Affective Brain-Computer Interfaces: Choosing a Meaningful Performance Measuring Metric

Md Rakibul Mowla[a,*], Rachael I. Cano[b], Katie J. Dhuyvetter[a] and David E. Thompson[a]

[a]Mike Wiegers Department of Electrical & Computer Engineering, Kansas State University, Manhattan, KS 66506 USA.
[b]Department of Mathematics, Kansas State University, Manhattan, KS, 66506.

ARTICLE INFO

Keywords:
affective brain-computer interfaces,
balanced accuracy,
electroencephalogram,
support vector machines,
emotion classification,
performance measurement.

ABSTRACT

Affective brain-computer interfaces are a relatively new area of research in affective computing. Estimation of affective states can improve human-computer interaction as well as improve the care of people with severe disabilities. To assess the effectiveness of EEG recordings for recognizing affective states, we used data collected in our lab as well as the publicly available DEAP database. We also reviewed the articles that used the DEAP database and found that a significant number of articles did not consider the presence of the class imbalance in the DEAP. Failing to consider class imbalance creates misleading results. Further, ignoring class imbalance makes the comparison of the results between studies using different datasets impossible, since different datasets will have different class imbalances. Class imbalance also shifts the chance level, hence it is vital to consider class bias while determining if the results are above chance. To properly account for the effect of class imbalance, we suggest the use of balanced accuracy as a performance metric, and its posterior distribution for computing credible intervals. For classification, we used features from the literature as well as theta beta-1 ratio. Results from DEAP and our data suggest that the beta band power, theta band power, and theta beta-1 ratio are better feature sets for classifying valence, arousal, and dominance, respectively.

1. Introduction

The term *affective* [64] is a psychological concept referring to the experience of human emotion or feeling. Brain-computer interfaces (BCIs) are usually defined as a direct means of communication between the brain and external devices or systems which enable the brain signal to control some external activity [93]. Yet BCIs also allow investigation of brain activity and analysis of brain state. Affective Brain-Computer Interfaces (aBCIs) can be defined as systems that estimate human affect from brain signals. The interest in automatic detection of people's affective states has increased over the last few decades. Studies have shown that affective states play an important role in human decision making [21]. The ability to manage one's affective states is also related to the abilities of logical reasoning, learning and extracting important information [70]. According to Goleman's model of emotional intelligence, having knowledge of your own affective states is a key factor behind personal and professional success [26].

However, estimation of the affective state is a difficult task for several reasons. Human subjects do not always reveal their true emotions, and often inflate their degree of happiness or satisfaction in self-reports [76]. Additionally, there is some ambiguity in understanding and defining affective states [63].

Facial expression analysis is one of the most popular methods for estimating affective states [61], but it is possible to deliberately fake facial expressions unrelated to one's true affective state. Therefore, as Picard argued, the estimation may have a high error rate if someone has the ability to disguise his or her emotion [63].

Nevertheless, there is a growing interest in relationships between affective states and brain activities. Investigating affective states using electroencephalography (EEG) is becoming popular among researchers because EEG is one of the most convenient, non-invasive forms of recording brain activity. EEG also has high temporal resolution, which makes it a preferable candidate for fast affective state estimation [58]. Before using EEG-based BCIs to estimate affective states, one major challenge is to model affective states in a measurable and understandable scale. A current, widely accepted affective state model is the circumplex model of affect (Fig. 1), which was initially proposed by J. A. Russel [69]. Finding distinct physiological patterns for each affective state has also always been a major topic of

*Corresponding author
rakibulmowla@ksu.edu (M.R. Mowla); davet@ksu.edu (D.E. Thompson)
www.mrmowla.com,rakib.raju05@gmail.com (M.R. Mowla)
ORCID(s): 0000-0001-5765-8856 (M.R. Mowla); 0000-0002-1897-2743 (D.E. Thompson)

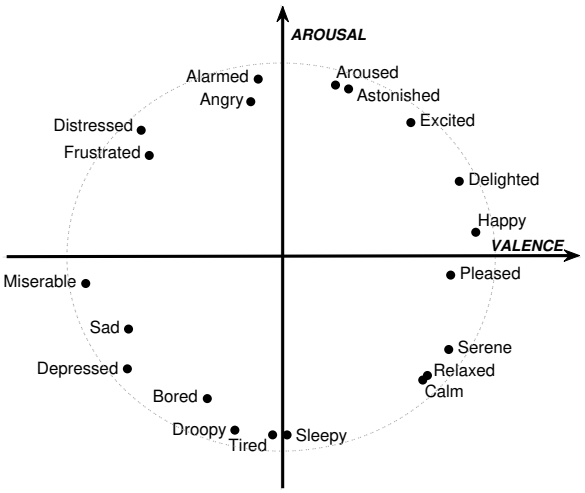


Figure 1: An example of the circumplex model where emotions are expressed in the valence and arousal dimensions. Valence refers to how pleasant or unpleasant an emotion is, and arousal refers to how exciting or boring it is. Words are placed according to direct circular scaling coordinates for 28 affect words from Russel's article [69].

interest for affective computing researchers [11]. Picard argued that emotion consists of more complex, underlying processes rather than outward physiological expression [63].

Interest in EEG-based emotion recognition has increased over time and is still growing. Searching "EEG emotion recognition" in Google scholar gave 115,000 results as of March 2020. Of these, 2100 were published just in the first quarter of 2020. Because these projects rely on individuals' emotional responses, the distribution of affective states (classes) is often uneven. However, most of these articles do not mention the class imbalance percentage, instead using only classification accuracy as a performance measuring metric. This practice creates a serious ambiguity and makes the results incomparable between different studies. For example, a publicly available database for emotion recognition known as the DEAP database [37] had been cited over 1600 times as of March 2020, and using the search keywords "EEG emotion recognition" within the DEAP-citing articles gave more than 1330 results. Out of those 1330 articles, at least 170 articles included the DEAP dataset in their analysis. Out of those 170 articles, only approximately 33 articles [98, 60, 90, 19, 59, 88, 28, 13, 65, 102, 75, 17, 97, 44, 99, 92, 22, 29, 1, 95, 94, 2, 15, 72, 38, 35, 40, 20, 3, 82, 25, 33, 77, 46, 101] mentioned or considered class imbalance. Classification accuracy, without considering class imbalance, is misleading for reasons we will present in this paper. Additionally, out of those 170 articles, only approximately 30 articles [49, 98, 97, 99, 7, 90, 24, 28, 67, 48, 88, 17, 14, 75, 39, 96, 95, 52, 101] discussed statistical significance. To us, these issues raised a few serious research questions:

1. Are the classification accuracies in these studies better than what could be achieved with unskilled classifiers?
2. If not, can it be said that these accuracies are significantly better than chance?
3. In the presence of class imbalance, what is the correct chance level?
4. What performance evaluation metric should be used in affect classification?

To investigate these questions, we present a case study of EEG-based detection of binary (high/low) valence, arousal, and dominance in response to different sets of stimuli. We use both our own data as well as the previously mentioned, publicly available DEAP database [37].

Affective states can be elicited through visual [43], auditory [42], and audio-visual stimuli[8], among other methods. The emotional experience is more profound when visual presentations are combined with auditory stimuli, intermediate under visual stimuli and minimal during auditory stimuli [27]. In our experiment, we used visual stimuli, the International Affective Picture System (IAPS) [43], to evoke emotions. The DEAP database used audio-visual stimuli.

2. Related Work

A large number of studies have been conducted on emotion recognition using EEG signals. With the improvement of dry electrodes, EEG is nearing or at the point of being a practical, out of the lab solution for affect recognition. More detailed EEG-based emotion recognition reviews can be found in [87, 23]. One major problem in EEG-based emotion recognition research is the lack of publicly available datasets. Consequently, researchers use their own data and as a result studies become more difficult to compare. To solve this problem, a few researchers developed publicly available datasets including the DEAP [37], USTC-ERVS [89] and MAHNOB-HCI datasets [74]. Among these datasets, the DEAP is the most cited and used for emotion recognition. Thus, we were motivated to use the DEAP dataset in this work.

Studies where DEAP was used as the benchmark dataset mostly used support vector machine (SVM) [65, 44, 75, 103, 88, 59, 85] for classification. The second most-used classification technique was the k-nearest neighbor (kNN) classifier [65, 103, 59]. Other classification techniques, such as deep convolutional neural network [45], decision tree [24], linear discriminate analysis (LDA) [4], logistic regression [103], discriminative graph regularized extreme learning machine (GELM) [103], back-propagation neural networks (BPNN) [67], probabilistic neural networks (PNN) [67], and multilayer perceptron (MLP) [85] have also been used to classify emotion on the DEAP dataset. Features used in these studies are statistical features: mean, standard deviation, variance, zero crossing rate [48, 78, 85, 53], Hjorth parameters [44, 54], fractal dimension [48, 57], Shannon entropy [48], spectral entropy [48, 85], kurtosis [30], skewness [98], different EEG band powers [78, 100], relative power spectral density (PSD) for delta, theta, alpha, beta and gamma frequency bands [90], differential entropy (DE), differential asymmetry (DASM), rational asymmetry (RASM), asymmetry (ASM) [103], wavelet coefficients [59], and higher order crossings (HOC) [65].

In the DEAP dataset, emotions are expressed in valence, arousal, and dominance dimensions on discrete 9-point scales. To design the classification model, those scales need to be labeled. Here also, inconsistencies exist between different studies. Not only are different numbers of classes chosen by different groups, but even within studies using the same number of classes, the thresholds are different. In these previously mentioned studies on the DEAP, classification labels were created by splitting the ratings into 3-class (1-3:negative, 4-6:neutral, and 7-9:positive) [34], 3-class (1-4.5:negative, 4.5-5.5:neutral, 5.5-9:positive) [85], 2-class (High/low, 4.5-9: high) [19], 2-class (negative: ratings ≤ 5, positive: ratings > 5) [90], 2-class (negative: ratings < 5, positive: ratings ≥ 5) [60, 28, 88], and 2-class (1-3: low and 7-9: high) [53]. Hence, the class imbalance in all these studies are different, based on their individual approaches to generating class labels.

Even though all the above-mentioned studies used the DEAP dataset, where significant class imbalance exists, very few studies have considered class imbalance while reporting results. Studies where class imbalance was considered mainly reported the F1 score [37, 99, 75, 24, 60]. A few other studies used receiver operating characteristic (ROC) [53, 65], area under ROC (AUC) [44] and balanced accuracy [17] along with the most common metric: accuracy. But AUC can be a misleading metric for a comparative study, especially in the presence of variable class imbalance [50]. Computing the F1 score for multiclass classification is also not straightforward, because F1 can be computed using macro-averaging or micro-averaging [80]. The difference between macro- and micro-averaged F1 can be large; if studies do not report which was used then comparing results is impossible. For example, [28] reported classification accuracies of 67% and 69% and F1 scores of 0.67 and 0.69 for valence and arousal, respectively. It is not clear whether macro- or micro-averaging was used, or if F1 scores were even calculated for both classes. Thus, any comparison between that study and others may lead to false conclusions.

To eliminate these above-mentioned problems, we are suggesting the field adopt balanced accuracy as the classification performance evaluation metric in high/low valence, arousal and dominance classification. To our knowledge, this has only been used in [17]. However, that study did not consider the lower bound of the credible intervals for balanced accuracy; here we will further discuss using the posterior distribution of balanced accuracy to compute credible intervals and perform statistical significance testing.

3. Data Description

We have used data from the publicly available DEAP dataset and EEG recordings from our lab.

3.1. Database for Emotion Analysis Using Physiological Signals (DEAP)

The DEAP is a publicly available, multimodal dataset consisting of 32-channel EEG, electrooculography (EOG), electromyography (EMG), galvanic skin response, respiration, plethysmograph, and temperature data [37]. We will

only use EEG recordings for the classification task. These signals were collected from thirty-two healthy participants, with an equal male-female ratio and an average age of 24.9 years. Data were recorded at a sampling rate of 512Hz and then pre-processed.

Minute-long music videos were used as emotional stimuli. After each video, participants were provided enough time to rate those videos for valence, arousal, and dominance on a discrete 9-point scale using self-assessment manikins (SAM) [9]. Each participant viewed forty videos.

3.2. Data collected at Brain and Body Sensing (BBS) lab

The BCI2000 [71] system was used to present picture stimuli to the participants. Each picture was displayed for 6.7 seconds, followed by a 20.8s pause for participants' self-report. A total of 244 pictures were selected from IAPS [43] images; the average valence and arousal ratings reported in the IAPS manual of the selected pictures are shown in Fig. 2. Pictures were presented in six blocks, with breaks for participant comfort. EEG data were recorded using a Cognionics Mobile-72 EEG system with a sampling frequency of 600Hz. Cognionics Mobile-72 EEG system is a high-density 64 channel EEG system with active Ag/AgCl electrodes. Reference and ground were on the right and left mastoids, respectively. In total, we had nine participants. Data from two participants have been excluded due to one

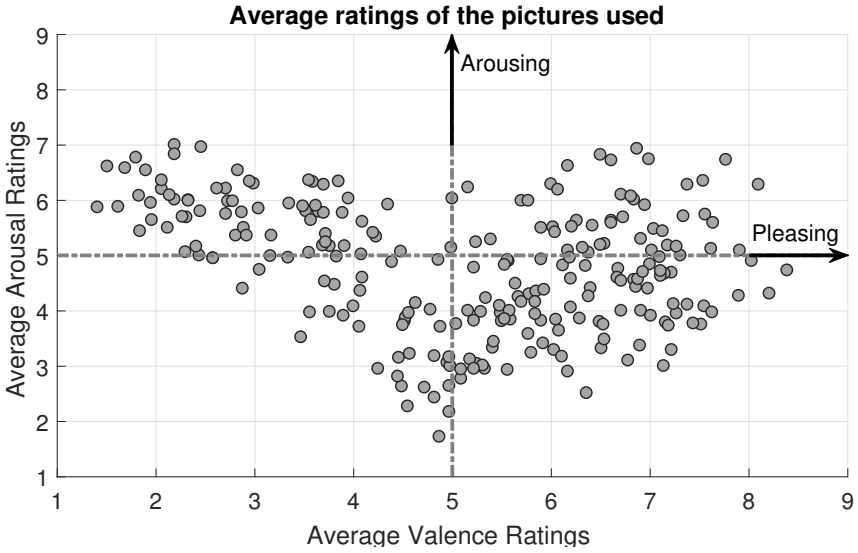


Figure 2: Visualization of average valence and arousal ratings (from the IAPS manual) [43] of the pictures used to collect data at the BBS lab.

data entry error and one battery failure. All participants were healthy college students with an age range of 21 to 22 years. Each participant was shown 244 pictures through two or three different sessions (i.e., visits). Most participants performed one session per day. However, a few participants performed multiple sessions on the same day (e.g., one session in the morning and one in the evening of the same day). Each participant rated each stimulus for valence, arousal, and dominance on a discrete 5-point scale using self-assessment manikins (SAM) [9].

3.3. Pre-processing

For the DEAP, both raw and pre-processed data are available for use. We used the MATLAB-ready preprocessed version of the data. The pre-processing steps were common-average referencing, down-sampling to 128Hz, band-pass filtering with cut-off frequencies of $(4.0-45.0)$ Hz, and eye blink artifact removal via independent component analysis. We then transformed data using scalp surface Laplacian or current-source density (CSD) because it has been argued that CSD transformation gives a more sensitive index of individual variations in frontal asymmetry than other EEG recording montages and also helps to reduce non-frontal contributions to frontal asymmetry [84, 6].

The data collected at the BBS lab was filtered using a finite impulse response (FIR) bandpass filter with corners at $(4.0 - 45.0)$ Hz. Data were then transformed into scalp surface Laplacian or current-source density (CSD) using the CSD toolbox [36] which provides a MATLAB implementation and uses the spherical spline algorithm [62].

4. Methods

Let $x(t) \in \mathcal{R}^T$ be the time series of a recording from a single electrode with N samples. The first and second derivatives of $x(t)$ with respect to time are $x'(t)$ and $x''(t)$, respectively. Standard deviation of $x(t)$, $x'(t)$ and $x''(t)$ are denoted as σ_x, σ_d, and σ_{dd}, respectively. Class labels are denoted by $c \in \{1, 2, \ldots, C\}$ and predicted class labels are denoted by y when classifying. $\mathbb{H}$ denotes entropy.

4.1. Feature sets

4.1.1. Frequency domain features

Power spectral density and signal power at different frequency bands are popular features for EEG-based affective state classification and have been used as features in several studies [47, 32]. Spectral density and band powers can be computed using various algorithms, including Fast Fourier Transform, short-Time Fourier Transform, or Welch's power spectral density estimation algorithm. We used Welch's power spectral density (PSD) estimation method [91] and then computed power in each band from the resulting PSD. The frequency ranges used for EEG bands varies slightly between different studies. In our analysis, the frequency ranges were theta: (4-8) Hz, alpha: (8-12) Hz, Beta-1: (12-18) Hz, Beta-2: (18-30) Hz, and Gamma: (31- 63) Hz.

It has been argued that frontal EEG asymmetry can be a moderator and mediator of affective state [18, 5]. By contrast, frontal alpha asymmetry is mostly used as a discriminator between depressed and healthy individuals [86], though it also can be used for affective state classification. Here, we will use both frontal EEG asymmetry and frontal alpha asymmetry (8-12 Hz) as features for classifying affective states. If R_p represents the signal power of electrodes located at the right frontal lobe and L_p represents the signal power of electrodes located at the left frontal lobe, then frontal EEG asymmetry can be calculated from

$$\text{Frontal asymmetry} = \ln\left(\frac{R_p}{L_p}\right) \tag{1}$$

Another form of the frontal asymmetry is the normalized version of equation (1) and is written as

$$\text{Frontal asymmetry} = \ln\left(\frac{R_p - L_p}{R_p + L_p}\right) \tag{2}$$

We used equation (1) to find the frontal asymmetry. We computed both the frontal asymmetry index (FAI) over $0 - 64$Hz and frontal alpha asymmetry index (FAAI) over the alpha band. Since the DEAP data was bandpass filtered using corner frequencies of 4 Hz and 45 Hz, we used the same band on our own data for consistency. Those filters would have reduced some frequencies in the range of calculations, but the effect of the reduction is consistent in all datafiles. FAI and FAAI were computed using the following symmetric pairs of electrodes: Fp1-Fp2, AF3-AF4, F3-F4, F7-F8, FC5-FC6, FC1-FC2.

We also used frontal theta-beta ratios (TBR) as frequency domain features. TBR has not been used previously for affective classification, but it has been reported to be related with affective traits [68]. To compute the frontal TBR we used equation (3)

$$\text{TBR} = \ln\left(\frac{\theta_p}{\beta_p}\right) \tag{3}$$

where θ_p represents the theta band power and β_p represents the beta band power of electrodes located over the frontal lobe. Frequency ranges for beta-1 and beta-2 were used in β_p to compute TBR1 and TBR2, respectively. TBR1 and TBR2 are computed for each frontal electrode (Fp1, Fp2, AF3, AF4, F3, F4, F7, F8, FC5, FC6, FC1, FC2).

4.1.2. Hjorth parameters

Hjorth parameters are time-domain features of EEG recordings, proposed by Bo Hjorth [31]. Hjorth parameters have been recently used in several studies [32, 54] as features for affective state estimation. The parameters are Activity, Mobility, and Complexity. Activity is simply the variance of the time signal. If the signal is denoted as $x(t)$, then Activity $= \sigma_x^2$ and is the measure of the squared standard deviation of amplitudes. Mobility measures the standard deviation of the slope with respect to the standard deviation of the amplitude. Mobility is defined as the square root of the ratio between the variances of the first derivative and the time signal. Complexity is a measure of how much the

time signal deviates from a pure sine shape and is defined as the ratio between the mobility of the first derivative of the time signal and the mobility of the time signal.

$$\text{Mobility} = \frac{\sigma_d}{\sigma_x}$$

$$\text{Complexity} = \frac{\sigma_{dd}/\sigma_d}{\sigma_d/\sigma_x}$$

We used mobility and complexity as features. For each trial, there was one value for mobility and complexity values for each EEG electrode.

4.1.3. Entropy

Entropy is a measure of disorder in a system. In the case of EEG, entropy measures the irregularity in the signal. Spectral entropy of EEG recordings has been used to discriminate different affective states in other studies [79] and it recently has been used in recognition of emotional states [103]. We used spectral entropy (SE), which is the normalized Shannon entropy of the power spectrum.

$$\text{Spectral Entropy} = -\frac{\sum_{i=1}^{N} p(X=i)\log_2 p(X=i)}{\log_2 N} \tag{4}$$

where X denotes the power spectrum of the time series $x(t)$, $p(X)$ is the spectral distribution such that $\sum_{i=1}^{N} p(X = i) = 1$, and N is the number of frequency bins.

4.1.4. Feature sets

For valence, arousal and dominance classification we used seventeen different feature sets: frontal asymmetry index (FAI), frontal alpha asymmetry index (FAAI), theta beta-1 ratio (TBR1), theta beta-2 ratio (TBR2), theta band power (ThetaP), alpha band power (AlphaP), beta band power (BetaP), gamma band power (GammaP), TBR1 and TBR2 together (TBR-C), theta, alpha, beta and gamma band power all together (TABG), Hjorth parameters (Hjorth), entropy (Entropy), power spectral density (PSD), beta alpha ratio (BARatio), all feature sets mentioned previously together (All), and principal components of all feature sets (All-PCA). For All-PCA, we used the principal components which contained 98% of total variability. These different feature sets, used with each of 32 participants, resulted in 17×32 classification results, in each affective dimension, for each classifier.

4.2. Classification

The ultimate goal for emotion estimation is a many-class classification or continuous-output regression. However, for this initial investigation, we focused on the easier binary classification problem, following multiple literature examples [19, 90, 60, 28, 88, 53]. Thus, we used a two-class classification system for each of valence, arousal, and dominance. Participants in our experiments rated each axis from 1 to 5; we labeled $ratings < 3$ as low valence, arousal, and dominance and $ratings \geq 3$ as high valence, arousal, and dominance. One participant never rated arousal less than 3, so for this participant (number 6) we shifted the split point from 3 to 4. In the DEAP database, participants rated each axis from 1 to 9; we labeled $ratings < 5$ as low and $ratings \geq 5$ as high, following several studies including the original work [37, 55, 48, 17].

We used both support vector machine (SVM) with a Gaussian/RBF kernel, and K-nearest neighbor (kNN) classifiers. These classifiers are the most commonly used techniques among published reports using the DEAP dataset [e.g. 55, 48, 17, 90, 85, 53, 65, 59]. We then confirmed the selection through an initial test with the classification learner app in MATLAB and a 70% train set, 15% validation set, and 15% test set strategy. Multiple kernel functions (i.e., linear, quadratic, cubic, RBF) for SVM classifiers, with auto-kernel scale parameter selection, were tested using the above-mentioned partition strategy for two participants. Similarly, for the kNN classifier, we chose the 'K'-value (K values were tested from 1 to 15) using the above-mentioned partition strategy and dataset. Once the classifier models and parameters were selected, we used cross-validation to estimate accuracy for each method on the full dataset. For the DEAP data, we used Leave-One-Out cross-validation to match the predominant approach in the literature [37, 19, 75, 17]. For our own data, which had more than six times the number of examples in the DEAP, we used 10-fold cross-validation.

4.2.1. Support vector machines (SVMs)

SVM uses a kernel trick and a separating hyperplane to classify new observations using support vectors from the training data. SVMs can be used for both regression and classification. In SVMs, with the observation vector $\mathbf{x}$ the predicted class label can be found using [56]

$$\hat{f}(\mathbf{x}) = \text{sgn}\Big(\hat{w}_0 + \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x})\Big) \tag{5}$$

Where $\alpha_i = \lambda_i y_i$, λ is the ℓ_1 regularization term and $k(\mathbf{x}_i, \mathbf{x})$ is the kernel function. For Gaussian kernel SVM, the kernel function is defined by

$$k(\mathbf{x}_i, \mathbf{x}) = \exp\Big(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x})^T \Sigma^{-1}(\mathbf{x}_i - \mathbf{x})\Big) \tag{6}$$

For implementation, we used the MATLAB built-in function `fitcsvm` for SVM with a Gaussian kernel.

4.2.2. K-Nearest Neighbours (KNN)

kNN is a simple classification algorithm where an example is classified based on the plurality vote of its k nearest neighbors. The nearest neighbours are chosen by a distance metric. Various metrics exist, including City block distance, Chebychev distance, Minkowski distance, Euclidean distance or Mahalanobis distance. We used the built-in MATLAB function `knnsearch` with $k = 9$ using Euclidean distance.

5. Performance Metrics

The most commonly used classification performance measurement metric is accuracy. Nevertheless, accuracy can be misleading, especially with the presence of class imbalance. In these situations, classifiers can learn from class label proportion rather than the features, a property sometimes known as "unskilled classification." In biased datasets, the unskilled performance is equal to the class imbalance. Thus, the same reported accuracy should be interpreted differently based on class bias. For example, consider a study reporting 80% accuracy in a two-class classification. This may be good performance on a balanced dataset but is at or below unskilled classification levels for biases $\geq 80\%$.

Comparing the performance of a similar classification task with different proportions of class labels is difficult. To make this kind of comparison meaningful, researchers suggest using other performance measuring metrics such as the Kappa statistic or area under the ROC curve (AUC) for imbalanced data. But since the multiclass ROC curve analysis is not well developed [41], AUC is not recommended for multiclass problems [73]. Moreover, the accuracy metric is the most widely used, and the most intuitive solution would be to make the accuracy metric meaningful by scaling down the baseline to be the performance of an unskilled classifier. One way to scale the baseline is to compute the balanced accuracy [83] where the accuracy in each class is considered separately.

5.1. Balanced Accuracy

If there are m number of classes, the balanced accuracy [83] is defined as

$$\textit{Balanced Accuracy} = \frac{1}{m}\sum_{k=1}^{m}\frac{C_{kk}}{n_k} \tag{7}$$

Here, n_k is the total number of observations in class k and C_{kk} is the number of correctly classified observations in that same class label.

Since our focus is on two-class classification, here, m=2. If the classifier performs equally well on both classes, then the balanced accuracy will be exactly equal to the conventional accuracy [83, 10]. Since balanced accuracy is the average accuracy of each class, it is unaffected by the class imbalance and is more meaningful than the traditional accuracy metric. Further, it has the convenient property that an unskilled classifier always achieves less than or equal to $1/m$ accuracy, regardless of class imbalance. For example, consider a majority-class classifier, a type of unskilled classifier, in the presence of 80% class bias in the training set of a 2-class classification task. That unskilled classifier will achieve an accuracy equal to the prevalence of the more likely class in the test set, roughly 0.8 if the training set was similar to the test set. However, the balanced accuracy will be $(1 + 0)/2 = 0.5 = 1/m$, regardless of the class proportions in the test set.

Although the traditional accuracy metric is a scaled binomial random variable, researchers often use a normal posterior distribution to compute credible intervals. The assumption behind the posterior normal distribution comes from the central limit theorem, where for a sufficiently large number of observations ($n \geq 30$), a binomial distribution can be approximated using the normal distribution. Nonetheless, this approximation becomes unreliable for small n. Particularly in the case of imbalanced data, the number of observations for the minority class can be smaller than the required number for the normal approximation. Therefore, finding chance performance and the credible interval of the classification rate for balanced accuracy is not as straightforward as it is in the case of traditional accuracy. For the two-class classification case, it is a combination of two separate distributions. In a multi-class scenario, accuracy in each class will have a separate distribution.

5.1.1. Credible intervals of Balanced Accuracy

If the probability of predicting correct classes of a classifier is denoted by $\mathcal{A}$ with a prior distribution $p(\mathcal{A})$, then the posterior is expressed as $p(\mathcal{A}|D)$ on observed data D. Let $y = 1$ and $y = 0$ represent correct and incorrect predictions, respectively. Now the classification predictions can be written as $y_1, y_2, \ldots, y_n$ which resembles the results of a Bernoulli experiment. So we can write

$$\begin{aligned} p(y_k|\mathcal{A}) &= \text{Bern}(y_k|p(\mathcal{A}) \\ &= \mathcal{A}^{y_k}(1-\mathcal{A})^{1-y_k} \end{aligned} \tag{8}$$

If the total number of success (correct predictions) of a Bernoulli trial $y_1, y_2, \ldots, y_n$ is c, then it follows a Binomial distribution.

$$\begin{aligned} p(c|\mathcal{A}, n) &= B(c|\mathcal{A}, n) \\ &= \binom{n}{c}\mathcal{A}^n(1-\mathcal{A})^{n-c} \end{aligned} \tag{9}$$

This suggests choosing Beta density as the prior of $\mathcal{A}$ since it is the conjugate prior of the Binomial distribution. This implies

$$\begin{aligned} p(\mathcal{A}) &= \text{Beta}(\mathcal{A}|a, b) \\ &= \text{Beta}(\mathcal{A}|1, 1) \end{aligned} \tag{10}$$

Now the posterior can be written using Bayes theorem as

$$\begin{aligned} p(\mathcal{A}|c, n) &= \frac{p(c|\mathcal{A}, n)p(\mathcal{A})}{p(c)} \\ &= \frac{B(c|\mathcal{A}, n) \times \text{Beta}(\mathcal{A}|1, 1)}{p(c)} \end{aligned} \tag{11}$$

From equation 11, we obtain the posterior $p(\mathcal{A}|c, n) = \text{Beta}(\mathcal{A}|c+1, n-c+1)$ and the posterior $(1-\alpha)100\%$ credible interval is [12]

$$\left[F^{-1}_{\text{Beta}(c+1,n-c+1)}(\alpha/2); F^{-1}_{\text{Beta}(c+1,n-c+1)}(1-\alpha/2)\right] \tag{12}$$

where $F^{-1}_{\text{Beta}(\cdot)}(\cdot)$ is the inverse density function of the Beta distribution and for 95% credible interval, $\alpha = 0.05$. In a multiclass scenario, each class has the distribution shown in equation (11). To find the posterior of the balanced accuracy $m-$fold convolution is used for m classes. Numerical approximations are used to compute the posterior since analytical forms are not available for the $m-$fold convolution. We used a MATLAB routine to compute the credible intervals of balanced accuracy provided in [10].

5.2. F1 measure

Another alternative performance evaluation metric is the F1-measure which has been used in some papers using the DEAP dataset [37, 19, 75]. The F-measure was originally proposed by Van Rijsbergen [81] and is defined as [16]

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R} \tag{13}$$

where P and R denotes precision and recall and are defined as $P = tp/(tp + fp)$, $R = tp/(tp + fn)$ ($tp \rightarrow$ true positive, $fp \rightarrow$ false positive, $fn \rightarrow$ false negative). β is a parameter to control balance between P and R. When $\beta = 1$, F_1 becomes the harmonic mean of precision and recall. Hence the F_1 measure is

$$F_1 = \frac{2PR}{P + R} \tag{14}$$

Since P and R are calculated considering one class as a positive class, P and R have to be calculated per class and hence the F1 measure as well. P and R per class can be calculated in two ways: microaveraging and macroaveraging. Microaveraging aggregates the individual true positives, false positives, and false negatives of each classes to calculate the P and R.

$$\begin{aligned}
miP &= \frac{\sum_{k=1}^{m} C_{kk}}{\sum_{k=1}^{m} C_{kk} + \sum_{k=1}^{m} \sum_{\substack{j=1 \\ j \neq k}}^{m} C_{jk}} \\
miR &= \frac{\sum_{k=1}^{m} C_{kk}}{\sum_{k=1}^{m} C_{kk} + \sum_{k=1}^{m} \sum_{\substack{j=1 \\ j \neq k}}^{m} C_{kj}} \\
miF_1 &= \frac{2 \cdot miP \cdot miR}{miP + miR}
\end{aligned} \tag{15}$$

An alternative technique is known as macroaveraging. In macroaveraging, P and R are calculated for each class and then the F_1 for each class is computed using P and R of individual classes. The macroaverage is the simple average of individual class F_1 scores.

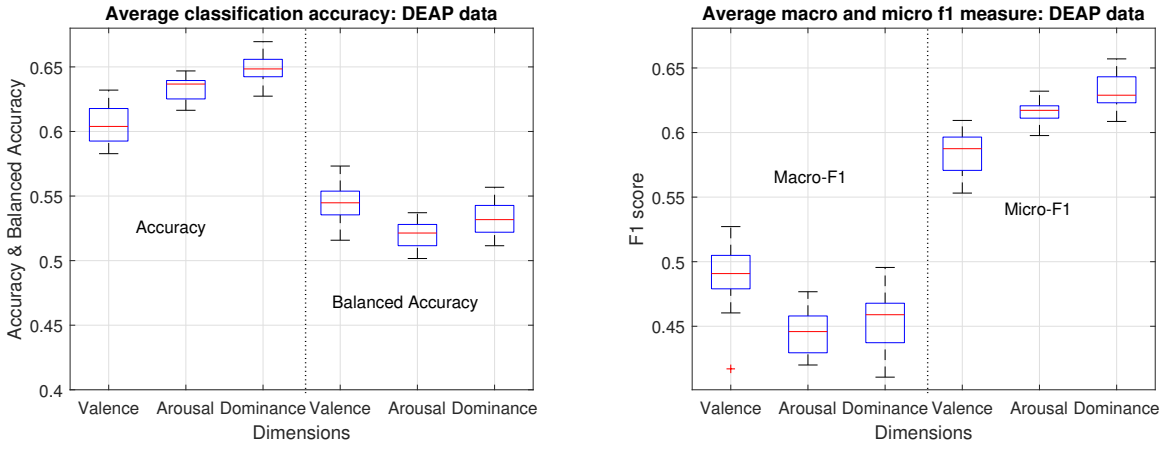
$$\begin{aligned}
P_k &= \frac{C_{kk}}{C_{kk} + \sum_{\substack{j=0 \\ j \neq k}}^{m} C_{jk}} = \frac{C_{kk}}{\sum_{j=1}^{m} C_{jk}} \\
R_k &= \frac{C_{kk}}{C_{kk} + \sum_{\substack{j=0 \\ j \neq k}}^{m} C_{kj}} = \frac{C_{kk}}{\sum_{j=1}^{m} C_{kj}} \\
maF_1 &= \frac{1}{m} \sum_{k=1}^{m} \frac{2 \cdot P_k \cdot R_k}{P_k + R_k}
\end{aligned} \tag{16}$$

The difference between miF_1 and maF_1 can be significant. Macroaveraging gives equal weight to each class, whereas microaveraging gives equal weight to each per-class classification decision. Since the F_1 measure ignores true negatives, the influence of large classes is higher than small classes in micro-averaging [51], which runs counter to the use of F1 in biased datasets. On the other hand, the F_1 measure's use of harmonic means suggest that the averaging should be over the per-class classification decision of each instances. In that sense, macro-averaging is not consistent with the original definition of the F_1 measure [66]. Hence, we do not yet have a convincing argument for choosing between miF_1 and maF_1 for multiclass classification.

6. Results

Since we have used seventeen different feature sets, it is not feasible to show all the results here. To summarize the results, we averaged the classification results over all participants for each feature set. Those average classification accuracies, and other performance metrics for different feature sets, are presented in Fig. 3, Fig. 5 and Table 1. All the results are for the SVM classifier, since it outperformed the kNN approach.

6.1. DEAP Dataset



(a) Accuracy and balanced accuracy

(b) Macro and micro f1 measure

Figure 3: Average classification rate of all participants in high/low recognition of valence, arousal and dominance for different features using the DEAP dataset.

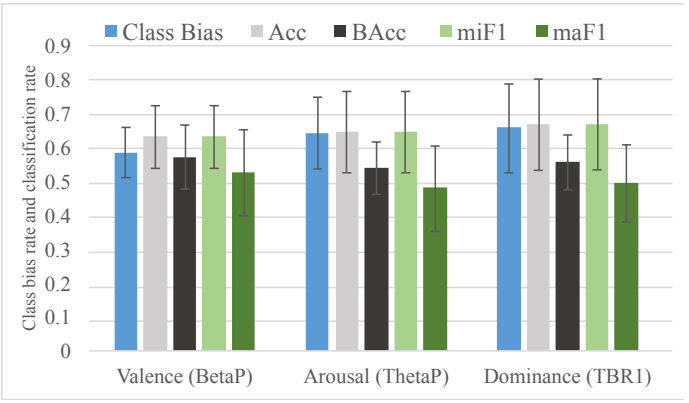


Figure 4: Class bias rate and average classification rate of all participants (DEAP data) using accuracy (Acc), balanced accuracy (BAcc), micro-F1 (miF1), and macro-F1 (maF1) metrics in high/low recognition of valence (features: BetaP), arousal (features: ThetaP) and dominance (features: TBR1).

Fig. 3(a) shows the average classification accuracies and balanced accuracies on the DEAP for different feature sets using SVM. The mean classification accuracies for all features are 0.604, 0.637, and 0.648 for valence, arousal, and dominance, respectively. These results are comparable with the results reported in the original DEAP paper [37] and other related studies [75, 19], and appear to be above chance performance. But the balanced accuracies on the right side of the Fig. 3(a) show very different results. The mean classification rate, in balanced accuracies, for all feature are 0.544, 0.521, and 0.531 for valence, arousal and dominance respectively. Only valence recognition appears to be much above chance. Notably, the average class bias rate in these three dimensions are 0.59, 0.64 and 0.66 for valence, arousal, and dominance (Blue colored bars in Fig. 4).

Fig. 3(b) shows the average macro- and micro-averaged F1 measure for different feature sets using SVM. The mean macro-F1 for all feature were 0.49, 0.445 and 0.46 for valence, arousal, and dominance, respectively. On the contrary, the mean micro-F1 for all features were 0.59, 0.62 and 0.63 for valence, arousal, and dominance, respectively. The best classification rate in the valence dimension was achieved using beta band power as a feature, as we found using balanced accuracy. For valence, the average across all participants macro-F1 for BetaP feature was 0.53 and the

Table 1
The average for all participants classification rate, in terms of balanced accuracy (BAcc), with ± standard deviation of the classification rate and the lower bound of the 95% credible intervals of balanced accuracies for different feature sets. Bold values are representing the maximum classification rate for the corresponding feature set.

Features	Valence		Arousal		Dominance	
	Balanced Accuracy (BAcc)	Lower bound of BAcc	Balanced Accuracy (BAcc)	Lower bound of BAcc	Balanced Accuracy (BAcc)	Lower bound of BAcc
PASI	0.545 ± 0.090	0.4297	0.528 ± 0.065	0.4227	0.532 ± 0.089	0.4262
FAI	0.522 ± 0.082	0.4089	0.512 ± 0.058	0.413	0.522 ± 0.070	0.4178
TBR1	0.548 ± 0.098	0.4267	0.525 ± 0.073	0.4235	**0.557 ± 0.08**	**0.4435**
TBR2	0.538 ± 0.070	0.4198	0.511 ± 0.072	0.4090	0.522 ± 0.066	0.4142
ThetaP	0.539 ± 0.070	0.4211	**0.537 ± 0.076**	**0.4336**	0.530 ± 0.092	0.4206
AlphaP	0.543 ± 0.078	0.4286	0.524 ± 0.078	0.4281	0.549 ± 0.071	0.4432
BetaP	**0.573 ± 0.093**	**0.4531**	0.530 ± 0.047	0.4263	0.537 ± 0.076	0.4247
GammaP	0.559 ± 0.096	0.4381	0.528 ± 0.050	0.4265	0.541 ± 0.074	0.4323
TBR-C	0.566 ± 0.093	0.4482	0.532 ± 0.066	0.4263	0.555 ± 0.084	0.4439
TABG	0.558 ± 0.071	0.4401	0.509 ± 0.067	0.4122	0.535 ± 0.087	0.4301
Hjorth	0.532 ±0.101	0.4159	0.527 ± 0.071	0.4268	0.520 ± 0.099	0.4104
PASI+FASI	0.547 ± 0.087	0.4355	0.521 ± 0.070	0.4207	0.534 ± 0.081	0.4307
Avg-Entropy	0.516 ± 0.065	0.4177	0.520 ± 0.060	0.4312	0.518 ± 0.064	0.4269
PSD	0.553 ± 0.091	0.4451	0.515 ± 0.068	0.4259	0.529 ± 0.041	0.4361
BARatio	0.523 ± 0.073	0.4077	0.502 ± 0.048	0.4054	0.512 ± 0.080	0.4059
All	0.552 ± 0.078	0.4447	0.507 ± 0.067	0.4178	0.523 ± 0.043	0.4290
All-PCA	0.537 ± 0.070	0.4160	0.517 ± 0.076	0.4086	0.548 ± 0.090	0.4329

micro-F1 was 0.63. For arousal, the average across all participants for macro-F1 from the ThetaP feature was 0.48 and the micro-F1 was 0.647. For dominance, the average across all participants' macro-F1 for the TBR1 feature was 0.495 and the micro-F1 was 0.67. Fig. 4 is included to further illustrate these results. Class bias rate, accuracy (Acc), balanced accuracy (BAcc), micro-F1 (miF1), and macro-F1 (maF1), all these are shown side-by-side using bar plots in Fig. 4.

Table 1 shows the average balanced accuracies and lower bound of the 95% credible intervals of balanced accuracies for different feature sets using equation (12). Values are in bold font represents the best feature set in terms of classification rate. All results are for the SVM classifier. The highest obtained balanced accuracy across all dimensions is 0.5732, achieved for valence recognition using beta band power. Unfortunately, the average lower limit of the credible intervals, in this case, is not above 0.5 (random chance). Though the average provides an overall recognition rate, it does not reflect the performance of individual participants. Explaining results for all features would be cumbersome; here we will explain classification results for each participant for only the best feature in each dimension. For valence, beta band power worked best. Using this feature, the balanced accuracy obtained for a participant (s10) with 0.75 and the lower bound of the credible interval is 0.622, which means that the valence classification rate is significantly above chance for this participant. Out of 32 participants, balanced accuracy is greater than 0.5 for 23 participants. For 8 of these participants, the lower bound of the credible interval is greater than 0.5. For arousal, theta band power worked best. Using the thetaP feature, the highest balanced accuracy obtained for a participant (s17) is 0.73 and the lower bound of the credible interval is 0.60, which means the arousal classification rate is significantly above chance for this participant. For 21 participants, observed balanced accuracy is greater than 0.5. However, only 4 participants were the lower bound of the credible interval greater than 0.5. For dominance, theta beta-1 ratio worked best. Using TBR1, the highest balanced accuracy obtained for a participant (s17) was 0.74 with a lower bound of 0.61, which means the

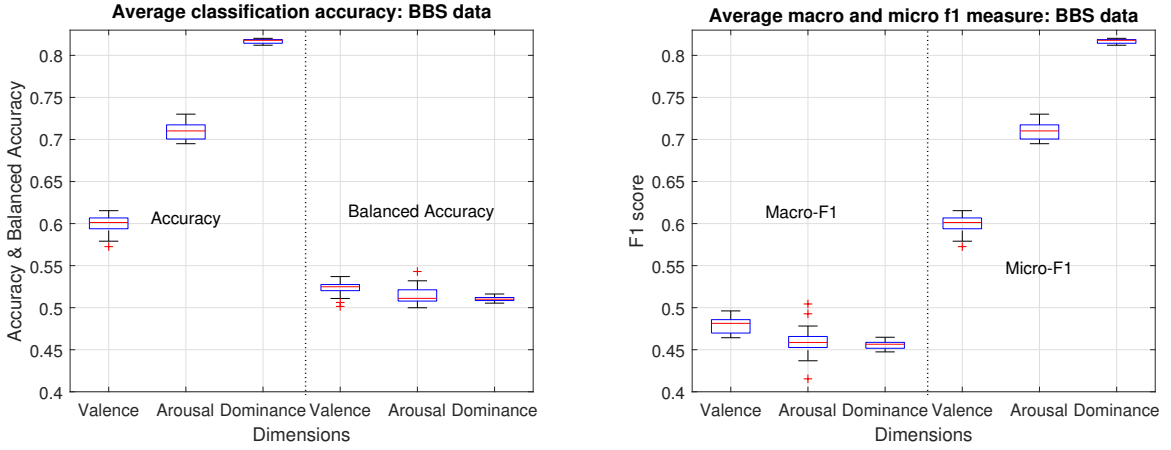
Table 2
The classification rate in terms of balanced accuracy, micro F1 (miF1) and macro F1 (maF1) scores of affect recognition compared to the DEAP dataset original work and related studies. The results shown here are average of all participants using beta band power (BetaP) features.

	Valence			Arousal			Dominance		
	bAcc	miF1	maF1	bAcc	miF1	maF1	bAcc	miF1	maF1
[37]	–	–	0.563	–	–	0.583	–	–	–
[19]	–	–	0.550	–	–	0.570	–	–	0.552
[75]	–	–	0.645	–	–	0.570	–	–	0.533
[17]	0.604	–	–	0.583	–	–	0.564	–	–
Current study	0.573	0.610	0.530	0.530	0.620	0.460	0.537	0.630	0.460

dominance classification rate is significantly above chance for this participant. For 24 participants, balanced accuracy is greater than 0.5. Yet again, only for 4 participants was the lower bound of the credible interval greater than 0.5.

Table 2 shows the affect recognition rate in terms of balanced accuracy, micro- and macro-averaged F1 score, compared with the original work [37] and some other related studies. Rather than presenting the best results in each dimension, we chose to present results for one specific feature set for consistency. The results presented under the current study are for beta band power (BetaP) feature using an SVM classifier. Note that our comparison studies seem to have picked the best result in each dimension for their reported results (though only Clerico et al. [17] unambiguously stated this).

6.2. Data at BBS lab



(a) Accuracy and balanced accuracy

(b) Macro and micro f1 measure

Figure 5: Average classification rate of all participants in high/low recognition of valence, arousal and dominance for different features using BBS data.

The data collected at the BBS lab using IAPS came from seven participants. For 2-class classification, the average class-biases were 0.60, 0.72, and 0.82 for valence, arousal, and dominance, respectively. For valence with SVM, the best 2-class classification results were obtained using gamma-band power considering the average of all participants. The obtained accuracy was 0.62 and the balanced accuracy was 0.54. The macro- and micro-averaged F1 scores were 0.49 and 0.60, respectively.

For arousal with SVM, the best 2-class classification results were obtained using the power asymmetry index (PASI) considering the average of all participants. The obtained accuracy was 0.73 and the balanced accuracy was 0.54. The macro- and micro-averaged F1 scores were 0.50 and 0.71, respectively.

For dominance with SVM, the best 2-class classification results were obtained using beta band power considering the average of all participants. The obtained accuracy was 0.82 and the balanced accuracy was 0.52. The macro- and micro-averaged F1 scores were 0.46 and 0.82, respectively.

7. Discussion

For the DEAP, the average class bias or majority class percentage in a 2-class classification scenario for valence, arousal and dominance are 0.59, 0.64 and 0.66 respectively. We have argued that class imbalance is important to understand the results of the classifier and should be reported. Performance metrics that include or account the class-biases are thus preferred to use. Any metric that ignores class imbalance will mislead readers. To illustrate this, consider the results from Table 1 where balanced accuracy and the lower bound of the 95% credible interval are presented for different feature sets for DEAP data using SVM. The best average classification accuracy for all participants in the valence dimension was 0.602 using beta band power as a feature, whereas the balanced accuracy, for this case, was 0.573. Without knowing the class bias and considering the accuracy metric, one might think the result is promising. But the lower bound of the 95% credible interval of balanced accuracy is below 0.5, so the classification rate cannot be claimed as statistically significant.

However, class imbalance for each participant for all three-dimension (valence, arousal, dominance) would be cumbersome and impractical to report. The biases mentioned earlier were averaged across all participants. Since affective state estimation is a participant-specific task, averaged results do not reflect individual performances. So comparisons using average results are not meaningful. Hence, we need something else which can address both the class imbalance problem and make the average performance meaningful. Considering those above-mentioned problems, balanced accuracy is a promising candidate since the baseline performance for balance accuracy is the same (50%) across all dimensions(valence, arousal, dominance) for all participants. Thus, balanced accuracy will make results easier to understand and compare. For example, just looking at the results in Table 1, we can easily conclude that the valence recognition rate is better than arousal and dominance recognition. Statistical comparison between the balanced accuracies for valence, arousal and dominance presented in Table 1 was done by using the MATLAB inbuilt function `ttest2`. Two-sample t-test resulted in the rejection of the null hypothesis (two groups are equal) when comparing valence and arousal. The valence recognition rate is significantly better than the arousal and dominance recognition rate with adjusted p-values 0.035 and $7.44e^{-06}$. The dominance recognition rate is also significantly better than arousal with adjusted p-value of 0.031. These three two-sample t-tests suggest that valence has the highest recognition rate and arousal has the lowest for the DEAP dataset.

Averages for all participants of the balanced accuracies, macro, and micro F1 measure are compared with other related studies in Table 2. Since they have not discussed the methods of statistical analysis, here we will use our obtained results shown in table 1 for discussion. Our average balanced accuracies are very similar to the highest balanced accuracy reported in [17]. In [17], it has been claimed that all the reported balanced accuracies were better than random voting classifiers with $p < 0.05$. This statement is true if we perform statistical analysis considering results from all participants as a group rather than individual participants. The number of participants with balanced accuracy above 0.5 is 25 for valence using all frequency band powers, 21 for arousal and 20 for dominance. In this case the probability that overall balanced accuracy is above chance are 0.66, 0.66 and 0.63 with intervals $(0.47-0.82)$, $(0.47-0.82)$, and $(0.44-0.79)$ for valence, arousal and dominance, respectively. But the significance of the experiment as a whole does not capture the significance of each participant's performance. Hence, just based on these statistics we are not comfortable to claim the accuracies are above chance. Rather we suggest using the probability of individual participants' performances being above chance to claim the results are significant. Using the number of participants that are significantly above chance, we have 6 for valence, 3 for arousal and 4 for dominance out of 32 participants. That tells us that the probabilities of a participant's classification accuracy being significantly above chance for valence, arousal and dominance are 0.19, 0.09 and 0.13 bounded by $(0.07-0.36)$, $(0.02-0.25)$ and $(0.04-.29)$, respectively. These are not very encouraging, as valence is only above the typical 0.05 threshold. This low rate of significant performance may be of concern for the EEG-based affective computing community, and as a community, we need to be more careful while reporting results.

8. Conclusion

We presented the experimental results for affective state estimation using the publicly available DEAP database and our lab data. We compared our results for DEAP data with the results reported in a few related studies. We used various features mentioned in the literature and also investigated theta-beta1 ratio as a novel feature for affect classification. Our findings showed that the beta band power is the most suitable for valence classification, theta band power for arousal classification, and theta beta-1 ratio for dominance classification.

In conclusion, we suggest using balanced accuracy and its posterior distribution as the performance evaluation metric for emotion estimation. Although F1 measure is a popular choice, it is not yet well established which F1 measure (macro/micro) we should use for multiclass classification. As our results demonstrate, that choice is important. Further, if macro-averaging is chosen, the statistical significance of the metric is not well understood.

In contrast to the F1 measure, balanced accuracy has several advantages. First, balanced accuracy does not have a "preferred class" and is thus comparable between groups. Second, the credible bounds can be calculated using known formulas. Third, the extension to large numbers of classes is straightforward. Fourth and finally, balanced accuracy is insensitive to class bias and always has the intuitive 1/m chance performance for unskilled classifiers.

We note that traditional accuracy metrics would have classified the performance of many more of our participants as statistically significant, relative to the number classified this way by balanced accuracy. Nevertheless, we maintain that balanced accuracy is far less misleading, and that the traditional accuracy metric substantially over-estimates performance is these unbalanced datasets.

Acknowledgment

This material is based upon work supported in part by Kansas State University faculty startup funds and in part by the National Science Foundation under Award No. 1910526. Opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. The authors would like to thank our participants for enduring long EEG sessions. The involvement of human participants with this research was approved by the Kansas State University Institutional Review Board under protocol No. 8328.

CRediT authorship contribution statement

Md Rakibul Mowla: Data collection software, Methodology, Analysis, Original draft preparation. **Rachael I. Cano:** Recruiting and preparing participants, Conducting experiments, Manuscript Revision. **Katie J. Dhuyvetter:** Recruiting and preparing participants, Conducting experiments, Manuscript Revision. **David E. Thompson:** Coordination, Design of Experiments, Training of Team Members, Manuscript Revision and Oversight.

References

[1] Ackermann, P., Kohlschein, C., Bitsch, J.Á., Wehrle, K., Jeschke, S., 2016. Eeg-based automatic emotion recognition: Feature extraction, selection and classification methods, in: 2016 IEEE 18th international conference on e-health networking, applications and services (Healthcom), IEEE. pp. 1–6.

[2] Al-Fahad, R., Yeasin, M., 2016. Robust modeling of continuous 4-d affective space from eeg recording, in: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE. pp. 1040–1045.

[3] Al-Fahad, R., Yeasin, M., Anam, A.I., Elahian, B., 2017. Selection of stable features for modeling 4-d affective space from eeg recording, in: 2017 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1202–1209.

[4] Al Zoubi, O., Awad, M., Kasabov, N.K., 2018. Anytime multipurpose emotion recognition from eeg data using a liquid state machine based framework. Artificial intelligence in medicine 86, 1–8.

[5] Allen, J.J., Coan, J.A., Nazarian, M., 2004. Issues and assumptions on the road from raw signals to metrics of frontal eeg asymmetry in emotion. Biological psychology 67, 183–218.

[6] Allen, J.J., Reznik, S.J., 2015. Frontal eeg asymmetry as a promising marker of depression vulnerability: Summary and methodological considerations. Current opinion in psychology 4, 93–97.

[7] Arnau-González, P., Arevalillo-Herráez, M., Ramzan, N., 2017. Fusing highly dimensional energy and connectivity features to identify affective states from eeg signals. Neurocomputing 244, 81–89.

[8] Baveye, Y., Dellandrea, E., Chamaret, C., Chen, L., 2015. Liris-accede: A video database for affective content analysis. IEEE Transactions on Affective Computing 6, 43–55.

[9] Bradley, M.M., Lang, P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry 25, 49–59.

[10] Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution, in: Pattern recognition (ICPR), 2010 20th international conference on, IEEE. pp. 3121–3124.
[11] Cacioppo, J.T., Tassinary, L.G., 1990. Inferring psychological significance from physiological signals. American psychologist 45, 16.
[12] Carrillo, H., Brodersen, K.H., Castellanos, J.A., 2014. Probabilistic performance evaluation for multiclass classification using the posterior balanced accuracy, in: ROBOT2013: First Iberian Robotics Conference, Springer. pp. 347–361.
[13] Chen, J., Hu, B., Moore, P., Zhang, X., Ma, X., 2015a. Electroencephalogram-based emotion assessment system using ontology and data mining techniques. Applied Soft Computing 30, 663–674.
[14] Chen, M., Han, J., Guo, L., Wang, J., Patras, I., 2015b. Identifying valence and arousal levels via connectivity between eeg channels, in: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE. pp. 63–69.
[15] Chen, T., Wang, S., Gao, Z., Wu, C., 2016. Emotion recognition from eeg signals enhanced by user's profile, in: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 277–280.
[16] Chinchor, N., 1992. Muc-4 evaluation metrics, in: Proceedings of the 4th conference on Message understanding, Association for Computational Linguistics. pp. 22–29.
[17] Clerico, A., Tiwari, A., Gupta, R., Jayaraman, S., Falk, T.H., 2018. Electroencephalography amplitude modulation analysis for automated affective tagging of music video clips. Frontiers in computational neuroscience 11, 115.
[18] Coan, J.A., Allen, J.J., 2004. Frontal eeg asymmetry as a moderator and mediator of emotion. Biological psychology 67, 7–50.
[19] Daimi, S.N., Saha, G., 2014. Classification of emotions induced by music videos and correlation with participantsâĂŹ rating. Expert Systems with Applications 41, 6057–6065.
[20] Fan, M., Chou, C.A., 2018. Recognizing affective state patterns using regularized learning with nonlinear dynamical features of eeg, in: 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), IEEE. pp. 137–140.
[21] Forgas, J.P., 1995. Mood and judgment: the affect infusion model (aim). Psychological bulletin 117, 39.
[22] Gao, Z., Wang, S., 2015. Emotion recognition from eeg signals using hierarchical bayesian network with privileged information, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 579–582.
[23] García-Martínez, B., Martinez-Rodrigo, A., Alcaraz, R., Fernández-Caballero, A., 2019. A review on nonlinear methods using electroencephalographic recordings for emotion recognition. IEEE Transactions on Affective Computing .
[24] García-Martínez, B., Martínez-Rodrigo, A., Zangróniz Cantabrana, R., Pastor García, J., Alcaraz, R., 2016. Application of entropy-based metrics to identify emotional distress from electroencephalographic recordings. Entropy 18, 221.
[25] Ghaemmaghami, P., Sebe, N., 2016. Brain and music: Music genre classification using brain signals, in: 2016 24th European Signal Processing Conference (EUSIPCO), IEEE. pp. 708–712.
[26] Goleman, D., 1996. Emotional intelligence. Why it can matter more than IQ. Learning 24, 49–50.
[27] Güntekin, B., Başar, E., 2014. A review of brain oscillations in perception of faces and emotional pictures. Neuropsychologia 58, 33–51.
[28] Gupta, R., Falk, T.H., et al., 2016. Relevance vector classifier decision fusion and eeg graph-theoretic features for automatic affective state characterization. Neurocomputing 174, 875–884.
[29] Hatamikia, S., Maghooli, K., Nasrabadi, A.M., 2014. The emotion recognition system based on autoregressive model and sequential forward feature selection of electroencephalogram signals. Journal of medical signals and sensors 4, 194.
[30] Hemanth, D.J., Anitha, J., et al., 2018. Brain signal based human emotion analysis by circular back propagation and deep kohonen neural networks. Computers & Electrical Engineering 68, 170–180.
[31] Hjorth, B., 1970. Eeg analysis based on time domain properties. Electroencephalography and clinical neurophysiology 29, 306–310.
[32] Jenke, R., Peer, A., Buss, M., 2014. Feature extraction and selection for emotion recognition from eeg. IEEE Transactions on Affective Computing 5, 327–339.
[33] Jia, X., Li, K., Li, X., Zhang, A., 2014. A novel semi-supervised deep learning framework for affective state recognition on eeg signals, in: 2014 IEEE international conference on bioinformatics and bioengineering, IEEE. pp. 30–37.
[34] Jirayucharoensak, S., Pan-Ngum, S., Israsena, P., 2014. Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation. The Scientific World Journal 2014.
[35] Kawde, P., Verma, G.K., 2017. Deep belief network based affect recognition from physiological signals, in: 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), IEEE. pp. 587–592.
[36] Kayser, J., Tenke, C.E., 2006. Principal components analysis of laplacian waveforms as a generic method for identifying erp generator patterns: I. evaluation with auditory oddball tasks. Clinical neurophysiology 117, 348–368.
[37] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I., 2012. Deap: A database for emotion analysis; using physiological signals. IEEE Transactions on Affective Computing 3, 18–31.
[38] Kraljević, L., Russo, M., Sikora, M., 2017. Emotion classification using linear predictive features on wavelet-decomposed eeg data, in: 2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN), IEEE. pp. 653–657.
[39] Kroupi, E., Vesin, J.M., Ebrahimi, T., 2013. Phase-amplitude coupling between eeg and eda while experiencing multimedia content, in: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE. pp. 865–870.
[40] Kuai, H., Xu, H., Yan, J., 2017. Emotion recognition from eeg using rhythm synchronization patterns with joint time-frequency-space correlation, in: International Conference on Brain Informatics, Springer. pp. 159–168.
[41] Lachiche, N., Flach, P.A., 2003. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using roc curves, in: Proceedings of the 20th International Conference on Machine Learning (ICML-03), pp. 416–423.
[42] Lang, P.J., Bradley, M.M., 1999. International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings. Technical Report B-2. The Center for Research in Psychophysiology, University of Florida, FL, USA.
[43] Lang, P.J., Bradley, M.M., Cuthbert, B.N., 2008. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. The Center for Research in Psychophysiology, University of Florida, FL, USA.
[44] Li, X., Song, D., Zhang, P., Zhang, Y., Hou, Y., Hu, B., 2018. Exploring eeg features in cross-subject emotion recognition. Frontiers in

neuroscience 12, 162.
[45] Li, Y., Huang, J., Zhou, H., Zhong, N., 2017. Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks. Applied Sciences 7, 1060.
[46] Liew, W.S., Loo, C.K., Obo, T., 2016. Genetic optimized fuzzy extreme learning machine ensembles for affect classification, in: 2016 Joint 8th International Conference on Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems (ISIS), IEEE. pp. 305–310.
[47] Lin, Y.P., Wang, C.H., Jung, T.P., Wu, T.L., Jeng, S.K., Duann, J.R., Chen, J.H., 2010. Eeg-based emotion recognition in music listening. IEEE Transactions on Biomedical Engineering 57, 1798–1806.
[48] Liu, J., Meng, H., Li, M., Zhang, F., Qin, R., Nandi, A.K., 2018. Emotion detection from eeg recordings based on supervised and unsupervised dimension reduction. Concurrency and Computation: Practice and Experience 30, e4446.
[49] Liu, Y., Sourina, O., 2014. Real-time subject-dependent eeg-based emotion recognition algorithm, in: Transactions on Computational Science XXIII. Springer, pp. 199–223.
[50] Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. Auc: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography 17, 145–151.
[51] Manning, C., Raghavan, P., Schütze, H., 2010. Introduction to information retrieval. Natural Language Engineering 16, 100–103.
[52] Martínez-Rodrigo, A., Alcaraz, R., García-Martínez, B., Zangróniz, R., Fernández-Caballero, A., 2016. Non-lineal eeg modelling by using quadratic entropy for arousal level classification, in: International Conference on Innovation in Medicine and Healthcare, Springer. pp. 3–13.
[53] Menezes, M.L.R., Samara, A., Galway, L., SantâĂŹAnna, A., Verikas, A., Alonso-Fernandez, F., Wang, H., Bond, R., 2017. Towards emotion recognition for virtual environments: an evaluation of eeg features on benchmark dataset. Personal and Ubiquitous Computing 21, 1003–1013.
[54] Mert, A., Akan, A., 2018. Emotion recognition from eeg signals by using multivariate empirical mode decomposition. Pattern Analysis and Applications 21, 81–89.
[55] Mohammadi, Z., Frounchi, J., Amiri, M., 2017. Wavelet-based emotion recognition system using eeg signal. Neural Computing and Applications 28, 1985–1990.
[56] Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning, MIT Press.
[57] Nakisa, B., Rastgoo, M.N., Tjondronegoro, D., Chandran, V., 2018. Evolutionary computation algorithms for feature selection of eeg-based emotion recognition using mobile sensors. Expert Systems with Applications 93, 143–155.
[58] Niemic, C.P., Warren, K., 2002. Studies of emotion: A theoretical and empirical review of psychophysiological studies of emotion. Journal of Undergraduate Research Rochester 1, 15–19.
[59] Özerdem, M.S., Polat, H., 2017. Emotion recognition based on eeg features in movie clips with channel selection. Brain informatics 4, 241.
[60] Padilla-Buritica, J.I., Martinez-Vargas, J.D., Castellanos-Dominguez, G., 2016. Emotion discrimination using spatially compact regions of interest extracted from imaging eeg activity. Frontiers in computational neuroscience 10, 55.
[61] Pantic, M., Rothkrantz, L.J.M., 2000. Automatic analysis of facial expressions: The state of the art. IEEE Transactions on pattern analysis and machine intelligence 22, 1424–1445.
[62] Perrin, F., Pernier, J., Bertrand, O., Echallier, J., 1989. Spherical splines for scalp potential and current density mapping. Electroencephalography and clinical neurophysiology 72, 184–187.
[63] Picard, R.W., Vyzas, E., Healey, J., 2001. Toward machine emotional intelligence: Analysis of affective physiological state. IEEE transactions on pattern analysis and machine intelligence 23, 1175–1191.
[64] Picard, R.W., et al., 1995. Affective computing.
[65] Piho, L., Tjahjadi, T., 2018. A mutual information based adaptive windowing of informative eeg for emotion recognition. IEEE Transactions on Affective Computing .
[66] Powers, D.M., 2015. What the f-measure doesn't measure: Features, flaws, fallacies and fixes. arXiv preprint arXiv:1503.06410 .
[67] Purnamasari, P., Ratna, A., Kusumoputro, B., 2017. Development of filtered bispectrum for eeg signal feature extraction in automatic emotion recognition using artificial neural networks. Algorithms 10, 63.
[68] Putman, P., van Peer, J., Maimari, I., van der Werff, S., 2010. Eeg theta/beta ratio in relation to fear-modulated response-inhibition, attentional control, and affective traits. Biological psychology 83, 73–78.
[69] Russell, J.A., 1980. A circumplex model of affect. Journal of personality and social psychology 39, 1161.
[70] Salovey, P., Mayer, J.D., 1990. Emotional intelligence. Imagination, cognition and personality 9, 185–211.
[71] Schalk, G., McFarland, D.J., Hinterberger, T., Birbaumer, N., Wolpaw, J.R., 2004. BCI2000: a general-purpose brain-computer interface (BCI) system. IEEE Transactions on biomedical engineering 51, 1034–1043.
[72] Shu, Y., Wang, S., 2017. Emotion recognition through integrating eeg and peripheral signals, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 2871–2875.
[73] Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45, 427–437.
[74] Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M., 2011. A multimodal database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing 3, 42–55.
[75] Soleymani, M., Villaro-Dixon, F., Pun, T., Chanel, G., 2017. Toolbox for emotional feature extraction from physiological signals (teap). Frontiers in ICT 4, 1.
[76] Strack, F., Schwarz, N., Chassein, B., Kern, D., Wagner, D., 1990. Salience of comparison standards and the activation of social norms: Consequences for judgements of happiness and their communication. British Journal of Social Psychology 29, 303–314.
[77] Thammasan, N., Fukui, K.i., Numao, M., 2016. Application of deep belief networks in eeg-based dynamic music-emotion recognition, in: 2016 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 881–888.
[78] Torres-Valencia, C., Álvarez-López, M., Orozco-Gutiérrez, Á., 2017. Svm-based feature selection methods for emotion recognition from

multimodal data. Journal on Multimodal User Interfaces 11, 9–23.
[79] Vakkuri, A., Yli-Hankala, A., Talja, P., Mustola, S., Tolvanen-Laakso, H., Sampson, T., Viertiö-Oja, H., 2004. Time-frequency balanced spectral entropy as a measure of anesthetic drug effect in central nervous system during sevoflurane, propofol, and thiopental anesthesia. Acta Anaesthesiologica Scandinavica 48, 145–153.
[80] Van Asch, V., 2013. Macro-and micro-averaged evaluation measures [[basic draft]]. Belgium: CLiPS , 1–27.
[81] Van Rijsbergen, C.J., 1979. Information retrieval. 2nd. newton, ma.
[82] Vateekul, P., Thammasan, N., Moriyama, K., Fukui, K.i., Numao, M., 2015. Item-based learning for music emotion prediction using eeg data, in: Principles and Practice of Multi-Agent Systems. Springer, pp. 155–167.
[83] Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M., Moore, J.H., 2007. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genetic epidemiology 31, 306–315.
[84] Velo, J.R., Stewart, J.L., Hasler, B.P., Towers, D.N., Allen, J.J., 2012. Should it matter when we record? time of year and time of day as factors influencing frontal eeg asymmetry. Biological psychology 91, 283–291.
[85] Verma, G.K., Tiwary, U.S., 2017. Affect representation and recognition in 3d continuous valence–arousal–dominance space. Multimedia Tools and Applications 76, 2159–2183.
[86] van der Vinne, N., Vollebregt, M.A., van Putten, M.J., Arns, M., 2017. Frontal alpha asymmetry as a diagnostic marker in depression: Fact or fiction? a meta-analysis. Neuroimage: clinical 16, 79–87.
[87] Wagh, K.P., Vasanth, K., 2019. Electroencephalograph (eeg) based emotion recognition system: A review, in: Innovations in Electronics and Communication Engineering. Springer, pp. 37–59.
[88] Wang, S., Chen, S., Ji, Q., 2017. Content-based video emotion tagging augmented by usersâĂŹ multiple physiological responses. IEEE Transactions on Affective Computing 10, 155–166.
[89] Wang, S., Zhu, Y., Wu, G., Ji, Q., 2014. Hybrid video emotional tagging using usersâĂŹ eeg and video content. Multimedia tools and applications 72, 1257–1283.
[90] Wang, S., Zhu, Y., Yue, L., Ji, Q., 2015. Emotion recognition with the help of privileged information. IEEE Transactions on Autonomous Mental Development 7, 189–200.
[91] Welch, P., 1967. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Transactions on audio and electroacoustics 15, 70–73.
[92] Wichakam, I., Vateekul, P., 2014. An evaluation of feature extraction in eeg-based emotion prediction with support vector machines, in: 2014 11th international joint conference on computer science and software engineering (JCSSE), IEEE. pp. 106–110.
[93] Wolpaw, J.R., Birbaumer, N., Heetderks, W.J., McFarland, D.J., Peckham, P.H., Schalk, G., Donchin, E., Quatrano, L.A., Robinson, C.J., Vaughan, T.M., 2000. Brain-computer interface technology: a review of the first international meeting. IEEE transactions on rehabilitation engineering 8, 164–173.
[94] Wu, S., Wang, S., Zhu, Y., Gao, Z., Yue, L., Ji, Q., 2016. Employing subjects' information as privileged information for emotion recognition from eeg signals, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE. pp. 301–306.
[95] Xu, H., Plataniotis, K.N., 2016a. Affective states classification using eeg and semi-supervised deep learning approaches, in: 2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP), IEEE. pp. 1–6.
[96] Xu, H., Plataniotis, K.N., 2016b. Eeg-based affect states classification using deep belief networks, in: 2016 Digital Media Industry & Academic Forum (DMIAF), IEEE. pp. 148–153.
[97] Yin, Z., Liu, L., Liu, L., Zhang, J., Wang, Y., 2017a. Dynamical recursive feature elimination technique for neurophysiological signal-based emotion recognition. Cognition, Technology & Work 19, 667–685.
[98] Yin, Z., Wang, Y., Liu, L., Zhang, W., Zhang, J., 2017b. Cross-subject eeg feature selection for emotion recognition using transfer recursive feature elimination. Frontiers in neurorobotics 11, 19.
[99] Yin, Z., Zhao, M., Wang, Y., Yang, J., Zhang, J., 2017c. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. Computer methods and programs in biomedicine 140, 93–110.
[100] Yoon, H.J., Chung, S.Y., 2013. Eeg-based emotion estimation using bayesian weighted-log-posterior function and perceptron convergence algorithm. Computers in biology and medicine 43, 2230–2237.
[101] Zhang, X., Cao, D., Moore, P., Chen, J., Zhou, L., Zhou, Y., Ma, X., 2012. A bayesian network (bn) based probabilistic solution to enhance emotional ontology, in: Human Centric Technology and Service in Smart Space. Springer, pp. 181–190.
[102] Zhang, Y., Zhang, S., Ji, X., 2018. Eeg-based classification of emotions using empirical mode decomposition and autoregressive model. Multimedia Tools and Applications 77, 26697–26710.
[103] Zheng, W.L., Zhu, J.Y., Lu, B.L., 2017. Identifying stable patterns over time for emotion recognition from eeg. IEEE Transactions on Affective Computing .