A mutual information based adaptive windowing of informative EEG for emotion recognition

Laura Piho, and Tardi Tjahjadi, *Senior Member, IEEE*

Abstract

Emotion recognition using brain wave signals involves using high dimensional electroencephalogram (EEG) data. In this paper, a window selection method based on mutual information is introduced to select an appropriate signal window to reduce the length of the signals. The motivation of the windowing method comes from EEG emotion recognition being computationally costly and the data having low signal-to-noise ratio. The aim of the windowing method is to find a reduced signal where the emotions are strongest. In this paper, it is suggested, that using only the signal section which best describes emotions improves the classification of emotions. This is achieved by iteratively comparing different-length EEG signals at different time locations using the mutual information between the reduced signal and emotion labels as criterion. The reduced signal with the highest mutual information is used for extracting the features for emotion classification. In addition, a viable framework for emotion recognition is introduced. Experimental results on publicly available datasets, DEAP and MAHNOB-HCI, show significant improvement in emotion recognition accuracy.

Index Terms

EEG, human emotions, mutual information, entropy, data reduction

1 Introduction

Emotion related studies are popular in neuroscience and psychology, and during the recent decade they have gathered research interest in computational science (e.g., [1], [2], [3]). Traditionally, work regarding emotion recognition use facial images (e.g., [4], [5]), speech (e.g., [6], [7]), or both modalities (e.g., [9], [10]). In recent years, emotion recognition using electroencephalogram (EEG) signals has become popular [11]. Such an approach consists of two main tasks: data collection, and analysis of the collected data. This paper concentrates on the second task.

Recording EEG signals is a non-invasive method for acquiring emotion data from the brain. Furthermore, EEG signals have good temporal resolution, making it a good option for emotion recognition. The stimuli used to evoke emotion responses are usually audio (e.g., [12]) or video (e.g., [13], [14]). In this paper, publicly available datasets DEAP [13] and MAHNOB-HCI [14] are used, where DEAP is the largest dataset for EEG-based emotion recognition.

The raw EEG signals have amplitude ranging between $10\mu V$ and $100\mu V$, and include a high level of noise from different sources, e.g., eye blinking, and muscular and vascular effects, etc. [15], [16], [17]. Both DEAP and MAHNOB-HCI recordings involved 32 active AgCl electrodes placed according to the international 10-20 system [13], [14]. In addition, the database includes participant self-assessment and frontal face videos of some of the participants.

Emotion recognition using EEG signals can be subject-dependent or subject-independent [18]. For the purpose of this paper, the subject-dependent approach has been chosen. Generally, subject-dependent approach achieves higher accuracy, but is slower as a classifier has to be trained for each subject.

The motivation of this paper is twofold. Firstly, the EEG dataset often includes recordings over long periods of time making the signal processing computationally costly. Secondly, there exist difficulties when dealing with noise in EEG signals, especially problematic is accurately identifying the signal components related to emotions and those related to other brain activities. In addition, emotions are subjective, and their recognition depend on their intensity evoked by the stimuli. This makes establishing ground truth difficult. Another challenge is removing any noise in the signal without removing the emotion-related EEG signal features.

To overcome these problems, this paper introduces a subject-dependent mutual information-based windowing method for extracting informative EEG features for robust and accurate classification of the associated emotions. The DEAP dataset includes recordings where 60s of music videos are used for stimuli. For each recording the first five seconds correspond to the baseline and the subsequent EEG recording is 60s long. The MAHNOB-HCI dataset uses video clips as stimuli. The EEG recordings are of varying length, with the shortest being 34.9s and the longest 117s. Assuming that during the length of EEG recording the intensity of the subject's emotion changes, the aim of the windowing method is to search for the most informative part of the signal for emotion recognition.

The contribution of this paper is a framework for subject dependent EEG-based emotion recognition using reduced signals via adaptive windowing. Traditional emotion recognition framework consists of pre-processing, feature extraction

• *L. Piho and T. Tjahjadi are with the School of Engineering, University of Warwick,Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom. E-mail: T.Tjahjadi@warwick.ac.uk*

Manuscript received xxxx; revised xxxx.

and selection, and classification steps. It is shown that reducing the signal length after pre-processing lowers the computational cost of the subsequent steps, i.e., feature extraction, selection and classification. Whilst the overall cost of the method is higher using windowing method to reduce the signal length, so is the accuracy. Furthermore, by providing the comparison of the emotion recognition performances on the use of different feature extraction and classification methods on different datasets, the effectiveness of these methods is determined. Finally, it is shown that the proposed framework outperforms four existing emotion recognition systems.

This paper is organised as follows. Section 2 presents the related work organised under signal pre-processing, feature extraction and selection, and classification. Section 3 presents the proposed windowing method and additional processes required for the emotion classification. Illustrative examples using DEAP and MAHNOB-HCI datasets together with comparison with existing methods are presented in Section 4. The conclusion follows in Section 5.

2 Related Work

There are several statistical and machine learning methods for EEG signal analysis. The processing from raw signals to classified emotions may be grouped into four tasks [15]: signal pre-processing, feature extraction, feature selection, and classification.

2.1 Signal pre-processing

The pre-processing of raw EEG data for further analysis involves the use of digital signal processing techniques. Noting that useful data for emotion recognition are at frequencies between 4-45 Hz [15], [22], for efficient pre-processing the data is downsampled to 128 Hz and bandpass filtered with a common bandwidth [12], [15], [19]. In common average reference (CAR) the value of the entire electrode montage is subtracted from the channel of interest resulting in a spatial voltage distribution of mean zero [20]. As it emphasizes components that are present in most electrodes, it reduces such components and thereby functions as a high-pass filter [20], [21].

The most challenging task of signal pre-processing is to remove noise from EEG without distorting the signals related to emotions. Noise sources, e.g., external and environmental noise from the EEG equipment, and electro-magnetic (EM) noise, are easy to address by ensuring the equipment is in good working order and removing EM sources from the recording room [23]. A more challenging task is to remove physiological noise, such as due to cardiac signals, movements caused by muscle contraction (electromyogram (EMG)), and signal caused by eyeball movement (electrooculogram (EOG)) [23]. Noise due to EMG can be minimised by asking the subjects to sit in a comfortable position, and noise due to EOG can be minimised by using stimuli that do not require eye movement [23]. However, it is not possible to completely avoid the aforementioned noise sources. In addition, using stimuli that do not invoke eye movement is not always practical, as they trigger a strong emotional response. Hence, a compromise between avoiding noise due to EOG and having good emotional stimuli has to be found.

To make noise removal easier, some datasets include additional physiological signals like EOG, EMG, plethysmograph, body temperature and measurement from respiration belt. There are two main approaches to EOG artefact correction, namely regression based methods, and methods based on spatial decomposition. An extensive review of EOG artefact removal methods for signals with and without prior knowledge is presented in [24].

The regression based methods smooth EEG by regressing out the reference EOG signals, e.g., [25], [26], [27], [28]. These require the recordings of EOG signals. In addition, several muscle groups require different reference channels, which makes this approach impractical [29].

The spatial decomposition based methods include principal component analysis (PCA), singular value decomposition (SVD), and blind source decomposition (BSS) such as independent component analysis (ICA). Similar to regression methods, EOG artefacts need to be known. PCA has been shown in [30] to be better in removing EOG artefacts than regression methods [29]. In [29], ICA has been shown to recover more brain activity signals than PCA, and therefore ICA based methods are more favourable.

In addition, one can use fully automatic methods, where artefacts are removed without any prior knowledge of the signals, e.g., the second order blind identification [24] which involves the ICA.

2.2 Feature extraction and feature selection

EEG signals are high dimensional, hence the computational processing of these signals is often complex and expensive. The purpose of feature extraction is to simplify the subsequent emotion classification task by identifying the important elements of the signal, creating a feature vector based on these elements and using the vector to classify the corresponding emotion. Available methods to extract features include wavelet transforms, higher order spectrum and higher order crossings (HOC). For a detailed overview of existing feature extraction methods, the reader is directed to [31] and [32].

The number of features extracted is largely dependent on the feature extraction method, and therefore it is helpful to identify the most relevant set of features to enable selection of the most appropriate feature extraction method for emotion recognition. In addition, feature selection is useful for reducing the dimensionality of the EEG signals. Existing feature selection methods include minimum redundancy maximum relevance (mRMR), Relief, and differential evolution feature

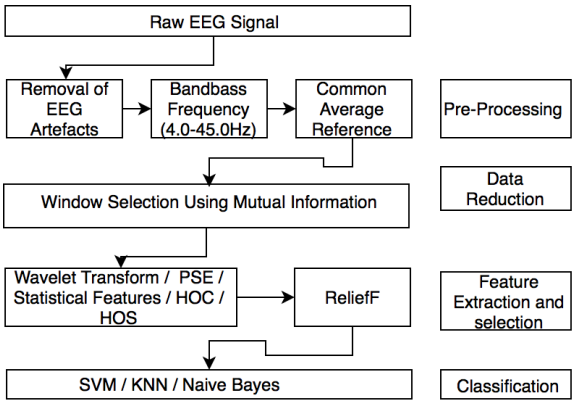

Figure 1: The proposed emotion recognition framework.

selection. Genetic algorithm and support vector machine (SVM) can also be used. More information about these methods can be found in [33], [34], [35].

Since there is no one feature extraction method that is universally accepted as best for EEG, in this paper a few more common feature extraction methods will be used and compared. Multivariate feature extraction methods have been shown to work slightly better in general than univariate methods [32].

2.3 Classification

Multiple classification methods can be considered for EEG-based emotion recognition. If the datasets include subject responses, then supervised classification methods can be used. It should be noted that emotions are subjective, and the subjects' evaluation of their emotions may not be the ground truth, and thus some correction of the emotion labels needs to be considered. In the absence of subject evaluations, it is appropriate to consider unsupervised classification.

The classification of EEG dataset into discrete sets for different brain-computer interface tasks can be achieved using for example nonlinear Bayesian classifiers [36], [37], neural networks [38] or SVMs [39], [40]. In this paper, multiple classification methods will be considered, namely SVM, naive Bayes (NB) classifier, and K-nearest neighbours (KNN). The overview, comparison, and guidelines on how to choose between EEG classification methods can be found in [36], [43]. Detailed overview of the multiclass SVM can be found in [41], and the use of multiclass SVM for EEG classification in [42].

3 Proposed Framework

The proposed emotion recognition framework is shown in Fig. 1. It consists of four processes: pre-processing, data reduction, feature extraction and selection, and classification. The proposed adaptive windowing method is experimented with a few feature extraction methods and the accuracy of the resulting emotional recognition compared.

3.1 Pre-Processing

Consider the raw EEG data, $\mathbf{X}$. Matlab automatic artefact removal (AAR) toolbox [28], [50] is used for EOG artefact removal as follows. First, using BSS, $\mathbf{X}$ is decomposed into spatial components with the aim of separating the artefacts due to cerebral activity. Second, artefact-related components are detected. Finally, the EEG data is reconstructed using only non-artefactual components [50].

The AAR toolbox was chosen for pre-processing because it is constructed in a way that if necessary, its methods can be easily adapted and extended. The data is downsampled to have a sampling rate of 128 Hz and averaged to CAR to reduce the components related to noise that are present in signals from a large proportion of the electrodes.

3.2 Data Reduction: Adaptive Windowing

Using all available EEG data is computationally expensive, and often will not give a viable emotion recognition. It has been shown that mutual information is a good criterion for measuring the importance of EEG based information [15], and has been used in feature extraction [15], [44]. These motivated us to develop an adaptive windowing which finds the most informative window of the pre-processed signal using mutual information for the subsequent emotion recognition. It is noted that EEG signals that correspond to emotions are noisy. Furthermore, the stimuli to evoke emotions is lengthy, e.g., the DEAP dataset uses a stimuli of one minute duration. During this time, a person can experience multiple emotions of different strength even though the stimuli has been tailored to evoke one emotion. This motivated us to find a window of short duration which extracts signal that better represents the emotion.

Consider a dataset consisting of N samples, of length M_j, $j = 1, .., N$, together with labels $\mathbf{y} = (\mathbf{y_1}, ..., \mathbf{y_N})$, where $y_i \in [1, ..., C]$, and C denotes the number of classes. The adaptive windowing method for data reduction, Algorithm 1,

works as follows. First, the maximum and minimum window size, and the change constant are chosen, and denoted as W_{max}, W_{min}, and c, respectively. Next, the window size is set to be W_{min}, and all possible combinations of signals of size W_{min} are found. That is, consider a new data matrix $\mathbf{X}'_{(i,W_{min})}$ of size $N \times W_{min}$, where $i = 1, 2, ...K_{W_{min}}$, and $K_{W_{min}}$ is the number of different possible reduced data matrices with signal length of W_{min}.The mutual information between $\mathbf{X}'_{(i,W_{min})}$ and $\mathbf{y}$, is

$$MI_{i,W_{min}} = I(\mathbf{X}'_{(i,W_{min})}, \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{X}'_{(i,W_{min})}) \tag{1}$$

where

$$H(\mathbf{y}) = -\sum_{i=1}^{N} p(y_i) \log p(y_i) \tag{2}$$

$$H(\mathbf{y}|\mathbf{X}) = -\sum_{j=1}^{N} \frac{1}{N} \sum_{i=1}^{C} p(y_i|\mathbf{x}_j) \log p(y_i|\mathbf{x}_j). \tag{3}$$

The conditional probability, $p(.)$, is estimated using Parzen Window density estimation, i.e.,

$$p(y|\mathbf{x}) = \frac{\sum_{\mathbf{k} \in \mathbf{K^y}} \exp(-\frac{(\mathbf{x}-\mathbf{x_k})\mathbf{\Sigma_X^{-1}}(\mathbf{x}-\mathbf{x_k})}{\mathbf{2h^2}})}{\sum_{\mathbf{i=1}}^{\mathbf{C}} \sum_{\mathbf{k} \in \mathbf{K^i}} \exp(-\frac{(\mathbf{x}-\mathbf{x_k})\mathbf{\Sigma_X^{-1}}(\mathbf{x}-\mathbf{x_k})}{\mathbf{2h^2}})}, \tag{4}$$

where h is the width of the window.

Algorithm 1 Determine signal window using mutual information

Input: $\mathbf{X}$ - Data consisting of N samples of size M_j, $j = 1, ..., N$. $\mathbf{y}$ - the emotional labels. W_{min} - minimum window size. W_{max} - maximum window size. c - change.
$f =$ sampling rate.
$W = W_{min}$
while $W \leq W_{max}$ **do**
$\mathbf{X}' \in \mathbf{X}^*_{\mathbf{W}}$, where

$$X^*_W = \{\mathbf{X}'|\mathbf{X}' = \mathbf{X}(:, \mathbf{a_j} : \mathbf{b_j}),$$
$$\forall \quad b_j - a_j = W, b_j \leq N, \text{ and } a_{j+1} = a_j + f\}$$

for $i = 1 : K_W, K_W = size(X^*)$ **do**
$MI_{(i_W,W)} = I(\mathbf{X}'_i, \mathbf{y})$
end for
$W = W + c$
end while
$M_{q_{MI},W_{MI}} = max(MI)$, where W_{MI} is the window size with highest average mutual information, and q_{MI} contains information about best window of size W_{MI} for all data samples.
$\tilde{\mathbf{X}} = \mathbf{X}'_{q_{MI}}$, such that $\mathbf{X}'_{q_{MI}} \in X^*_{W_{MI}}$

Next, the size of the window is increased by the change constant c. Similarly, all possible combinations of signals of size $W_{min} + c$ are found and the mutual information $MI_{(i,W_{min}+c)}$ between new data matrices $\mathbf{X}'_{(i,W_{min}+c)}$ and emotional labels $\mathbf{y}$ calculated, $i = 1, 2, ...K_{(W_{min}+c)}$. This process is repeated until the window size is greater than or equal to W_{max}. Iterating this process assures that all possible signal time locations are considered.

Choosing the reduced signal matrix is achieved in two steps. First, for the sake of simplicity later on, the window size is chosen to be uniform over all data samples. That is, from all tested window sizes the one with the highest average mutual information is chosen and denoted as W_{MI}. Second, the data matrix, $\mathbf{X}'_{q_{MI}}$, where $q_{MI} \in [1, ..., K_{W_MI}]$, for which has the length W_{MI} and the highest mutual information is identified as the reduced data matrix. The reduced data matrix with the highest mutual information is assumed to consist of signals with the greatest emotion intensity, and is chosen for further analysis. The new data matrix is given by

$$\tilde{\mathbf{X}} = \mathbf{X}'_{q_{MI}} | \mathbf{X}'_{q_{MI}} \in X^*_{W_{MI}}. \tag{5}$$

3.3 Feature Extraction

The feature extraction methods chosen for this paper have been shown to give good results in studies that used EEG signals [33], [46], [49], [61]. Also, the features have been extracted from five frequency bands using wavelet transform (WT). To extract features one of the four feature extraction techniques was used: statistical features (SF), power spectral entropy (PSE), higher order crossing (HOC), and higher order spectral (HOS).

3.3.1 Wavelet transform (WT)

WT [45] maps one dimensional signal to a two-dimensional function by decomposing a signal as a superposition of simple units from which the original signal can be reconstructed. WT is a spectral estimation technique where all functions are expressed as an infinite series of wavelets. The decomposition of the EEG signal via WT leads to a set of wavelet coefficients which represents its energy distribution in time and frequency.

The bandlimited EEG signal will be decomposed through "db4" wavelet up to five levels. This wavelet is chosen due to its near optimal time-frequency localisation property [46]. The decomposition levels A5, D5, D4, D3, D2 and D1 correspond to delta (δ, $0-4$ Hz), theta (θ, $4-8$ Hz), alpha (α, $8-14$ Hz), beta (β, $14-32$ Hz), gamma (γ, $32-64$ Hz), and noise ($64-128$ Hz) frequency bands [46].

3.3.2 Statistical Features (SF)

Using statistical measures for feature extraction involves mean (μ), standard deviation (σ), first and second order differences (Δ and Γ, respectively), and normalised first and second order differences (i.e., $\bar{\Delta}$ and $\bar{\Gamma}$, respectively). These measures are calculated for each frequency bands of all EEG signals s_i , $i = 1, ..., N$. For a single signal, the features can be ordered and written as

$$\begin{aligned} FV_{s_i} = \quad & [\mu_{\delta_{s_i}}, \sigma_{\delta_{s_i}}, \Delta_{\delta_{s_i}}, \bar{\Delta}_{\delta_{s_i}}, \Gamma_{\delta_{s_i}}, \bar{\Gamma}_{\delta_{s_i}}, ... \\ & \mu_{\gamma_{s_i}}, \sigma_{\gamma_{s_i}}, \Delta_{\gamma_{s_i}}, \bar{\Delta}_{\gamma_{s_i}}, \Gamma_{\gamma_{s_i}}, \bar{\Gamma}_{\gamma_{s_i}}] \ . \end{aligned} \tag{6}$$

Note that the order of frequency bands from which the features are extracted are for all signals δ, θ , α, β, and γ. Therefore, for each subject the frequency vector that includes all signals can be written as

$$FV = \quad [FV_{s_1}, FV_{s_2}, ..., FV_{s_N}] \ . \tag{7}$$

3.3.3 Power Spectral Entropy (PSE)

Another method that is used to extract features from signals is PSE [61]. PSE can be calculated by first calculating the power spectral density $\hat{P}(\omega_i)$ using the discrete Fourier Transform of the signal, where ω_i is the frequency variable. Next, the PSD is normalised to obtain PSD distribution function

$$p_i = \frac{\hat{P}(\omega_i)}{\sum_i \hat{P}(\omega_i)}. \tag{8}$$

Finally, as information entropy can be given by

$$H = -\sum_{i=1}^{n} p_i \ln p_i \ , \tag{9}$$

the PSE is found by substituting equation (8) in (9).

3.3.4 Higher Order Spectral (HOS)

HOS are spectral representations of higher order moments, or cumulants of a signal. There are several reasons for using HOS in signal processing. For example, to suppress Gaussian noise when its mean and variance are unknown, to reconstruct the phase and magnitude response of signals, and to find and characterise the nonlinearities of the signal. Decomposing signals using WT into different frequency bands and analysing the decomposed signals using HOS give information in multiple scales, which has been shown to provide accurate assessment of emotional stress [33].

EEG signal processing uses third order correlation, i.e., bispectrum. Consider a signal $x(t)$ with a discrete Fourier transform [47] evaluated on N data points, i.e.,

$$X(f) = \sum_{t=0}^{t=N-1} x(t) \exp^{j2\pi fn} \ , \tag{10}$$

where f is the frequency variable. The Fourier transform of the bispectrum of a signal is [33]

$$Bis(f_1, f_2) = E[X(f_1)X(f_2)X^*(f_1 + f_2)] \ , \tag{11}$$

where $X^*(f)$ denotes the complex conjugate of $X(f)$, and $E[.]$ is the statistical expectation operator. The normalised bispectrum, i.e., bicoherence, is [33]

$$Bic(f_1, f_2) = \frac{Bis(f_1, f_2)}{\sqrt{P(f_1)P(f_2)P(f_1 + f_2)}} \ , \tag{12}$$

where the power [33], [48]

$$P(f) = E[X(f)X*(f)] \ . \tag{13}$$

The following five features are computed: sum of the bispectrum magnitudes (f^1), sum of the squares of the bispectrum magnitudes (f^2), sum of the bicoherence magnitudes (f^3), sum of the squares of the bicoherence magnitudes (f^4), and test

of Gaussianity (f^5). These features are computed for five frequency bands, namely γ, β, α, θ, and δ. These are arranged into the following feature vector,

$$\begin{aligned} FV_{HOS} = \quad & [f^1_{\gamma,s_1}, f^2_{\beta,s_1}, f^3_{\alpha,s_1}, f^4_{\theta,s_1}, f^5_{\delta,s_1}, ..., \\ & f^1_{\gamma,s_N}, f^2_{\beta,s_N}, f^3_{\alpha,s_N}, f^4_{\theta,s_N}, f^5_{\delta,s_N}], \end{aligned} \tag{14}$$

where the subscript pair respectively denotes the wavelet frequency band and signal from an electrode, and the superscript denotes the different HOS features extracted.

3.3.5 Higher Order Crossing (HOC)

Simple HOC applies a sequence of high-pass filters to the zero-mean time series $X(t)$, [49]

$$\mathfrak{T}_k\{X(t)\} = \nabla^{k-1}X(t)\ , \tag{15}$$

where ∇ is the iterative difference operator. We use $\nabla \equiv X(t) - X(t-1)$, and $k = 1, ..., L$, where L is the number of filters. The HOC sequence D_k, i.e., the resulting k features, comprises the number of zero-crossings of the filtered time series by counting its sign changes, i.e.,

$$T_k\{X(t)\} = \quad \sum_{j=1}^{k} \begin{pmatrix} k-1 \\ j-1 \end{pmatrix} (-1)^{j-1} X(t-j+1). \tag{16}$$

We construct a binary time series

$$Y_t(k) = \begin{cases} 1 & \text{if } \mathfrak{T}_k\{X(t)\} \geq 0 \\ 0 & \text{if } \mathfrak{T}_k\{X(t)\} < 0 \end{cases}, \quad k = 1, 2, ...; t = 1, ..., N. \tag{17}$$

Hence, the simple HOC is estimated by counting the symbol changes in binary time series $Y_t(k)$, giving the feature vector

$$V_{HOC} = [D_1, ..., D_L]\ , \tag{18}$$

where $D_k = \sum_{t=2}^{N} [Y_t(k) - Y_{t-1}(k)]^2$. The different HOC features are computed to represent the oscillatory patterns present in the EEG data.

3.4 Feature Selection

The extraction of features from 32 signals using 5 wavelet bands resulted in 1120 SF features, 160 PSE features, 800 HOS features, and 320 HOC features. Due to the limited number of data points in the datasets available for EEG emotion recognition, the number of features is significantly higher than the number of data points, resulting in model over-fitting. To overcome over-fitting, feature selection can be used to reduce the number of features used to train the model, where the aim of the feature selection algorithm is to find a new set of the most informative features. The rule of thumb in feature selection is to make the number of features fewer than the number of observations.

The feature selection method considered in this work is the ReliefF algorithm [55], which is an extension of Relief algorithm [56]. Relief is not dependent on heuristics, runs in low-order polynomial time, and is noise-tolerant and robust to feature interactions. It is simple and has low computational time. However, it does not behave well with small set of training instances. To address this, ReliefF runs the outer loop of the algorithm over all available training instances [55]. In addition, it can be extended to the multi-class problem.

In our feature selection method, ReliefF is applied to select features by first selecting an instance and finding k near misses and hits. The hits are instances corresponding to the same class, and misses are instances from different classes. These are used to calculate the weight vector corresponding to the quality of features, based on their feature values, near hit, and near misses [55]. Finally, using the weight vector, the features with the highest quality are chosen so that the number of selected features is fewer than the number of samples.

3.5 Classification

In our emotion recognition framework, emotion classification is achieved using SVM, KNN and NB.

3.5.1 Support Vector Machine (SVM)

SVM [39] is a binary classifier which can be extended into a multiclass classifier. It is chosen due its high-generalisation ability, and it has been shown to work well for classification. Consider a training set $(\mathbf{x}_j, y_j),\ 1 \leq j \leq N$, where $\mathbf{x_j}$ denotes the feature vectors extracted from EEG signals, y_j denotes the corresponding emotion labels, and N is the number of data.

The SVM decision function can be written as

$$f(\mathbf{x}) = \sum_{\mathbf{i}}^{\mathbf{N}} \alpha_{\mathbf{i}} \mathbf{y_i} \mathbf{k}(\mathbf{s_i}, \mathbf{x}) + \mathbf{b}, \tag{19}$$

where $\mathbf{x}$ is the input vector (in this case the feature vector extracted from EEG signals), k is the kernel function, s_i denotes support vectors, α_i are the weights and b is the bias. In the scope of this paper, we used the Gaussian kernel.

To train the SVM, weights α_i are found for existing data such that

$$f(\mathbf{x_i}) = \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}, \tag{20}$$

where $+1$ and -1 denote positive and negative emotion classes, respectively.

3.5.2 K-Nearest Neighbours (KNN)

KNN has been shown to work well with EEG signals in [46]. The classification is based on user-defined constant integer k, where new case will be assigned to the class most common amongst its k nearest neighbours measured by a distance metric. Most commonly, Euclidean distance is used as the distance metric, but Manhattan, Minkowski, and Hamming distances can also be used. The problem with KNN classification is that training imbalanced sets may result in the classes with more examples dominating the classification.

For the purpose of this paper, MATLAB inbuilt function *fitcknn* is used to fit KNN model to the data. This function attempts to minimize the cross-validation loss for the *fitcknn* by varying its parameters, including the number of neighbours and distance metric depending on the dataset. The distance metrics available to this function are City block distance, Chebychev distance, Minkowski distance, Euclidean and standardised Euclidean distance, Hamming distance, Jaccard coefficient, Mahalanobis distance, and Spearman's rank correlation.

3.5.3 Naive Bayes (NB)

A NB classifier [57], [58] assumes all input variables are independent. It aims to find the conditional probability that data points belong to a specific class given the input features and chooses the class with the highest probability. Thus, the goal of NB classifier is to find the probability $p(C|F_1, ..., F_n)$, where C is the class variable and $F_1, ..., F_n$ are the data points. This probability is difficult to compute, and thus the Bayes theorem is used instead. Furthermore, all input variables F_i are assumed to be independent.

Hence, the conditional probability that a given data point belong to a specific class given the input features can be written as

$$p(C|F_1, ..., F_n) = \frac{p(C)\prod_{i=1}^{n} p(F_i|C)}{p(F_1, ..., F_n)} \tag{21}$$

$$= \frac{1}{Z}p(C)\prod_{i=1}^{n} p(F_i|C), \tag{22}$$

where Z is a scaling factor dependent on $F_1, ..., F_n$. Despite the assumption, a NB classifier still performs surprisingly well even when the assumption is not entirely accurate.

3.6 Complexity Analysis

Table 1: Reduced data results for valence using DEAP dataset.

# of Features		30	31	32	33	34	35	36	37	38	39	AVG	MAX
HOC	SVM	79.06%	78.89%	79.72%	82.81%	80.39%	79.20%	78.89%	77.87%	78.26%	80.89%	79.60%	82.81%
	KNN	**84.38%**	**84.92%**	**84.06%**	**83.75%**	82.66%	**84.69%**	**84.14%**	**84.30%**	81.80%	82.73%	**83.74%**	**84.92%**
	NB	82.03%	82.73%	82.50%	82.73%	**83.44%**	82.27%	82.66%	82.73%	**82.97%**	**82.89%**	82.70%	83.44%
SF	SVM	85.55%	85.70%	85.94%	85.55%	84.92%	85.08%	86.56%	86.09%	87.03%	87.11%	85.95%	87.11%
	KNN	**89.61%**	**89.61%**	**89.06%**	**89.38%**	**88.98%**	**88.05%**	**88.98%**	**88.20%**	**88.59%**	**89.06%**	**88.95%**	**89.61%**
	NB	86.02%	86.48%	86.95%	86.95%	87.19%	87.03%	86.95%	87.42%	86.88%	87.11%	86.90%	87.42%
PSE	SVM	69.92%	71.88%	71.17%	70.78%	70.39%	72.73%	71.25%	70.08%	70.78%	71.09%	71.01%	72.73%
	KNN	**76.17%**	73.44%	**75.16%**	**74.84%**	72.19%	**75.63%**	**74.53%**	**74.45%**	73.20%	72.58%	74.22%	**76.17%**
	NB	75.39%	**74.61%**	74.38%	74.61%	**74.77%**	73.75%	74.38%	74.06%	**73.91%**	**73.91%**	**74.38%**	75.39%
HOS	SVM	75.94%	74.49%	78.16%	77.19%	79.22%	77.66%	77.58%	77.97%	77.58%	76.09%	77.19%	79.22%
	KNN	**83.52%**	**82.50%**	**82.73%**	**83.98%**	**81.87%**	**83.05%**	**82.66%**	**84.14%**	**82.89%**	**81.80%**	**82.91%**	**84.14%**
	NB	81.25%	81.72%	81.64%	81.17%	81.64%	81.48%	81.88%	81.80%	80.86%	80.94%	81.44%	81.88%

One of the major problems with EEG-based emotion recognition system is that it is computationally very complex, and therefore the required training time is long. The computational complexity is $H(\mathbf{y}|\mathbf{X}) \propto (\mathbf{n^2} \times \mathbf{d})$, where $\mathbf{y}$ is the vector containing information about the dataset classes, $\mathbf{X}$ is the dataset matrix of size $n \times d$, n is the number of samples and d is the dimensionality of the feature vector. Noting that for each EEG sample there are 32 electrodes, the computational complexity is increased further by a factor of 32. Thus, decreasing the size of the dataset in this step decreases the computational cost of the subsequent steps.

Computational complexity varies for different feature extraction methods. The computational complexity of WT using discrete wavelet transform is $O(N)$, where N is the size of the signal. Using HOS, namely bispectrum, invokes a

Table 2: All data results for valence using DEAP dataset.

# of Features		30	31	32	33	34	35	36	37	38	39	MEAN	MAX
HOC	SVM	75.68%	75.15%	71.87%	72.09%	68.20%	70.35%	75.92%	71.68%	70.82%	73.36%	72.51%	75.92%
	KNN	**77.03%**	**77.73%**	**77.11%**	**78.05%**	**77.19%**	76.09%	**78.44%**	76.72%	**77.58%**	**77.03%**	**77.30%**	**78.44%**
	NB	76.33%	76.09%	75.78%	76.17%	76.17%	**76.33%**	76.25%	**76.80%**	75.94%	76.25%	76.21%	76.80%
SF	SVM	80.00%	81.17%	80.23%	81.02%	81.48%	80.78%	79.38%	80.70%	79.38%	80.47%	80.46%	81.48%
	KNN	**81.41%**	**83.98%**	**81.88%**	**82.42%**	**82.81%**	**82.66%**	**84.38%**	**82.66%**	81.64%	**83.75%**	**82.76%**	**84.38%**
	NB	80.78%	80.94%	81.48%	81.33%	81.95%	81.64%	82.58%	82.42%	**82.50%**	83.28%	81.89%	83.28%
PSE	SVM	67.03%	69.14%	68.13%	66.72%	67.42%	67.27%	65.94%	66.64%	68.13%	67.27%	67.37%	69.14%
	KNN	**75.86%**	**72.66%**	**72.73%**	**75.78%**	**72.58%**	**74.14%**	**72.27%**	**73.05%**	**74.69%**	**71.88%**	**73.56%**	**75.86%**
	NB	71.80%	71.88%	71.02%	70.63%	70.78%	69.84%	70.23%	70.31%	68.52%	69.53%	70.45%	71.88%
HOS	SVM	75.89%	71.19%	75.55%	75.98%	75.76%	75.78%	75.55%	75.54%	72.86%	73.67%	74.78%	75.98%
	KNN	79.14%	78.52%	77.89%	78.20%	**79.77%**	76.48%	78.28%	77.66%	**78.91%**	**79.84%**	78.47%	**79.84%**
	NB	**79.53%**	**79.22%**	**79.77%**	**79.45%**	79.69%	**79.61%**	**79.14%**	**78.98%**	78.83%	79.53%	**79.38%**	79.77%

computational cost of $O(N^2)$. For our emotion recognition framework, there is a need to compute bispectrum multiple times, which results in large computation complexity. In addition, to calculate bicoherence, there is a need to compute PSD, which has the computational complexity of $O(N)$ for a window size of N. Therefore, reducing the size of the dataset reduces the computational cost of these methods. On the other hand, HOC and statistical feature extraction have lower computational complexity, namely $O(N)$.

4 EXPERIMENTS

The experiments conducted with the reduced data include the methods in Section 3.3, and include the use of statistical features, PSD, HOS and HOC. Furthermore, to validate emotion recognition using reduced data, all studies were also run using non-reduced data for comparison. Similarly to [15], in this paper the emotion recognition using EEG will be considered as a subject dependent problem.

To calculate the error, leave-on-out cross validation (LOOCV) was used. The reason to use LOOCV comes from the limited data availability. That is, the sets used for training are relatively small to split into training and testing sets. When small datasets are used, the LOOCV is a good alternative to use for model validation. In addition, leave-on-out cross validation is suggested to avoid over fitting. It will be emphasised, that when using LOOCV, the data was split into training and test set repeatedly, where the test set is unseen by the model each time. Furthermore, Receiver Operating Characteristic (ROC) curves were plotted to show the trade off between selectivity and sensitivity.

4.1 Dataset

4.1.1 DEAP

DEAP dataset [13] includes recordings from 32 participants. Each participant was asked to look at 40 music videos and their EEG signals were recorded at 32 electrodes set according to the international 10-20 system [52]. In addition, the dataset includes recordings from 12 peripheral channels, 3 unused channels and 1 status channel. In some cases, the data from peripheral channels can be used for emotion recognition. For the purpose of this study, the peripheral channels were used to remove noise from EEG channels, and only EEG signals were used for classifying emotions. The DEAP dataset includes pre-processed dataset, but for the purpose of this paper, the raw data was used and pre-processed separately as in Section 3.1.

To label the data as corresponding to an emotion, participant self-assessment was performed by each participant at the end of each trial. For this assessment, the valence-arousal-dominance scale [53] was used, where the scales range from unhappy/sad to happy/joyful for valence, calm/bored to stimulated/excited for arousal, and submissive to dominant. In addition, participants rated the videos according to their liking as follows. For each video, the participant rated the valence, arousal, dominance and liking on a continuous scale between 1 and 10.

4.1.2 MAHNOB-HCI

MAHNOB-HCI [14] database consists of two experiments, emotion recognition and implicit tagging. For the purpose of this paper, only stimulated emotion recognition data was used. The emotion recognition experiment has recordings from 30 participants, from which 25 participants had complete EEG recordings. EEG signals were recorded of a participant watching 20 video clips, and at 32 electrodes set according to the international 10-20 system. The raw data was pre-processed as in Section 3.1.

Participants self-assessed the videos and were asked to rate their valence, arousal, and dominance on a nine-point scale. In addition, participants were asked to give emotional labels. The labels included neutral, anxiety, amusement, sadness, joy, disgust, anger, surprise, and fear. For the purpose of consistency, experiments were performed using valence and arousal results. The labels were split into two classes, positive and negative.

4.2 Results

There are a few things to note regarding EEG emotion recognition. First, the EEG signals are subject dependent. This will be taken into account and classifiers will be trained separately for all participants. Second, the available number of trials to train the classifier for each participant is limited. The number of trials for each participant is 40 in DEAP and 20 in MAHNOB. With small dataset, it is more difficult to avoid over-fitting. To overcome this, the number of selected features is smaller than the number of trials. For DEAP dataset the number of features has been selected to be between 30 and 39 features, and for MAHNOB dataset between 10 and 19 features.

In MATLAB Statistics and Machine Learning Toolbox [59] the hyperparameter optimisation has been included for SVM, KNN, and NB MATLAB functions. This toolbox was used with the aim of finding suitable hyperparameters for the problem. Furthermore, the results were validated using leave-one-out cross validation for all classifiers. The leave-one-out cross validation, reserves one observation as validation data, and trains the model using the remaining observations. This will be performed for all observations. The two-class classification was performed using only valence and arousal, and in both dimensions separately.

4.2.1 DEAP dataset

Table 3: Reduced data results for arousal using DEAP dataset.

# of Features		30	31	32	33	34	35	36	37	38	39	AVG	MAX
HOC	SVM	78.24%	76.63%	80.00%	78.88%	77.40%	77.91%	76.39%	75.13%	75.39%	77.94%	77.39%	80.00%
	KNN	**82.03%**	**81.72%**	**81.64%**	**81.80%**	**80.86%**	**80.70%**	**81.56%**	**80.78%**	**81.09%**	**81.80%**	**81.40%**	**82.03%**
	NB	77.66%	77.34%	77.50%	77.19%	77.03%	77.42%	77.03%	76.95%	76.64%	76.64%	77.14%	77.66%
SF	SVM	82.97%	84.45%	83.44%	83.83%	83.36%	83.83%	84.77%	84.84%	84.69%	84.61%	84.08%	84.84%
	KNN	**89.77%**	**89.38%**	**89.38%**	**89.53%**	**89.84%**	**89.53%**	**88.83%**	**89.30%**	**89.84%**	**89.14%**	**89.45%**	**89.84%**
	NB	85.55%	85.55%	85.47%	86.72%	86.09%	85.63%	85.78%	86.33%	86.17%	86.48%	85.98%	86.72%
PSE	SVM	72.66%	71.41%	72.58%	71.64%	70.70%	71.56%	71.56%	70.47%	69.77%	71.56%	71.39%	72.66%
	KNN	75.23%	75.55%	**74.84%**	75.70%	**76.17%**	**75.78%**	**76.09%**	**75.00%**	**74.84%**	**74.69%**	**75.39%**	76.17%
	NB	**76.33%**	**75.70%**	74.69%	**75.78%**	**76.17%**	75.16%	75.23%	74.61%	74.38%	74.45%	75.25%	**76.33%**
HOS	SVM	78.67%	79.33%	78.66%	78.13%	78.33%	78.11%	78.40%	78.40%	77.50%	84.61%	79.01%	79.33%
	KNN	**84.69%**	**85.94%**	**85.86%**	**83.91%**	**85.55%**	**86.95%**	**85.08%**	**85.55%**	**85.70%**	**87.34%**	**85.66%**	**86.95%**
	NB	81.48%	81.41%	81.72%	81.56%	82.11%	81.88%	81.41%	81.95%	82.11%	86.41%	82.20%	82.11%

Table 4: All data results for arousal using DEAP dataset.

# of Features		30	31	32	33	34	35	36	37	38	39	MEAN	MAX
HOC	SVM	71.60%	67.55%	67.34%	69.20%	68.87%	73.26%	65.97%	67.79%	72.24%	71.64%	69.55%	73.26%
	KNN	**77.03%**	**76.80%**	**75.63%**	**74.53%**	**75.78%**	**75.47%**	**77.11%**	**76.72%**	**76.56%**	**74.53%**	**76.02%**	**77.11%**
	NB	73.75%	73.91%	72.89%	73.44%	73.36%	72.81%	73.13%	73.36%	72.27%	73.05%	73.20%	73.91%
SF	SVM	80.47%	81.09%	**81.72%**	81.80%	81.48%	81.41%	**82.73%**	80.23%	81.56%	81.09%	81.36%	82.73%
	KNN	**83.28%**	**83.44%**	81.48%	**82.73%**	**83.83%**	**82.42%**	82.50%	**82.50%**	**82.42%**	**83.13%**	**82.77%**	**83.83%**
	NB	81.80%	81.09%	**81.72%**	81.25%	81.17%	81.09%	80.55%	81.17%	81.95%	80.78%	81.26%	81.95%
PSE	SVM	71.48%	68.59%	68.91%	67.19%	67.97%	68.91%	68.83%	68.59%	69.53%	68.05%	68.81%	71.48%
	KNN	**73.52%**	**72.66%**	70.63%	**72.03%**	**71.33%**	**72.27%**	**75.00%**	**72.42%**	**72.19%**	**72.81%**	**72.49%**	**75.00%**
	NB	71.95%	71.25%	**71.64%**	71.25%	70.47%	71.09%	70.78%	70.00%	71.33%	70.23%	71.00%	71.95%
HOS	SVM	75.68%	73.80%	73.36%	74.59%	76.64%	74.77%	70.31%	75.41%	74.30%	77.19%	74.61%	76.64%
	KNN	78.13%	**80.39%**	79.38%	**80.23%**	**79.22%**	**80.39%**	**80.00%**	**80.39%**	**79.14%**	**81.25%**	**79.85%**	**80.39%**
	NB	**79.45%**	78.91%	**79.45%**	78.75%	78.98%	78.75%	78.59%	79.22%	78.91%	79.45%	79.05%	79.45%

Overall, the reduced data give better results than non-reduced data for both valence and arousal. For both reduced and non-reduced data, four different feature extraction methods were used. To avoid overfitting, features were selected using ReliefF [60]. The number of features selected was chosen to be between 30 and 39 features. The experiments were run for all feature sets and compared. Furthermore, three different classifiers were used.

The results for reduced data are shown in Table 1 for valence and in Table 3 for arousal. Similarly, the all data results are given in Table 2 and Table 4 for valence and arousal, respectively. The results have been shown for all combinations of feature extraction methods, number of features and classification methods. The best classification method for each feature extraction method and number of features selected are denoted in bold. The highest accuracy per number of features are highlighted in grey. The average for each method are given in the final column with the highest highlighted.

Higher Order Crossing Feature Extraction

The HOC method was chosen because it gave good results when Ekman's picture set was used in [49]. The experiments were run using both, reduced data and all data. Tables 4 and 2 show that the average accuracy using HOC and SVM classification with all data is 72.51% and 69.55% for valence and arousal, respectively. The best results achieved for valence using SVM classification is 75.92%. The highest average accuracy of 77.3% was obtained using KNN classifier with 32 features. The overall highest accuracy for all data is 78.44%. For NB classification, the average overall classification is 76.21% and the highest accuracy is 76.8%. All these results can be seen in Table 2 for valence and Table 4 for arousal.

Comparing results in classifying valence using all data and reduced data, Table 1 shows that the results achieved using reduced data are superior. The average increase in accuracy is 6.67%. The highest accuracy of 84.92% was obtained using KNN classification and 31 features.

Furthermore, the reduced data preforms better for arousal classification. Reduced data results for arousal are shown in Table 3. The overall average increase in accuracy is 5.72%. The highest accuracy is 77.11% on all data, and 82.03% on reduced data when classifying arousal. Both were obtained using KNN with 36 and 30 features, respectively.

Statistical Features feature Extraction

The best overall results, for both reduced data and all data, were achieved using statistical features for valence. The complete set of results are given in Table 1 and Table 3 for reduced data, and in Table 2 and Table 4 for all data. Using KNN and all data, the highest accuracy when classifying valence is 84.38%. Using reduced data, the average accuracy when using KNN increased by 6.19%. The highest accuracy overall achieved in classifying valence was obtained using reduced data, statistical features and KNN, namely 89.61%.

When using SVM and NB with statistical features, the increase in accuracy is not as high as the increase in accuracy when using KNN. The best result obtained using all data and SVM is 81.48%. Using NB with all data results in accuracy of 83.28%. Table 1 shows that the highest accuracy achieved for SVM and NB on reduced data are 87.11% and 87.42%, respectively.

Similar results were obtained in classifying arousal. For both, all data and reduced data, the best results were achieved using statistical features and KNN. The highest accuracy achieved in using reduced data and all data are 89.84% and 83.83%, respectively. Thus, reducing data improves the classification accuracy. Using SVM and NB to classify arousal on all data resulted in accuracy of 82.73% and 81.95%, respectively. The same methods respectively resulted in 84.84% and 86.72% accuracy using reduced data.

Power Spectral Entropy Feature Extraction

Table 5: Reduced data results for valence using MAHNOB dataset.

# of Features		10	11	12	13	14	15	16	17	18	19	AVG	MAX
HOC	SVM	85.60%	85.60%	85.00%	84.40%	85.80%	86.60%	84.80%	84.80%	85.40%	85.40%	85.34%	86.60%
	KNN	**93.60%**	**90.00%**	**92.20%**	**91.60%**	**90.60%**	**90.80%**	**90.00%**	**90.60%**	**90.20%**	**90.40%**	**91.00%**	**93.60%**
	NB	87.00%	86.80%	86.80%	88.60%	88.40%	87.80%	87.20%	88.00%	88.00%	88.40%	87.70%	88.60%
SF	SVM	**93.40%**	92.40%	93.00%	91.20%	91.40%	90.00%	91.60%	92.40%	93.20%	91.40%	92.00%	93.40%
	KNN	93.20%	**93.20%**	**94.00%**	**93.00%**	**93.80%**	**93.60%**	**93.60%**	**92.80%**	**94.60%**	**94.40%**	**93.62%**	**94.60%**
	NB	91.60%	91.80%	92.00%	91.60%	91.80%	91.60%	91.20%	92.40%	92.80%	93.20%	92.00%	93.20%
PSE	SVM	87.40%	86.40%	86.40%	87.40%	87.00%	87.40%	87.00%	89.20%	88.00%	85.80%	87.20%	89.20%
	KNN	**91.20%**	**91.40%**	**90.20%**	**89.40%**	89.60%	**90.60%**	89.40%	90.00%	89.40%	89.80%	**90.10%**	**91.40%**
	NB	89.00%	90.00%	**90.20%**	89.20%	**90.00%**	89.60%	**90.40%**	**90.20%**	**90.40%**	**90.40%**	89.94%	90.40%
HOS	SVM	87.40%	88.60%	89.60%	88.20%	89.80%	89.00%	89.40%	90.40%	90.40%	90.60%	89.34%	90.60%
	KNN	**91.40%**	**92.40%**	**91.80%**	**90.60%**	**91.80%**	**93.00%**	**91.40%**	**93.60%**	**93.80%**	**93.40%**	**92.32%**	**93.80%**
	NB	89.60%	88.60%	88.40%	89.40%	88.80%	88.00%	86.80%	88.20%	88.00%	88.60%	88.44%	89.60%

Table 6: All data results for valence using MAHNOB dataset.

# of Feature		10	11	12	13	14	15	16	17	18	19	MEAN	MAX
HOC	SVM	62.40%	64.00%	64.00%	63.80%	63.80%	62.00%	63.60%	64.60%	63.80%	62.20%	63.42%	64.60%
	KNN	**73.80%**	**72.00%**	**78.00%**	**74.60%**	**74.20%**	**72.40%**	67.80%	**73.40%**	**72.40%**	**73.00%**	**73.16%**	**78.00%**
	NB	71.80%	71.60%	73.00%	74.00%	72.80%	**72.40%**	**70.40%**	70.60%	70.20%	70.80%	71.76%	74.00%
SF	SVM	86.00%	86.20%	85.20%	83.40%	84.60%	84.00%	86.40%	87.60%	87.20%	88.20%	85.88%	88.20%
	KNN	**88.20%**	**89.20%**	**91.00%**	**89.20%**	**89.60%**	**89.20%**	**90.40%**	**89.20%**	**88.60%**	**88.80%**	**89.34%**	**91.00%**
	NB	85.00%	84.20%	85.40%	84.60%	84.80%	85.60%	85.20%	84.80%	85.40%	85.40%	85.04%	85.60%
PSE	SVM	66.20%	68.80%	69.60%	69.60%	68.80%	66.80%	68.40%	65.40%	65.60%	66.80%	67.60%	69.60%
	KNN	**76.20%**	**77.40%**	**77.60%**	**75.60%**	**75.40%**	**75.20%**	**75.00%**	**70.80%**	**74.40%**	**77.40%**	**75.50%**	**77.60%**
	NB	71.80%	71.80%	69.40%	71.00%	69.80%	70.40%	69.40%	68.60%	68.80%	68.80%	69.98%	71.80%
HOS	SVM	73.60%	72.00%	73.00%	69.80%	70.80%	71.40%	71.60%	69.20%	70.20%	68.20%	70.98%	73.60%
	KNN	**78.40%**	**77.80%**	**76.20%**	**76.80%**	**78.00%**	**75.80%**	71.60%	75.00%	**76.20%**	**75.60%**	**76.14%**	**78.40%**
	NB	76.80%	75.20%	75.20%	75.20%	75.00%	74.80%	**73.60%**	**75.80%**	75.20%	73.20%	75.00%	76.80%

The PSE method was chosen due its good performance in extracting features from EEG for imagined left and right-hand movements in [61]. Our results show that PSE feature extraction is outperformed by all of our chosen feature extraction methods.

Classifying valence with PSE feature extraction gives the accuracy of 69.14% using SVM, 75.86% using KNN, and 70.88% using NB. Reducing the data improves the accuracy of valence on average by 2.74%. The best results were obtained when using KNN with 30 features, namely 76.17%. The accuracies in using SVM and NB on reduced data are 72.73% and 75.39%, respectively.

Using all data in classifying arousal gives the accuracy of 71.48 using SVM, 75.0% using KNN, and 71.95% using NB (see Table 4). As before, reducing the data increases the accuracy. The accuracies achieved using reduced data are 72.66%,

76.17%, and 76.33% for SVM, KNN, and NB, respectively. It will be noted that when using PSE to calculate arousal, the best results were achieved using NB.

Higher Order Spectral Feature Extraction

An accuracy of 82% for two-class classification using HOS feature extraction was achieved in [33]. Motivated by this, HOS was chosen as one of the feature extraction methods. Using the HOS feature extraction method proposed in [33] gives the highest accuracy of 79.84% for valence and 80.39% for arousal on all data. Reducing the data increases the accuracy of valence to 84.14% and to 86.95% for arousal. All these results have been observed when using KNN.

When using HOS feature extraction, the overall average increase in accuracy is 4.46% for arousal and 2.97% for valence. For both valence and arousal, the highest increase in accuracy is when using KNN.

4.2.2 MAHNOB dataset

Similarly to DEAP dataset, different classification methods were trained to compare their performance using the reduced data and all data. For all methods, the reduced data give the higher accuracy for both valence and arousal. The highest accuracy achieved in classifying valence is 94.6% and for arousal 94%.

Likewise, the classification methods were trained using four different feature extraction methods, including HOC, HOS, SF, and PSE. Using each of these methods, features were selected. Since the number of observations per subject is 20, the number of features selected was between 10 and 19. The classification methods used were SVM, KNN, and NB. All results for reduced data are shown in Table 5 for valence, and Table 7 for arousal. In addition, the all-data results are given in Table 6 for valence and Table 8 for arousal.

Table 7: Reduced data results for arousal using MAHNOB dataset.

# of Features		10	11	12	13	14	15	16	17	18	19	AVG	MAX
HOC	SVM	87.00%	86.60%	89.00%	87.80%	86.80%	88.00%	87.60%	87.20%	90.20%	88.20%	87.84%	90.20%
	KNN	**91.20%**	**91.20%**	**92.60%**	**92.80%**	**91.60%**	**90.40%**	**93.00%**	**89.20%**	**91.60%**	**92.23%**	**91.58%**	**93.00%**
	NB	89.60%	89.00%	89.40%	90.20%	88.80%	89.40%	89.00%	88.80%	89.40%	89.40%	89.30%	90.20%
SF	SVM	89.60%	91.40%	**92.00%**	91.40%	91.60%	91.00%	90.20%	91.40%	90.60%	91.00%	91.02%	92.00%
	KNN	**93.60%**	91.60%	91.40%	**93.60%**	**93.40%**	**93.20%**	**92.40%**	**92.40%**	**94.00%**	**92.40%**	**92.80%**	**94.00%**
	NB	90.20%	**92.40%**	90.60%	91.00%	91.40%	90.40%	92.00%	92.00%	93.20%	92.00%	91.52%	93.20%
PSE	SVM	87.40%	87.40%	88.00%	86.80%	87.40%	86.20%	85.40%	86.40%	86.80%	85.80%	86.76%	88.00%
	KNN	89.60%	88.20%	87.60%	86.80%	85.60%	88.60%	87.60%	86.20%	85.20%	88.40%	87.38%	89.60%
	NB	**90.60%**	**90.20%**	**91.20%**	**90.80%**	**90.40%**	**90.20%**	**89.60%**	**89.00%**	**90.60%**	**90.60%**	**90.32%**	**91.20%**
HOS	SVM	88.80%	89.40%	89.60%	89.40%	90.20%	91.40%	90.00%	90.40%	91.00%	**91.80%**	90.20%	91.80%
	KNN	**89.60%**	**90.27%**	**91.00%**	**91.40%**	**93.60%**	**92.60%**	**92.00%**	**91.00%**	**93.60%**	91.20%	**91.63%**	**93.60%**
	NB	88.40%	88.20%	88.60%	87.40%	86.80%	88.20%	88.20%	87.80%	89.20%	88.40%	88.12%	89.20%

Table 8: All data results for arousal using MAHNOB dataset.

# of Features		10	11	12	13	14	15	16	17	18	19	MEAN	MAX
HOC	SVM	66.00%	66.40%	66.80%	65.60%	66.40%	67.00%	67.00%	69.40%	65.80%	68.00%	66.84%	69.40%
	KNN	**75.40%**	**76.00%**	**78.80%**	**77.60%**	**77.40%**	**79.00%**	**77.60%**	**74.80%**	**78.00%**	**77.93%**	**77.25%**	**79.00%**
	NB	74.20%	74.60%	75.40%	74.60%	75.00%	74.40%	74.60%	74.60%	75.40%	74.00%	74.68%	75.40%
SF	SVM	87.40%	87.80%	87.40%	88.00%	88.40%	88.40%	89.60%	**89.00%**	87.40%	**89.40%**	88.28%	89.60%
	KNN	**90.20%**	**90.20%**	**89.40%**	**92.20%**	**89.40%**	**88.20%**	**89.80%**	88.80%	**88.40%**	89.20%	**89.58%**	**92.20%**
	NB	86.60%	84.80%	85.80%	83.60%	86.20%	86.00%	85.20%	85.40%	86.60%	86.00%	85.62%	86.60%
PSE	SVM	70.20%	70.20%	69.80%	68.80%	67.80%	67.40%	68.40%	67.80%	67.80%	68.80%	68.70%	70.20%
	KNN	**79.80%**	**76.60%**	**77.60%**	**78.00%**	**77.60%**	**77.60%**	**76.40%**	**75.80%**	**75.40%**	**76.20%**	**77.10%**	**79.80%**
	NB	75.80%	74.40%	74.80%	76.00%	73.80%	74.20%	74.20%	73.60%	74.00%	70.80%	74.16%	76.00%
HOS	SVM	81.20%	81.00%	84.40%	85.80%	82.80%	86.00%	86.60%	87.80%	85.40%	86.40%	84.74%	87.80%
	KNN	**89.00%**	**86.20%**	**85.40%**	**88.80%**	**86.40%**	**86.80%**	**88.40%**	**88.60%**	**89.60%**	**90.00%**	**87.92%**	**90.00%**
	NB	82.00%	81.40%	83.00%	81.60%	82.60%	83.20%	83.40%	83.00%	83.00%	81.80%	82.50%	83.40%

Higher Order Crossing Feature Extraction

When using all data to classify valence, the best results were obtained using KNN with 12 features, resulting in accuracy of 78.0%. The performance of NB and SVM on all data are 74.0% and 64.6%, respectively. These results were obtained using 13 features for NB and 17 features for SVM, Table 6.

Table 5 shows that using reduced data improves accuracy on average by 18.56%. This takes into account different number of features and classification methods. The highest increase of 21.92% was achieved using SVM. The lowest increase in accuracy was achieved using NB, namely 15.94%. The highest accuracy achieved using reduced data is 86.6%, 93.6%, and 88.6% for SVM, KNN, and NB, respectively.

In addition to classifying valence, HOC was used to classify positive and negative arousal. Like in the valence case, the best results were obtained for both on all data and reduced data using KNN. Table 8 shows the highest accuracy achieved on all data is 79.0% and Table 5 shows the highest accuracy on reduced data is 93.0%. The average increase in accuracy due to data reduction is 16.65%.

Comparing with the results obtained for DEAP, the accuracy achieved is much higher on MAHNOB-HCI.

Statistical Features feature Extraction

Similarly to results on DEAP, statistical features give the best overall results for both all data and reduced data. The results on reduced data are shown in Table 5 for valence and Table 7 for arousal. Similarly, all-data results are given in Tables 6 and 8 for valence and arousal, respectively.

When using all data, the best results for valence were obtained using KNN with 12 features giving 91.0%. Reducing the data increases the accuracy to 94.6% with KNN. The highest accuracy achieved using SVM on all data is 88.2% and using NB is 85.6%. Using reduced data, the average accuracies are 93.4% for SVM and 93.2% for NB. The overall average increase is 5.8%. For KNN the average increase in accuracy is 4.3%, for SVM is 6.1%, and for NB is 7.0%.

The highest accuracy of 94.0% for arousal was reached using KNN with 18 features using reduced data. SVM and NB achieved 92.0% and 93.2%, respectively. The average increase in accuracy is lower than for valence classification. For arousal, the accuracy increases on average by 3.95%. The lowest average increase per classification method was achieved using SVM with 2.74%, and highest average when using NB with 5.9%.

Power Spectral Entropy Feature Extraction

Similarly to statistical features and HOC, the highest accuracy on all data with PSE was achieved using KNN. The highest accuracy achieved for valence is 77.6% using KNN with 12 features. Using SVM and NB respectively gives accuracies of 67.6% and 70.0%, Table 6.

Reducing the data increases the accuracy with the average increase for valence being 18.1%. For KNN, the average increase is 14.6%. The best result using KNN is 91.4%. The best results on reduced data using SVM is 89.2%, and using NB is 90.4%, Table 5.

Likewise, when classifying arousal, reducing the data increases the accuracy. Using all data, the highest accuracy reached is 79.8%, and after data reduction the classification accuracy increases to 91.2%. Note that in using all data the best result was achieved using KNN, but for reduced data the best result was obtained using NB. The average increase in accuracy for all classifiers is 14.83%.

When classifying the DEAP dataset, PSE was outperformed by all of our chosen feature extraction methods. It is noted that even though not giving the best results on the MAHNOB-HCI dataset, PSE feature extraction results are comparable to the other feature extraction methods.

Higher order Spectral Feature Extraction

Tables 5 and 7 show the accuracies achieved on reduced data, and Tables 6 and 8 show accuracies **for** on all data.

The best results were obtained using SVM. HOS is the only feature extraction method that works better with SVM and NB. Using all data, the highest accuracy in classifying valence is 78.4% using SVM, 76.8% using NB and 73.6% using KNN.

On reduced data, the best result was obtained using KNN with an accuracy of 93.8%. The corresponding best results for SVM and NB are 90.6% and 89.6%, respectively. The average increase in valence accuracy is 16.0% The highest increase in accuracy when using KNN is 21.3%. The increases in using SVM and NB are 13.2% and 13.4%, respectively.

The highest accuracy in classifying arousal using all data is 90.0%. This is much higher than using the same method in classifying valence. Using reduced data, the classification accuracy increases to 93.6%. Similar to valence classification, the best results were achieved using KNN for both all data and reduced data. The increase in accuracy, even though highly noticeable, is much smaller than in classifying valence. The average increase in accuracy in classifying valence is 16.0% and in classifying arousal is 4.93%.

4.3 Analysis and Comparison with other Emotion Recognition Systems

Accuracy may not always be the best way to evaluate classification performance, especially when the dataset is imbalanced, as accuracy treats all examples the same. Thus, Receiver Operating Characteristic (ROC) curve has also been used to visualise the performance of the classifiers.

ROC curve shows classifier performance as a trade off between specificity and sensitivity, giving a good estimate on how well the classifiers separate the classes. A good classifier generates a ROC curve close to the upper left corner, whereas a poor classifier has a curve close to a diagonal line, corresponding to a random guess.

ROC curves were generated for all feature extraction methods. Using the DEAP dataset the ROC curves in classifying valence and arousal are shown in Figure 2 and Figure 3, respectively. Similarly, Figure 4 and Figure 5 show the curves using the MAHNOB-HCI dataset. These figures show that the ROC curves of all classifiers are very close to the upper left corner, indicating a good separation between the classes. Furthermore, Figures 3 and 5 show that for arousal calculation, HOS features separate the classes slightly better than statistical features for both datasets. Also, Figure 2 shows that using HOC features separates classes better than using statistical features.

The overall trend as observed from our experimental results indicates that reducing data increases the accuracy in emotion recognition. The mRMR-SVM method for emotion classification is proposed in [15] and evaluated using the DEAP dataset. Table 9 shows the comparison of this method with the proposed. The accuracy for two-class classification using mRMR-SVM is 73.06% for valence and 73.14% for arousal. Whereas, the proposed method achieves accuracies of 89.61% and 89.84% for valence and arousal, respectively. Therefore, it can be concluded that using mutual-information windowing to reduce the data has a positive effect on accuracy. The reduced data was also used for three-class classification as well as

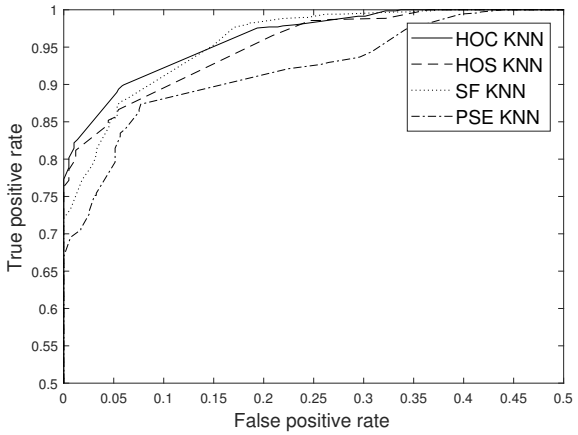

Figure 2: ROC curve for DEAP Valence Classification.

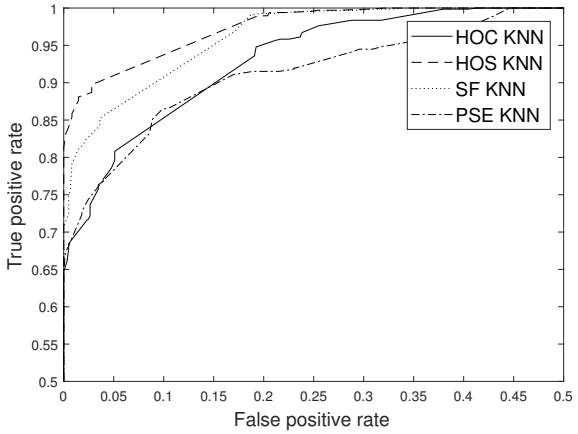

Figure 3: ROC curve for DEAP Arousal Classification.

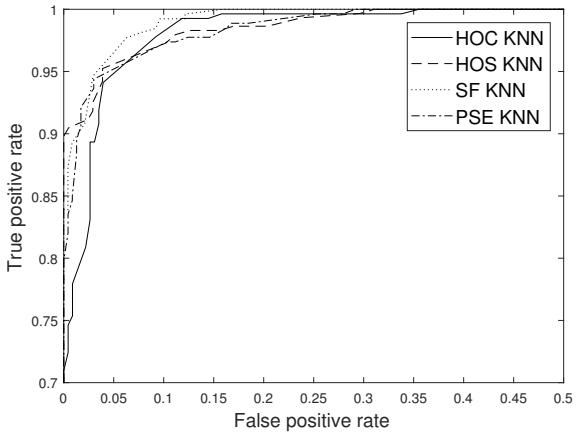

Figure 4: ROC curve for MAHNOB Valence Classification.

five-class classification. Table 9 shows that the proposed method achieves better performance for larger number classes as well. The average accuracy of the proposed method is 61.88% for valence.

We compared with two other existing methods. The accuracy for four-class classification achieved by the method in [62] reached 81.3% using IAPS (2D emotional space) dataset. For our proposed framework, both three-class classification and five-class classification reached 92.50% accuracy for a single subject, i.e., it is far more superior. Furthermore, for two-class classification the method in [63] achieved 94.4% for a single subject. For single subject classification, our framework reached 99.9% for both DEAP and MAHNOB-HCI dataset for a number of subjects. Note that to give a more representative performance of our framework, the results in Section 4.2 are presented as average accuracies over all subjects.

Finally, three-class classification was performed using MAHNOB-HCI dataset. Table 10 shows that using this dataset

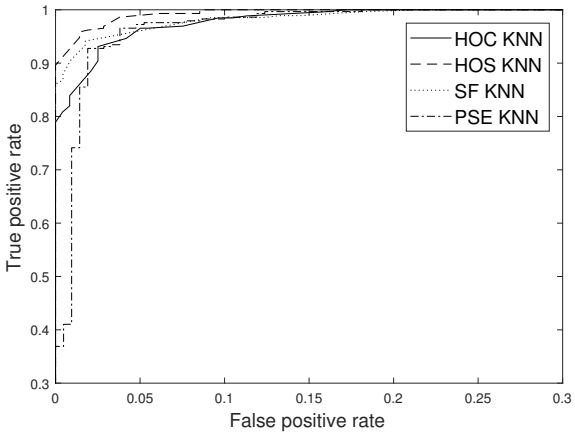

Figure 5: ROC curve for MAHNOB Arousal Classification.

Table 9: Using DEAP dataset to compare proposed method with method in [15].

Method	No. of classes	Valence	Arousal
Reduced Data, Statistical Features, KNN,	2	**89.61%**	**89.84%**
	3	**75.02%**	**75.70%**
	5	**61.88%**	**67.19%**
mRMR-SMV [15]	2	73.14%	73.06%
	3	62.33%	60.70%
	5	45.32%	46.69%

gives more superior results than using DEAP dataset. There are two explanations for this. First, DEAP used music videos, whereas MAHNOB-HCI used mostly clips from motion pictures. It can be argued, that the stimuli used in MAHNOB-HCI are more effective in evoking emotional response. Second, both datasets are noisy and the pre-processing may not have been as effective on DEAP dataset.

In addition the method was evaluated on a dataset generated by a commercial wireless EEG recording device, i.e., DREAMER [64]. For two class classification, the average accuracy of valence and arousal are respectively 91.7% and 90.4%, i.e., adequate accuracy for emotion recognition. Since the dataset is smaller that DEAP and MAHNOB, no in-depth results are presented.

5 Conclusion

This paper introduces a mutual information based data reduction via windowing for the purpose of increasing the accuracy of subject-dependent emotion recognition. Several feature selection methods as well as classification methods were used with the data reduction method. Overall, selecting an appropriate reduced signal improves accuracy.

PSE features did not perform as well as other features. When using accuracy as a measure to validate the methods, statistical features gave better results. However, when the experimental results were analysed more thoroughly using sensitivity and specificity of the models, it can be concluded that HOS and HOC feature extraction methods gave better results. Furthermore, using ROC curves to analyse different classifiers, SVM and NB classifiers should not be discarded, as on average when using statistical features, NB and SVM separate the classes better than KNN.

There are multiple factors to consider when dealing with emotion recognition using EEG signals. First, the EEG signal is noisy. Furthermore, with most noise removal techniques there is no good way to identify which part of the signal is related to emotions. Another factor is the required training time. For a single participant, full training using the proposed data reduction took around 12 hours, depending on the size of the maximum data window. This was overcome by parallelising the code, so that all training was performed simultaneously. From the experimental results, reducing the data gives more accurate emotion recognition, and therefore is an option to consider. Depending on the feature extraction method used the proposed data reduction method can still be a faster way of training the EEG emotion classifier.

Another challenge is the size of the datasets. The largest publicly available dataset, DEAP, has been used together with a smaller similar dataset. In addition, due to subject dependency of the signals, the classification algorithms are trained for each subject separately. The classification task is more challenging when higher number of emotions are involved.

Table 10: Classification using MAHNOB-HCI dataset

Method	No. of classes	Valence	Arousal
Reduced Data, Statistical Features, KNN	2	94.6	94.00%
	3	86.00%	87.20%

Appropriate steps have been taken to address small datasets, namely leave-one-out cross validation and ROC curves which give the trade-off between sensitivity and specificity. Reducing the data showed promising results in 3-class classification as well as 5-class classification.

6 Acknowledgement

The authors would like to thank Warwick School of Engineering for providing the funds for this research.

References

[1] J. H. Gruzelier,"EEG-neurofeedback for optimising performance. I: a review of cognitive and affective outcome in healthy participants," Neuroscience & Biobehavioral Reviews, vol. 44, pp. 124–141, (2014).
[2] F. L. da Silva, "EEG and MEG: relevance to neuroscience," Neuron, vol. 80, no. 5, pp. 1112–1128, 2013.
[3] J. Frey, C. Mühl, F. Lotte, M.Hachet, "Review of the use of electroencephalography as an evaluation method for human-computer interaction," In Proc. Int. Conf. on Physiological Computing Systems, PhyCS, Lisbonne, Portugal, 2014.
[4] G. Recio, A. Schacht, W. Sommer. "Recognizing dynamic facial expressions of emotion: Specificity and intensity effects in event-related brain potentials," Biological psychology, vol. 96, pp. 111–125, 2014.
[5] H. Ali, M. Hariharan, S. Yaacob, A.H. Adom, "Facial emotion recognition using empirical mode decomposition," Expert Systems with Applications, vol.42, no. 3, pp. 1261–1277, 2015.
[6] C. M. Lee, S. Yildirim, M. Bulut, et al., "Emotion recognition based on phoneme classes," Interspeech, 2004, pp. 205–211.
[7] K. Han, D. Yu, I. Tashev. "Speech emotion recognition using deep neural network and extreme learning machine," Interspeech, 2014, pp. 223–227.
[8] F. Dellaert, T. Polzin and A. Waibel, "Recognizing emotion in speech," in Proc. ICSLP, Philadelphia, PA, 1996, pp. 1970–1973.
[9] L. Kessous, G. Castellano and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," in J. Multimodal User Interfaces, vol. 3, no. 1 pp. 33–48, 2010.
[10] T. B"anziger, D. Grandjean and K. R. Scherer, "Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT)," Emotion, vol. 9, no. 5, pp. 691–704, 2009.
[11] M. K. Kim, et al. "A review on the computational methods for emotional state estimation from the human EEG," Computational and Mathematical Methods in Medicine, vol. 2013, Article ID 573734, 13 pages, 2013.
[12] D. Sammler, M. Grigutsch, T. Fritz, S. Koelsch, "Music and emotion: electrophysiological correlates of the processing of pleasant and unpleasant music," in Psychophysiology, vol. 44, no. .2, pp. 293-304, 2007.
[13] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," IEEE Trans. Affective Comput., vol. 3, no. 1, pp. 18–31, 2012.
[14] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, "A multimodal database for affect recognition and implicit tagging," IEEE Trans. Affective Comput., vol. 3, no. 1, pp. 42–55, 2012.
[15] J. Atkinson, D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," Expert Systems with Applications, vol. 47, pp. 35–41, 2016.
[16] J. W. Matiko, Y. Wei, R. Torah, N. Grabham, G. Paul, S. Beeby, J. Tudor. "Wearable EEG headband using printed electrodes and powered by energy harvesting for emotion monitoring in ambient assisted living." Smart Materials and Structures, vol. 24, no.12, (2015), Article ID 125028.
[17] H. Aurlien, I.O. Gjerde, J. H. Aarseth, G. Eldøen, B. Karlsen, H. Skeidsvoll, N. E. Gilhus, "EEG background activity described by a large computerized database," Clinical Neurophysiology, vol. 115, no. 3, pp. 665–673, 2004.
[18] Y. Liu and O. Sourina, "Real-Time Subject-Dependent EEG-Based Emotion Recognition Algorithm," in Transactions on Computational Science XXIII, ed: Springer, 2014, pp. 199–223.
[19] S. Sanei, J. A. Chambers, "Fundamentals of EEG Signal Processing," in EEG signal processing, John Wiley & Sons, 2013, pp. 35–125.
[20] D. J. McFarland, L. M. McCane, S.V. David, J. R. Wolpaw, "Spatial filter selection for EEG-based communication," Electroencephalography and clinical Neurophysiology, vol. 103, no. 3, 1997, pp. 386-394.
[21] P.L. Nunez, R.B. Silberstein, P.J. Cadusch, R.S. Wijesinghe, A.F. Westdorp, R. Srinivasan, "A theoretical and experimental study of high resolution EEG based on surface Laplacians and cortical imaging," Electroencephalography and Clinical Neurophysiology, vol. 90, no. 1 pp. 40-57, 1994.
[22] Y. Liu, O. Sourina, and M. K. Nguyen, "Real-time EEG-based human emotion recognition and visualization," in Proc. 2010 Int. Conf. on Cyberworlds, Singapore, 2010, pp. 262–269.
[23] G. Repovš, "Dealing with noise in EEG recording and data analysis," Informatica Medica Slovenica, vol. 15, no. 1, pp. 18–25 2010.
[24] J. A. Urigüen, B. Garcia-Zapirain, "EEG artifact removal- state-of-the-art and guidelines," J. Neural Engineering, vol. 12, no. 3, 2015, Article ID 031001.
[25] R. J. Croft, R. J. Barry, "Removal of ocular artifact from the EEG: a review," Neurophysiol. Clin., vol. 49, pp. 5–19, 2000.
[26] G. L. Wallstrom, R. E. Kass, A. Miller, J. F. Cohn, N. A. Fox. "Automatic correction of ocular artifacts in the EEG: a comparison of regression-based and component-based method," Int. J. of Psychophysiol., vol. 53, pp. 105–119, 2004.
[27] S. Puthusserypady, T. Ratnaraja, "H^{∞}adaptive filters for eye blink artifact minimization from electroencephalogram," IEEE Sig. Proc. Lett., vol. 12, no. 12, pp. 816–819, 2005.
[28] G. Gómez-Herrero, W. De Clercq, H. Anwar, O. Kara, K. Egiazarian, S. Van Huffel, W. Van Paesschen, "Automatic removal of ocular artifacts in the EEG without an EOG reference channel," in Proc. 7th Nordic Signal Process. Symp., Jun. 2006, pp. 130–133.
[29] T. P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski, "Removing electroencephalographic arti- facts by blind source separation," Psychophysiol., vol. 37, pp. 168–178, 2000.
[30] O. G. Lins, T. W. Picton, P. Berg, and M. Scherg, "Ocular artifacts in recording EEGs and event-related potentials, II: Source dipoles and source components," Brain Topography, vol. 6, pp. 65–78, 1993.
[31] A. S. Al-Fahoum and A. A. Al-Fraihat, "Methods of EEG Signal Features Extraction Using Linear Analysis in Frequency and Time- Frequency Domains," ISRN Neuroscience, vol. 2014, Article ID 730218, 7 pages, 2014.
[32] R. Jenke, A. Peer, M. Buss, "Feature extraction and selection for emotion recognition from EEG," IEEE Trans. Affective Comput., vol. 5, no. 3, pp. 327–339, 2014.
[33] S. A. Hosseini, M. A. Khalilzadeh, M. B. Naghibi-Sistani, V. Niazmand, "Higher order spectra analysis of EEG signals in emotional stress states," in Proc. IEEE Int. Conf. Inform. Technol. Comput. Sci., Jul. 2010, pp. 6063.
[34] J. Yang, V. Honavar, "Feature subset selection using a genetic algorithm," IEEE Intell. Sys., vol. 13, no. 2, pp. 44–49, 1998.
[35] D. Garrett, D. A. Peterson, C. W. Anderson, M. H. Thaut, "Comparison of linear, nonlinear, and feature selection methods for EEG signal classification," IEEE Trans. Neural Syst. Rehab. Eng, vol. 11, pp. 141–144, June 2003
[36] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," J. Neural Eng., vol. 4, pp. R1-R13, 2007.

[37] S. Zhong, J. Ghosh, "HMMs and coupled HMMs for multi-channel EEG classification," In Proc. Int. Joint Conf. Neural Networks, vol. 2, pp. 1154-1159, 2002.
[38] G. Pfurtscheller, J. Kalcher, C. Neuper, D. Flotzinger, M. Pregenzer, "On-line EEG classification during externally-paced hand movements using a neural network-based classifier," Electroenceph. Clin. Neuro- physiol., vol. 99, pp. 416–425, 1996.
[39] V. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.
[40] B. E. Boser, I. M. Guyon, V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proc. 5th Annu. Wkshp. Comput. Learning Theory. Pittsburgh, PA: ACM, 1992, pp. 144–152.
[41] K. B. Duan, S. S. Keerthi, "Which is the best multiclass SVM method? An empirical study," in Proc. 6th Int. Workshop Multiple Classifier Syst., 2005, pp. 278–285.
[42] I. Guler, E. D. Ubeyli, "Multiclass support vector machines for EEG-signals classification," IEEE Trans. Inf. Technol. Biomed., vol. 11, no. 2, pp. 117–126, Mar. 2007.
[43] A. Khorshidtalab, M. J. E. Salami, "EEG signal classification for real-time brain-computer interface applications: A review," in Proc. 4th Int. Conf. Mechatronics, Kuala Lumpur, Malaysia, 2011, pp. 1–7.
[44] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," J. Machine Learning Research, vol. 3, pp. 1415-1438, 2003.
[45] I. C. Daubechies, Ten Lectures on Wavelets. Philadelphia, PA: SIAM, 1992.
[46] M. Murugappan, R. Nagarajan, and S. Yaacob, "Classification of human emotion from EEG using discrete wavelet transform," J. Biomed. Sci. Eng., vol. 3, no. 4, pp. 390-396, 2010.
[47] R. N. Bracewell, "The Fourier transform and its applications," New York: McGraw-Hill, 1986.
[48] P. Stoica, and R. L. Moses, "Spectral analysis of signals," Upper Saddle River, NJ: Prentice-Hall, 2005.
[49] P. C. Petrantonakis, and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings." IEEE Trans. Inf. Technol. Biomed., vol. 14, no. 2, pp. 186–197, 2010.
[50] G. G'omez-Herrero, "Automatic artifact removal (AAR) toolbox v1. 3 (Release 09.12. 2007) for MATLAB," Tampere University of Technology, 2007.
[51] N.Kwak, Nojun, and C. H. Choi, "Input feature selection by mutual information based on Parzen window," IEEE Trans. on Patt. Anal. and Mach. Intell., vol. 24, no. 12, pp. 1667–1671, 2002.
[52] H. Jasper, "The ten twenty electrode system of the international federation," Electroencephalogr. Clin. Neurophysiol., vol. 10, pp. 371-375, 1958.
[53] J. A. Russell, "A circumplex model of affect," J. Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178, (1980).
[54] J. D. Morris. "Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response." J. of Advertising Research, vol. 35, no. 6, pp. 63–68, 1995.
[55] I. Kononenko, S. Robnik, and U. Pompe, "ReliefF for estimation and discretization of attributes in classification, regression and ILP problems," in Proc. Artif. Intell. Methodol. Syst. Appl. (AIMSA) 1996, IOS Press, pp. 31-40.
[56] K. Kira and L.A. Rendell. "The feature selection problem: Traditional methods and a new algorithm," In Proc. of the Tenth National Conference on Artificial Intelligence, pages 129-134, 1992.
[57] R. Duda and P. Hart, Pattern Classification and Scene Analysis. New York: Wiley, 1973.
[58] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," Machine Learning, vol. 29, pp. 131–163, 1997.
[59] MATLAB and Statistics and Machine Learning Toolbox Release 2017a, The MathWorks, Inc., Natick, Massachusetts, United States.
[60] G. Roffo, S. Melzi, and M. Cristani, "Infinite feature selection," in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 4202-4210.
[61] A. Zhang, B. Yang, and L. Huang, "Feature extraction of EEG signals using power spectral entropy," in Proc. Int. Conf. BioMed. Eng. Inform., Sanya, China, May 2008, pp. 435-439.
[62] C. Frantzidis, C. Bratsas, C. Papadelis, E. Konstantinidis, C. Pap- pas, and P. Bamidis, "Toward emotion aware computing: An inte- grated approach using multichannel neurophysiological recordings and affective visual stimuli," IEEE Trans. Inf. Technol. Biomed., vol. 14, no. 3, pp. 589-597, May 2010.
[63] P. C. Petrantonakis and L. J. Hadjileontiadis, "A novel emotion elic- itation index using frontal brain asymmetry for enhanced EEG-based emotion recognition," IEEE Trans. Inf. Technol. Biomed. vol. 15, no. 5, pp. 737-746, Sep. 2011.
[64] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices", IEEE Journal of Biomedical and Health Informatics, vol. PP, 2017.

Laura Pihol reiceived her MMath Mathematics degree from University of Leicester in 2015. She is now working towards her PhD degree in the School of Engineering, University of Warwick, UK. Her research interests include affective computing and machine learning.

Tardi Tjahjadi (SM'02) received B.Sc. in Mechanical Engineering from University College London in 1980, and M.Sc. in Management Sciences in 1981 and Ph.D. in Total Technology in 1984 from UMIST, U.K. He has been an associate professor at the University of Warwick since 2000 and a reader since 2014. His research interests include image processing and computer vision.