	$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - \bar{x}_{\cdot\cdot})^2$				$nk-1$	

when there is absolute agreement among the grouped data, i.e., $MS_W = 0$. A smaller ICC value suggests a lower similarity level. ICC value can drop below zero in the case when MS_W is larger than MS_B, accounting for dissimilarity among the grouped data.

4.3 Stable Feature Selection

A stable affective EEG feature should give consistent measurements of the same emotion on the same subject over the course of time, therefore there is the possibility to reduce the need of re-calibration by using the stable features. To this end, we propose a stable feature selection method based on ICC score ranking. The proposed method consists of three steps: ICC assessment, ICC score ranking, and iterative feature selection.

We assess the long-term stability of different EEG features with ICC. Let $\boldsymbol{X}$ be the matrix of feature parameters of a specific feature, rows of $\boldsymbol{X}$ correspond to different emotions, and columns of $\boldsymbol{X}$ correspond to different repeated measurements over the course of time. Intuitively, we want the feature parameters to be consistent when measuring the same emotion repeatedly over the course of time. Therefore, we want the parameters within the same row to be similar to each other. Moreover, we want the parameters measuring different affective states to be discriminative, so that different affective states are distinguishable. Therefore, we want different rows to be dissimilar to each other. The ICC measurement takes both considerations into account. The ICC is computed as per (18), which is based on ANOVA. For clarity, we display $\boldsymbol{X}$ in the ANOVA table as in Table 5. In Table 5, we refer treatment to different emotions induced by specific affective stimuli. x_{ij} is the feature parameter of the j-th measurement of emotion i. $x_{i\cdot}$ is the sum of all measurements of emotion i, $x_{i\cdot} = \sum_{j=1}^{k} x_{ij}$. $\bar{x}_{i\cdot}$ is the average of all measurements of emotion i, $\bar{x}_{i\cdot} = (1/k)\sum_{j=1}^{k} x_{ij}$. $x_{\cdot\cdot}$ is the sum of all measurements over all emotions, $x_{\cdot\cdot} = \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}$. $\bar{x}_{\cdot\cdot}$ is the average of all measurements over all emotions, $\bar{x}_{\cdot\cdot} = (1/nk)\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}$.

We can obtain the stability score of each feature by computing the ICCs, thereafter, we rank the feature according to the stability score in descending order. Features with higher ICC are more stable over the course of time, and exhibit better discriminability among different emotions. Our proposed feature selection method consists in iteratively selecting the top stable features and validating the inter-session emotion recognition accuracy. The feature subset that yields the best accuracy is retained.

Table 6 Referenced state-of-the-art affective EEG features

Feature (dimension, abbreviation)	**Reference**
6 statistics (30, STAT)	[12, 13, 34, 49, 50]
36 higher order crossings (180, HOC)	[34, 36-38]
Fractal dimension + 6 statistics + 36 higher order crossings (215, FD1)	[13, 34]
Fractal dimension + 6 statistics (35, FD2)	[13, 34]
3 Hjorth (15, HJORTH)	[40, 41]
Signal energy (5, SE)	[39]
Spectral power of δ, θ, α, and β bands (20, POW)	[4, 7, 12, 51]

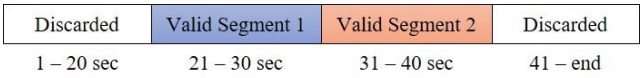

Fig. 3 Division of the EEG trial. EEG data at both ends are discarded. The middle part is retained and divided into two valid segments of the same length. Only valid segments are used for the subsequent processing.

5 Experiments

Based on our dataset, we carry out three simulations of aBCI under different paradigms. In the first simulation, we evaluate the recognition performance of aBCI when it can be re-calibrated from time to time. In the second simulation, we evaluate the long-term recognition performance of aBCI, especially when it operates without re-calibration during the course of usage. In the third simulation, we evaluate our proposed stable feature selection method.

5.1 Simulation 1: With Re-calibration

In this experiment, we simulate the recognition performance of an affective BCI where re-calibration of the system can be carried out each time before the subject uses the system. Specifically, we evaluate the within-session cross-validation recognition accuracy using the state-of-the-art affective EEG features referenced in Table 6.

We base the simulation on the EEG data we collected in Section 3. Each EEG trial lasts for 76 seconds. We discard both ends of the EEG trial and retain the middle part of the EEG trial for the subsequent processing, based on the assumption that emotions are better elicited in the middle of the trial. The division of the EEG trial is illustrated in Fig. 3. EEG features are extracted out of the valid segments of the EEG trials on a sliding-windowed basis. The final feature vector is a concatenation of the feature vectors from channel AF3, F7, FC5, T7, and F4, which were justified in [33] to be the top five discriminative

Table 7 Four-emotion recognition accuracy of Simulation 1, mean accuracy (%) ± standard deviation (%)

Feature	Subject					
	1	***2***	***3***	***4***	***5***	***6***
STAT	56.81 ± 10.52	44.75 ± 16.66	43.64 ± 13.89	71.43 ± 14.32	47.92 ± 15.44	73.88 ± 15.29
HOC	32.25 ± 10.50	30.25 ± 10.05	28.46 ± 10.24	43.53 ± 12.20	28.37 ± 10.95	36.61 ± 12.29
FD1	43.08 ± 13.98	37.39 ± 12.58	33.59 ± 8.12	58.59 ± 13.40	39.58 ± 12.05	54.58 ± 11.03
FD2	**57.14** ± 9.93	**46.88** ± 17.25	**45.76** ± 13.01	72.54 ± 14.49	**48.91** ± 15.42	**76.23** ± 15.51
HJORTH	53.24 ± 11.81	46.65 ± 14.30	41.41 ± 14.39	**72.77** ± 17.82	47.92 ± 15.67	72.54 ± 18.78
SE	45.54 ± 15.95	40.63 ± 12.67	41.96 ± 17.57	59.49 ± 16.23	41.96 ± 18.90	62.83 ± 20.02
POW	48.66 ± 12.21	46.88 ± 17.72	36.05 ± 14.70	69.20 ± 15.83	42.26 ± 18.03	62.72 ± 16.00
Upp Chan Lvl	42.79	42.80	42.79	39.36	42.70	42.79

Table 8 Four-emotion recognition accuracy of Simulation 2, mean accuracy (%) ± standard deviation (%)

Feature	Subject					
	1	***2***	***3***	***4***	***5***	***6***
STAT	**37.95** ± 5.01	24.79 ± 1.77	25.61 ± 1.65	39.49 ± 6.95	27.00 ± 3.98	30.39 ± 6.24
HOC	26.55 ± 4.27	24.78 ± 2.72	25.51 ± 2.63	28.68 ± 4.01	25.68 ± 2.78	27.01 ± 3.05
FD1	28.93 ± 3.98	24.52 ± 2.27	25.13 ± 2.83	33.68 ± 5.58	25.82 ± 3.01	28.45 ± 3.67
FD2	37.38 ± 6.05	25.25 ± 2.68	25.16 ± 2.62	**39.70** ± 7.10	27.52 ± 3.88	29.61 ± 6.25
HJORTH	31.77 ± 6.05	25.85 ± 3.33	27.05 ± 3.84	35.19 ± 8.13	26.32 ± 3.96	28.18 ± 4.82
SE	28.07 ± 2.83	25.80 ± 3.04	26.99 ± 2.79	38.35 ± 5.97	**27.96** ± 4.37	28.53 ± 3.84
POW	30.49 ± 4.30	**28.41** ± 4.25	**28.01** ± 3.55	39.42 ± 6.44	27.63 ± 4.53	**31.49** ± 6.94
Upp Chan Lvl	29.33	29.09	28.83	28.30	27.75	28.85

Table 9 Four-emotion recognition accuracy of Simulation 3 using the top n stable features. Mean accuracy (%) ± standard deviation (%) (# of stable features)

Feature	Subject					
	1	***2***	***3***	***4***	***5***	***6***
Our Selected Stable Feature	41.55 ± 4.31 (2)	30.24 ± 5.14 (7)	33.87 ± 3.55 (5)	45.22 ± 4.57 (1)	30.68 ± 3.43 (42)	33.63 ± 7.99 (34)

channels concerning emotion recognition. The width of the window is 4-second, and the step of the move is 1-second, as was used in [33]. Thus, each valid segment yields 7 samples.

In this within-session cross-validation evaluation, the training data and test data are from the EEG trials within the same session. As the time gap between the acquisition of training and test data is minimal, the evaluation can approximate the performance of the BCI where calibration is carried out shortly before use. We use one valid segment as the training data and the other as the test data, and repeat the process until each segment has served as the test data for once. The per-session recognition accuracy is averaged across all possible runs. In this very case, the evaluation is repeated twice per session, which is referred to as a two-fold cross validation. As we recognize four emotions in each session, the training data comprise 7×4 = 28 samples for four emotions, totally. Likewise, the test data consist of 28 samples for four emotions. We adopt the Logistic Regression (LR) [46] classifier. The simulation is implemented in MATLAB R2017a, where we use the MATLAB built-in toolbox of the LR classifier with the default hyperparameters. The evaluation is carried out for each of the subjects on a session-by-session basis. The mean classification accuracy over 16 sessions and the standard deviations are displayed in Table 7.

5.2 Simulation 2: Without Re-calibration

In this experiment, we simulate the recognition performance where no re-calibration is allowed during the long-term use of the BCI. We evaluate the inter-session leave-one-session-out cross-validation accuracy of the system for this purpose. Recall that in our dataset, we have 16 recording sessions per subject throughout the course of eight days. In this evaluation, we reserve one session as the calibration session whose EEG data are used to train the classifier, and pool together the data from the remaining 15 sessions as test data. We repeat the evaluation until each session has served as calibration session for once. In this very case, the process will be repeated 16 times per subject, and the reported recognition accuracy is the mean accuracy of 16 runs. This evaluation is to simulate the system performance in the long run, since there is a longer time gap between the training session and testing sessions—up to eight days. We adopt the features referenced in Table 6 in this simulation, in the same sliding-windowed manner as in Section 5.1. We use only the valid segment 1 (see Fig. 3) of each EEG trial and reserve the valid segment 2 for the testing purpose in Simulation 3 introduced in the following section. The sliding-windowed feature extraction yields 7 samples per valid segment. The training data consist of 7×4 = 28 samples for four emotions recorded in the same session. The test data comprise 7×4×15 = 420 samples pooled together from the remaining 15 sessions. The mean classification accuracy over 16 runs and the standard deviations are displayed in Table 8.

5.3 Simulation 3: Stable Feature Selection

In this experiment, we validate the effect of our proposed stable feature selection algorithm based on the simulation of emotion recognition where no re-calibration is allowed during the long-term use of the BCI. This simulation is similar to simulation 2, with the focus on the comparison between the state-of-the-art feature set and the stable feature set we propose.

We propose to find the stable features on a subject-dependent basis. The subject-dependent evaluation intends to find subject-specific stable features for each subject. We quantify the long-

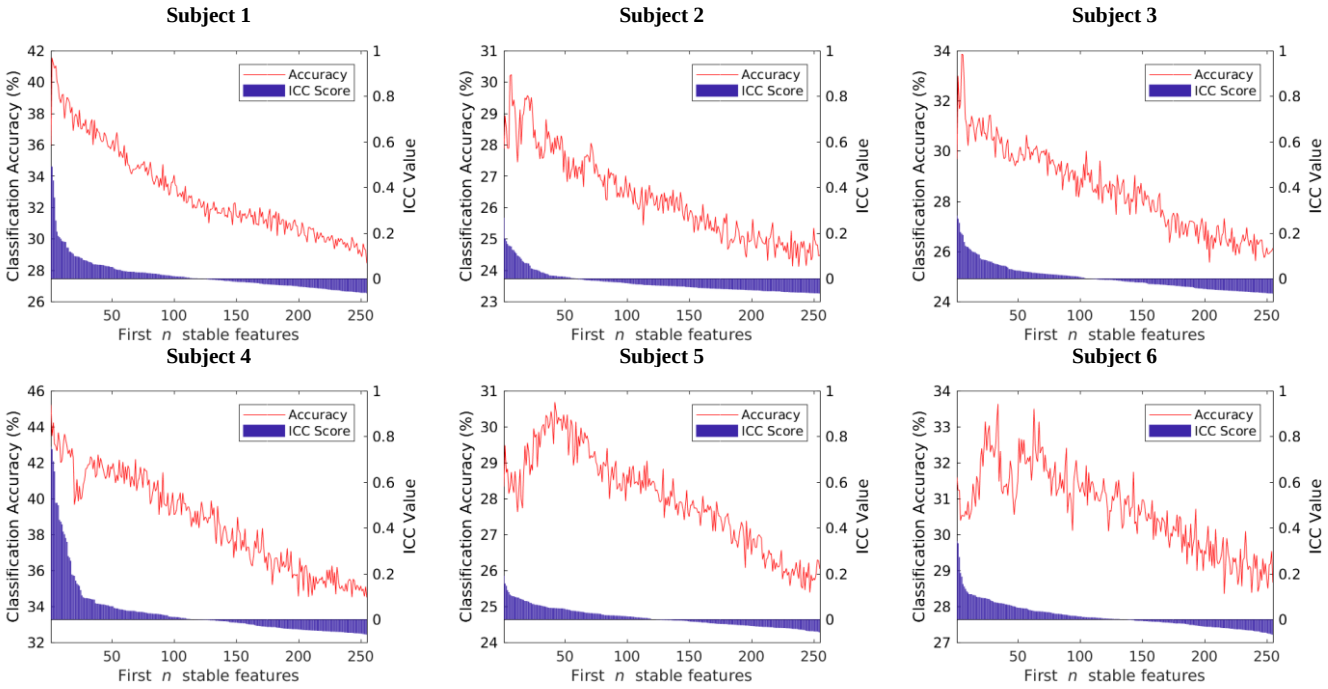

Fig. 4 ICC scores of each feature and the inter-session leave-one-session-out cross-validation accuracy using the top n stable features, $1 \leq n \leq 255$. The features are ranked by the ICC score in descending order.

term feature stability by computing the ICC scores on the training set consisting of the valid segment 1 (see Fig. 3) from all available trials (16 trials per subject), rank the feature according to the stability scores, and retain the optimal subset of features pertinent to the subject in question that maximizes the recognition accuracy when iteratively evaluating the inter-session leave-one-session-out cross-validation accuracy using the top n stable features. The results are shown in Table 9 and Fig. 4. After we find the stable features, we evaluate the performance of the stable features on the test set comprising the valid segment 2 (see Fig. 3) from all available trials. The recognition performance on the test set is shown in Table 10.

6 Results and Discussions

6.1 Simulation 1: With Re-calibration

Table 7 shows the mean accuracy ± standard deviation per subject based on the 2-fold cross-validation evaluation, which simulates the use case where re-calibration is allowed each time before a subject uses the BCI. The recognition accuracies vary between subjects and features, ranging from 28.37 % (Subject 5, HOC) to 76.23 % (Subject 6, FD2). HOC is found to be inferior to other referenced features on all subjects. The best performing feature varies between subjects. For subject 1, 2, 3, 5, and 6, referenced feature set FD2 yield better recognition accuracy than other referenced features in most cases. For subject 2, FD2, POW and HJORTH features give similar performance, outperforming other referenced features. For subject 4, STAT, FD2 and HJORTH features yield comparable results, being better than other referenced features. In general, FD2 performs well on all subjects in this simulation, which may suggest that FD2 is good for the use case where re-calibration is allowed from time to time.

For a four-class classification task, the theoretical chance level of random guess is 25.00 %. However, it is known that the real chance level is dependent on the classifier as well as the number of test samples. For an infinite number of test samples, the real chance level approaches the theoretical value. For a finite number of test samples, the real chance level is computed based on repeated simulations of classifying samples with randomized class label, as is suggested in [47, 48]. We carry out such simulation and present also in Table 7 the upper bound of the 95 % confidence interval of the simulated chance level for the best performing feature (in bold) for each classifier. Results show that the best-performing features yield recognition accuracy higher than the upper bound of the chance level. We assert that the best-performing features perform significantly better than chance level at a 5 % significance level.

6.2 Simulation 2: Without Re-calibration

Table 8 shows the mean accuracy ± standard deviation per subject based on inter-session leave-one-session-out cross-validation evaluation, which simulates the long-term recognition performance of the BCI when no re-calibration is permitted during use. Notable accuracy drop can be observed, compared to when re-calibration is allowed at each new session. This experiment establishes that intra-subject variance of affective feature parameters does exist and does have a negative impact on the recognition performance, though the severity varies from subject to subject. For subject 2 and 3, the recognition performance is severely affected by the variance—the best recognition performance has dropped and fallen within the 95 %

confidence interval of the simulated chance level. We therefore assert that subject 2 and 3 are performing at random guess level. For subject 1, 4 and 6, the best performance remains significantly better than the chance level at 5 % significance level, which seems to suffer from the variance problem to a lesser extent. Subject 5 gives mediocre performance. We loosely categorize subject 1, 4, and 6 as good performer, subject 5 as moderate performer and subject 2 and 3 as weak performer.

6.3 Simulation 3: Stable Feature Selection

To improve the long-term recognition accuracy, we propose to use stable features to mitigate the intra-subject variance of the affective feature parameters. Ideally, stable feature should give consistent measurement of the same affective state over the course of time, therefore there is the possibility to mitigate the variance among repeated sessions on different days. We propose a feature selection method that consists in quantifying the long-term stability of features with ICC model, ranking the features according to stability scores and iteratively selecting the topmost stable feature for inclusion into the stable feature subset. We propose to find the subject-dependent stable features.

Fig. 4 presents the results of subject-dependent stable feature selection. The bar plot in Fig. 4 indicates the stability score given in ICC values. The higher the stability score, the less variance the feature exhibits. The stability scores are ranked in descending order. Table 11 shows the ranking of the top 10 most stable features and their respective ICC scores. As we can see, the feature stability varies from subject to subject. For subject 1 and 4, the stability scores of the topmost stable features are notably higher than that of the other subjects. Generally, we observe that only a fraction of the features carries positive stability scores. For those with negative stability score, it suggests that the variance of the feature parameters over the course of time is even larger than the variance of the feature parameters between different emotions. Intuitively, these unstable features contribute to the deterioration of long-term recognition performance.

The curves superimposed on the bar plots indicate the inter-session leave-one-session-out cross-validation accuracy for classifying four emotions using only the first n stable features, with n varying from 1 to 255. As we can see, the curves exhibit similar trend among all subjects. The accuracy peaks at a small subset of stable features, then deteriorates when more and more unstable features are included into the feature subset being examined as n increases. For subject 2, 3, 4, 5, and 6, we can clearly see that the accuracy quickly deteriorates as features that carry negative stability scores are included into the feature subset being examined. This experiment shows the advantage of stable features over unstable features when the long-term performance is the utmost concern, and establishes the effectiveness of our proposed feature selection method. The peak recognition accuracy (peak of the accuracy curves in Fig. 4) and the number of stable features needed to achieve the peak performance is given in Table 9. Comparing Table 9 with Table 8, we can see that stable features selected by our algorithm have outperformed nearly all referenced features. Comparing our features to the best-performing referenced features in Table 8 (bold values), our features improve the accuracy by 3.60 %, 1.83 %, 5.86 %, 5.52 %, 2.72 %, and 2.14 %, for subject 1, 2, 3, 4, 5, and 6, respectively. Moreover, our selected features have a smaller dimension than the referenced state-of-the-art features, mitigating the burden of classifier training.

In addition, we observe that ICC value is in direct correlation with the long-term recognition performance, which validates our hypothesis that using stable features improves the accuracy. As can be seen from Fig. 4 (and also Table 11), the stability scores of the top stable features for subject 1 and subject 4 are notably higher than that for the other subjects. The long-term recognition performance of selected stable features of subject 1 and subject 4 are also notably higher than that of the other subjects. Generally, the higher the stability score, the better the recognition accuracy.

Looking at the subject-dependent feature ranking in Table 11, we can see that the feature ranking exhibits similar pattern among subject 1, 4, and 6. Statistic features top the stability ranking, together with Hjorth features and some HOCs. However, for subject 2, 3 and 5, different ranking patterns are observed. HOCs are found to be more stable, mixed with some power features and Hjorth features. Interestingly, HOC features have been frequently selected given their relatively high stability scores, despite their mediocre performance in Simulation 1 in Table 7. It may suggest that HOC features exhibit good stability and are suitable for the use case where the long-term recognition performance shall be put into consideration. However, it is not the optimal features if re-calibration is allowed before using the BCI from time to time.

6.4 Comparison on the Test Data

We further examine the performance of the stable features on unseen test data comprising Segment 2 (see Fig. 3) of all available trials. To simulate the long-term recognition performance, the same inter-session leave-one-session-out cross-validation evaluation scheme is applied. The stable feature set remains the same as was found on the training data. The recognition accuracy using our proposed stable features as well as the referenced state-of-the-art features is presented in Table 10. The results are principally consistent with the findings based on training data set. Our stable features outperform the best-performing referenced features by 2.54 %, 0.23 %, 3.12 %, 1.92 %, and 1.62 %, for subject 1, 3, 4, 5, and 6, respectively.

6.5 Limitation

In this study, we have proposed and validated a stable feature selection method for EEG-based emotion recognition on a dataset comprising six subjects. Further studies are needed to conclude the performance on a larger dataset. We have taken a subject-dependent approach to finding the subject-specific stable features. Compared to our previous studies [15, 16] where we had taken a subject-independent approach, subject-specific stable features are found to be more effective. However, since the effective stable feature set is subject-dependent, to find which requires ample labeled affective EEG data recorded over a long course of time. The acquisition of such data may post a burden to the subjects. Although the stable features perform relatively better than the referenced state-of-the-art in the long run, the absolute recognition accuracy is still admittedly low. It remains an open question as to how we can effectively mitigate or even eliminate the need of frequent re-calibrations of the BCI.

Table 10 Comparison of inter-session leave-one-session-out cross-validation accuracy on the test data between using referenced state-of-the-art feature set and stable feature set selected by our proposed algorithm. Mean accuracy (%) ± standard deviation (%).

Feature	Subject					
	1	***2***	***3***	***4***	***5***	***6***
STAT	36.79 ± 6.04	26.80 ± 3.87	26.88 ± 3.97	38.68 ± 5.92	28.38 ± 4.06	31.29 ± 7.76
HOC	28.68 ± 3.11	24.51 ± 2.84	25.55 ± 3.87	28.62 ± 3.74	25.90 ± 2.67	27.23 ± 4.30
FD1	30.92 ± 3.58	24.64 ± 3.56	25.95 ± 4.43	35.51 ± 5.57	26.41 ± 2.87	29.99 ± 5.22
FD2	35.61 ± 5.47	26.44 ± 4.22	27.50 ± 3.57	40.54 ± 5.89	27.47 ± 3.49	31.82 ± 7.93
HJORTH	31.65 ± 5.86	26.62 ± 2.80	26.82 ± 3.15	38.47 ± 5.85	26.76 ± 2.84	29.64 ± 3.78
SE	26.28 ± 3.97	26.61 ± 5.40	26.64 ± 2.93	36.98 ± 8.46	28.89 ± 3.40	27.49 ± 5.36
POW	33.41 ± 7.11	**27.95** ± 3.66	28.04 ± 3.14	38.85 ± 8.02	27.65 ± 3.94	31.92 ± 7.68
Ours	**39.33** ± 6.13	26.52 ± 4.23	**28.27** ± 3.72	**43.66** ± 6.09	**30.81** ± 5.11	**33.54** ± 6.93

Table 11 Feature ranking of the top 10 stable features and their respective ICC scores.

Rank	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6	
	Feature	***Score***	***Feature***	***Score***	***Feature***	***Score***	***Feature***	***Score***	***Feature***	***Score***	***Feature***	***Score***
1	hoc1_T7	0.4921	beta_F4	0.2671	hoc9_FC5	0.2771	stat5_T7	0.7548	hoc2_FC5	0.1909	stat3_T7	0.3430
2	stat3_T7	0.4913	hoc31_T7	0.1751	alpha_F7	0.2630	stat3_T7	0.7443	hoc32_AF3	0.1570	stat5_T7	0.3331
3	stat5_T7	0.4302	hoc33_T7	0.1651	hoc10_FC5	0.2445	beta_T7	0.6914	hoc28_AF3	0.1469	beta_T7	0.2726
4	hoc2_T7	0.3557	hoc34_T7	0.1523	hoc11_FC5	0.2040	stat2_T7	0.6473	hoc29_AF3	0.1279	hoc1_T7	0.2030
5	mblty_T7	0.2547	hoc32_T7	0.1450	hoc3_T7	0.1973	se_T7	0.5098	hoc30_AF3	0.1194	stat2_T7	0.1872
6	stat4_T7	0.2057	hoc2_F4	0.1428	hoc4_T7	0.1911	actvt_T7	0.5098	mblty_F4	0.1086	mblty_T7	0.1544
7	cpxty_T7	0.1874	beta_F7	0.1413	hoc8_FC5	0.1607	hoc1_T7	0.4992	hoc8_F4	0.1048	stat4_T7	0.1438
8	hoc13_AF3	0.1815	beta_AF3	0.1269	mblty_F4	0.1439	hoc5_T7	0.4353	hoc19_AF3	0.1015	hoc25_T7	0.1334
9	hoc29_AF3	0.1749	hoc2_FC5	0.1249	alpha_AF3	0.1398	alpha_T7	0.4272	hoc33_AF3	0.1011	stat6_T7	0.1233
10	stat6_T7	0.1652	hoc35_T7	0.1173	stat4_F4	0.1385	hoc4_T7	0.4161	hoc34_F4	0.0980	hoc26_T7	0.1173

7 Conclusion

aBCI is an affective interface between the user and the computer that relies on spontaneous EEG signals to function. In many existing aBCI studies, machine learning techniques are leveraged to recognize the affective states, which consist in acquiring the affective EEG signals from the user and calibrating the classifier to the affective pattern of the user. However, affective neural patterns are volatile over time even within the same subject, and intra-subject variance exist in the affective feature parameters. Due to these challenges, the recognition accuracy cannot be maintained if the usage of aBCI prolongs without recalibration. We propose a stable feature selection method to select the optimal feature set that maximize the recognition accuracy for the long run of an aBCI. The proposed method consists in modeling the feature stability by ICC, feature ranking and iterative selection of stable features. We hypothesize that unstable features contribute to the accuracy deterioration when the aBCI operates without re-calibration over the course of time, and by using stable features, the recognition accuracy can be improved. We carry out extensive comparison between our stable features and the state-of-the-art features. In Simulation 1, we show the recognition accuracy of an aBCI using the state-of-the-art features, where the aBCI is allowed to be re-calibrated from time to time. In Simulation 2, we simulate the long-term usage of an aBCI and establish that accuracy deterioration will occur when the aBCI operates without re-calibration. In Simulation 3, we analyze the performance of stable features selected by our proposed method. We demonstrate the accuracy trajectory when we iteratively include features into the selected feature subset. Experimental results show that recognition accuracy peaks at a small subset of stable features, and as more unstable features are included, the recognition accuracy quickly deteriorates. The experiment results validate our hypothesis. Comparisons between our stable features and the referenced state-of-the-art features show that our stable features yield better accuracy than the best-performing referenced features by 1.83 % – 5.85 % on the training set, and by 0.23 % – 2.54 % on the test set.

We stress that existing studies have overlooked the performance evaluation of aBCI during long-term use, which may partly be due to the fact that few existing datasets contain long-term affective EEG recordings. In this paper, we present a dataset which includes multiple recording sessions spanning across several days for each subject. Multiple sessions across different days were recorded so that the long-term recognition performance of aBCI can be evaluated. We stress that it is equally important to inspect the long-term recognition performance of aBCI. We invite other researchers to test the performance of their aBCI algorithms on this dataset, and especially to evaluate the long-term performance of their algorithms.

Acknowledgment

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative.

References

[1] B. M. Appelhans and L. J. Luecken, "Heart rate variability as an index of regulated emotional responding," *Review of general psychology,* vol. 10, no. 3, p. 229, 2006.

[2] M. Najström and B. Jansson, "Skin conductance responses as predictor of emotional responses to stressful life events," *Behaviour Research and Therapy,* vol. 45, no. 10, pp. 2456-2463, 2007.

[3] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett, "The brain basis of emotion: A meta-analytic review," *Behavioral and Brain Sciences,* vol. 35, no. 3, pp. 121-143, 2012.

[4] K. Ishino and M. Hagiwara, "A feeling estimation system using a simple electroencephalograph," in *IEEE International Conference on Systems, Man and Cybernetics*, 2003, vol. 5, pp. 4204-4209.

[5] K. Schaaff, "EEG-based Emotion Recognition," Diplomarbeit am Institut fur Algorithmen und Kognitive Systeme, Universitat Karlsruhe (TH), 2008.

[6] Y. P. Lin, C. H. Wang, T. L. Wu, S. K. Jeng, and J. H. Chen, "EEG-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 489-492.

[7] K. Schaaff and T. Schultz, "Towards an EEG-based emotion recognizer for humanoid robots," in *IEEE International Workshop on Robot and Human Interactive Communication*, 2009, pp. 792-796.

[8] G. Chanel, J. J. M. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *International Journal of Human-Computer Studies*, vol. 67, no. 8, pp. 607-627, Aug 2009.

[9] M. Li and B. Lu, "Emotion classification based on gamma-band EEG," in *IEEE International Conference on Engineering in Medicine and Biology Society*, 2009, pp. 1223-1226.

[10] Y. P. Lin *et al.*, "EEG-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798-1806, 2010.

[11] M. Murugappan, R. Nagarajan, and S. Yaacob, "Combining spatial filtering and wavelet transform for classifying human emotions using EEG Signals," *Journal of Medical and Biological Engineering*, vol. 31, no. 1, pp. 45-51, 2011.

[12] X.-W. Wang, D. Nie, and B.-L. Lu, "EEG-based emotion recognition using frequency domain features and support vector machines," in *Neural Information Processing*, 2011, pp. 734-743.

[13] Y. Liu and O. Sourina, "EEG Databases for Emotion Recognition," in *2013 Internation Conference on Cyberworlds (CW)*, Yokohama, 2013, pp. 302-309.

[14] M. Kwon, J.-S. Kang, and M. Lee, "Emotion classification in movie clips based on 3D fuzzy GIST and EEG signal analysis," in *International Winter Workshop on Brain-Computer Interface (BCI)*, 2013, pp. 67-68.

[15] Z. Lan, O. Sourina, L. Wang, and Y. Liu, "Stability of features in real-time EEG-based emotion recognition algorithm," in *2014 International Conference on Cyberworlds (CW)*, 2014, pp. 137-144.

[16] Z. Lan, O. Sourina, L. Wang, and Y. Liu, "Real-time EEG-based emotion monitoring using stable features," *The Visual Computer*, vol. 32, no. 3, pp. 347-358, 2016.

[17] Z. Lan, O. Sourina, L. Wang, and Y. Liu, "Stable feature selection for EEG-based emotion recognition," in *2018 Internation Conference on Cyberworlds (CW)*, 2018, pp. 1-8. In Press.

[18] A. Savran *et al.*, "Emotion detection in the loop from brain signals and facial images," in *Proc. eNTERFACE*, 2006.

[19] 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, 2012.

[20] S. Koelstra *et al.*, "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18-31, 2012.

[21] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49-59, 1994.

[22] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162-175, 2015.

[23] M. M. Bradley and P. J. Lang, "The International Affective Digitized Sounds (2nd Edition; IADS-2): Affective ratings of sounds and instruction manual," University of Florida2007.

[24] A. Mehrabian, "Pleasure-Arousal-Dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261-292, 1996.

[25] N. A. Badcock *et al.*, "Validation of the Emotiv EPOC EEG system for research quality auditory event-related potentials in children," *PeerJ*, vol. 3, p. e907, 2015.

[26] N. A. Badcock, P. Mousikou, Y. Mahajan, P. De Lissa, J. Thie, and G. McArthur, "Validation of the Emotiv EPOC EEG gaming system for measuring research quality auditory ERPs," *PeerJ*, vol. 1, p. e38, 2013.

[27] S. Debener, F. Minow, R. Emkes, K. Gandras, and M. De Vos, "How about taking a low ‐cost, small, and wireless EEG for a walk?," *Psychophysiology*, vol. 49, no. 11, pp. 1617-1621, 2012.

[28] M. De Vos, M. Kroesen, R. Emkes, and S. Debener, "P300 speller BCI with a mobile EEG system: comparison to a traditional amplifier," *Journal of neural engineering*, vol. 11, no. 3, p. 036008, 2014.

[29] M. De Vos, K. Gandras, and S. Debener, "Towards a truly mobile auditory brain–computer interface: exploring the P300 to take away," *International journal of psychophysiology*, vol. 91, no. 1, pp. 46-53, 2014.

[30] M. Duvinage, T. Castermans, M. Petieau, T. Hoellinger, G. Cheron, and T. Dutoit, "Performance of the Emotiv Epoc headset for P300-based applications," *Biomedical engineering online*, vol. 12, no. 1, p. 56, 2013.

[31] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D: Nonlinear Phenomena*, vol. 31, no. 2, pp. 277-283, 1988.

[32] Y. Liu and O. Sourina, "Real-Time Fractal-Based Valence Level Recognition from EEG," *Transactions on Computational Science XVIII*, vol. 7848, pp. 101-120, 2013.

[33] Y. Liu, "EEG-based Emotion Recognition for Real-time Applications," Ph.D. Thesis, Nanyang Technological University, 2014.

[34] Y. Liu and O. Sourina, "Real-Time Subject-Dependent EEG-Based Emotion Recognition Algorithm," in *Transactions on Computational Science XXIII*: Springer, 2014, pp. 199-223.

[35] B. Kedem and E. Slud, "Time series discrimination by higher order crossings," *The Annals of Statistics*, pp. 786-794, 1982.

[36] P. C. Petrantonakis and L. J. Hadjileontiadis, "Adaptive Emotional Information Retrieval From EEG Signals in the Time-Frequency Domain," *IEEE Transactions on Signal Processing*, vol. 60, no. 5, pp. 2604-2616, 2012.

[37] P. C. Petrantonakis and L. J. Hadjileontiadis, "A novel emotion elicitation index using frontal brain asymmetry for enhanced EEG-based emotion recognition," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 5, pp. 737-746, 2011.

[38] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 186-197, 2010.

[39] F. Feradov and T. Ganchev, "Ranking of EEG time-domain features on the negative emotions recognition task," *Annual Journal of Electronics*, vol. 9, pp. 26-29, 2015.

[40] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalography and clinical neurophysiology*, vol. 29, no. 3, pp. 306-310, 1970.

[41] K. Ansari-Asl, G. Chanel, and T. Pun, "A channel selection method for EEG classification in emotion assessment based on synchronization likelihood," in *15th European Signal Processing Conference*, 2007, pp. 1241-1245.

[42] R. Horlings, D. Datcu, and L. J. Rothkrantz, "Emotion recognition using brain activity," in *9th international conference on computer systems and technologies and workshop for PhD students in computing*, 2008, pp. II. 1-6: ACM.

[43] J. J. Allen, H. L. Urry, S. K. Hitt, and J. A. Coan, "The stability of resting frontal electroencephalographic asymmetry in depression," *Psychophysiology*, vol. 41, no. 2, pp. 269-280, 2004.

[44] S. Gudmundsson, T. P. Runarsson, S. Sigurdsson, G. Eiriksdottir, and K. Johnsen, "Reliability of quantitative EEG features," *Clinical Neurophysiology*, vol. 118, no. 10, pp. 2162-2171, 2007.

[45] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological methods*, vol. 1, no. 1, pp. 30-46, 1996.

[46] F. C. Pampel, *Logistic regression: A primer*. Sage Publications, 2000.

[47] G. Müller-Putz, R. Scherer, C. Brunner, R. Leeb, and G. Pfurtscheller, "Better than random: a closer look on BCI results," *International Journal of Bioelectromagnetism*, vol. 10, no. 1, pp. 52-55, 2008.

[48] E. Combrisson and K. Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *Journal of Neuroscience Methods*, vol. 250, pp. 126-136, 2015.

[49] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175-1191, 2001.

[50] K. Takahashi and A. Tsukaguchi, "Remarks on emotion recognition from multi-modal bio-potential signals," in *IEEE International Conference on Systems, Man and Cybernetics*, 2003, vol. 2, pp. 1654-1659.

[51] Y. Liu and O. Sourina, "EEG-based Dominance Level Recognition for Emotion-enabled Interaction," in *IEEE International Conference on Multimedia and Expo*, Melbourne, 2012, pp. 1039-1044.

Declaration of Interest

The authors declare that there is no conflict of interest.