CrossMark

ORIGINAL PAPER

SVM-based feature selection methods for emotion recognition from multimodal data

Cristian Torres-Valencia[1] · **Mauricio Álvarez-López[1]** · **Álvaro Orozco-Gutiérrez[1]**

Abstract Multimodal emotion recognition is an emerging field within affective computing that, by simultaneously using different physiological signals, looks for evaluating an emotional state. Physiological signals such as electroencephalogram (EEG), temperature and electrocardiogram (ECG), to name a few, have been used to assess emotions like happiness, sadness or anger, or to assess levels of arousal or valence. Research efforts in this field so far have mainly focused on building pattern recognition systems with an emphasis on feature extraction and classifier design. A different set of features is extracted over each type of physiological signal, and then all these sets of features are combined, and used to feed a particular classifier. An important stage of a pattern recognition system that has received less attention within this literature is the feature selection stage. Feature selection is particularly useful for uncovering the discriminant abilities of particular physiological signals. The main objective of this paper is to study the discriminant power of different features associated to several physiological signals used for multimodal emotion recognition. To this end, we apply recursive feature elimination and margin-maximizing feature elimination over two well known multimodal databases, namely, DEAP and MAHNOB-HCI. Results show that EEG-related features show the highest discrimination ability. For the arousal index, EEG features are accompanied by Galvanic skin response features in achieving the highest discrimination power, whereas for the valence index, EEG features are accompanied by the heart rate features in achieving the highest discrimination power.

✉ Cristian Torres-Valencia
cristian.torres@utp.edu.co

[1] Department of Electrical Engineering, Faculty of Engineering, Universidad Tecnológica de Pereira, Pereira, Colombia

1 Introduction

Emotion recognition plays an important role in the different affective sciences that are related to characterizing and determining the responses of an individual to some stimulus from the environment. In recent years the improvement in the knowledge of the changes of different biological systems in the human body, as a response to an emotional stimulus, has potentiated the development of several applications for automatic emotion recognition in fields like psychological treatment, posttraumatic rehab, human-machine interfaces, and some emerging fields like marketing analysis based on emotional responses [24,29,35].

Multimodal emotion recognition (MER) uses the information of different signals such as electroencephalogram, electromyogram, electrooculogram, galvanic skin response, heart rate, skin temperature, video and audio signal to assess the affective state of an individual. Emotional states can either be quantified using a discrete space, or a continuous space. For a discrete space, the usual categories include emotions like happiness, sadness, anger, surprise, fear [7,35]. For a continuous space, emotions are assessed in terms of an arousal index and a valence index [12,20,35].

Recent efforts in multimodal emotion recognition have been based on developing pattern recognition systems with emphasis in feature extraction and classifier design. A common setup for such systems is to extract statistical features from a range of physiological signals [21,24,25,28], and then to combine those features in a large feature vector used to feed a particular classifier. Examples of classifiers include Radial basis function-support vector machine (SVM-

RBF) [24,28], KNN classifier [22] and neural networks [21].

Framing the multimodal emotion recognition within the realm of the statistical pattern recognition methodology, it is noticeable how the problem of feature selection has received less attention. Feature selection is a common step in different supervised learning problems that aims at reducing the dimensionality of the input space for easing the process of developing the predictive model. It also helps to avoid the well known problem of the "curse of the dimensionality", where no amount of data seems enough to properly train a predictive model in spaces of high dimensionality. The stage of feature selection also looks for those features that are relevant for discrimination purposes. Feature selection is meaningful in the context of multimodal emotion recognition for uncovering the discrimination ability that different physiological signals may exhibit.

Different methods for feature selection have been proposed in the literature. They can be categorized as wrapper methods, filter methods, projection methods and embedded methods [8]. For the wrapper methods, a predictive model is used to score subsets of features. Each subset selected is used to train a predictive model, and compute its performance [2,8]. Filter methods asses the relevance of the features directly from the data, ignoring the effects of each subset into the learning algorithm performance. The selection of features from filter methods could be developed for example by computing the correlation of the features with the output [2]. Projection methods perform the feature selection by projecting the original features to a lower space where each feature is a combination of two or more of them. In embedded methods, the feature selection and the learning algorithm interacts in order to determine the optimal subset that leads to obtain the largest generalization for an specific classifier. Recursive feature elimination (RFE) and margin-maximizing feature elimination (MFE) are embedded methods based on discriminant methods such as support vector machines (SVM) [1,11]. For example in support vector machine-recursive feature elimination (SVM-RFE) proposed by Guyon in [11], the features are selected based on a ranking criterion related to the weights of the trained SVM [11]. Embedded methods reduce the computational cost of the wrapper methods, and bring higher accuracy rates for the subset of features selected, than the filter methods [8].

Most of the attempts to include a feature selection stage for MER have focussed on how the selected set of features reflect over the accuracy of the classifier. For example, in [22], Gu et al. proposed the use of Genetic algorithms (GA's) for selecting features and classifying emotions from some physiological signals such as ECG, GSR, respiratory pattern and EMG. Statistical moments were used to extract features from the signals in order to differentiate between high and low levels of arousal and valence. A similar work

is also developed in [10], where the emotion recognition is carried out from a set of acquired physiological signals [galvanic skin response (GSR), ECG and skin temperature (SKG)] for a discrete classification space for seven emotions. A work presented in [32] develop the emotion assessment from physiological signals such as electromyograph (EMG), ECG, GSR and respiratory signal with a stage of feature selection from a bunch of statistical features using the Tabu Search (TS) algorithm. In [3], Cheng et. al. develop the emotion recognition task with the assessment of the relevance of 19 physiological parameters, from EMG, ECG, GSR and temperature signals using a commercially statistical package (SPSS) for applying the paired t-test [3]. In [21], the authors present a study of the relevance of the features extracted from audio and video signals, using neural networks in a discrete space of emotions. The results shows that different combinations of the features extracted leads to different accuracy from the classifier. In [31], a similar feature space reduction is performed based on PCA over a set of nonlinear features from the RQA analysis. To the best of our knowledge, the study and analysis of which physiological signals seem more relevant for the classification problem have not received the desired attention. Works in the state-of-the-art have focused on performing a feature space reduction searching for the higher possible accuracy rate without a deep insight of the features that are more relevant and from which signals they were extracted. An exhaustive analysis of the relevance of the features and the signals involved in MER would contribute to the development of systems that extract features from the more significant signals.

The aim of this work is to perform feature selection using discriminant-based algorithms for further automatic emotion recognition within a multimodal approach using different biosignals such as EEG and peripheral signals. The feature extraction is performed from linear and non-linear analysis. For the emotion classification, a discrete space based in the arousal and valence indexes is chosen. The main contribution of the methodology is the feature selection stage where two embedded methods such as RFE and MFE based in discriminant methods such as SVM's are applied in order to determine the most relevant features for emotion recognition. Several biclass experiments using two databases are performed to assess the efficiency of the feature selection methods and the emotion recognition methodology.

2 Materials and methods

In this section, the different databases and implementation details are depicted and also the theoretical background of the methods for feature extraction and selection is included. A brief description of feature extraction following some state of art methodologies for linear [14] and non-linear analysis

Table 1 Features from EEG, peripheral and video signals [14]

Signal	Extracted features
GSR	Average skin resistance, average of derivative, average of derivative for negative values only, proportion of negative samples in the derivative vs. all samples, number of local minima in the GSR signal, average rising time of the GSR signal, 10 spectral power in the [0–2.4] Hz bands, zero crossing rate of skin conductance slow response (SCSR) [0–0.2] Hz, zero crossing rate of skin conductance very slow response (SCVSR) [0–0.08] Hz, SCSR and SCVSR mean of peaks magnitude
Skin temperature	Average, average of its derivative, spectral power in the bands ([0–0.1] Hz, [0.1–0.2] Hz)
Respiration pattern	Average respiration signal, mean of derivative (variation of the respiration signal), standard deviation, 10 spectral power in the bands from 0 to 2.4 Hz
Blood volume pressure	Average and standard deviation of HR, HRV, and inter beat intervals, energy ratio between the frequency bands [0.04–0.15] Hz and [0.15–0.5] Hz, spectral power in the bands ([0.1–0.2] Hz, [0.2–0.3] Hz, [0.3–0.4] Hz), low frequency [0.01–0.08] Hz, medium frequency [0.08–0.15] Hz and high frequency [0.15–0.5] Hz components of HRV power spectrum
EEG	Theta, slow alpha, alpha, beta, and gamma spectral power for each electrode. The spectral power asymmetry between 14 pairs of electrodes in the four bands of alpha, beta, theta and gamma
EMG and EOG	Eye blinking rate, energy of the signal, mean and variance of the signal
Video	Mean shape (shape of the face in all the frames from each video)

of signals documented in [31] based on recurrence plots (RP) is presented.

2.1 Databases

Different databases have been published in recent years for performing multimodal emotion analysis. In this work two databases are used, the database for emotion analysis using physiological signals (DEAP) [14], and the multimodal database for affect recognition and implicit tagging (MAHNOB-HCI) [26]. The two databases were obtained via emotion elicitation with a set of different videos presented to every subject for self-assessment of different indexes such as arousal, valence, dominance and liking. These indexes could form a dimensional space of emotion representation, where the arousal dimension represents the intensity of the emotion and goes from aphatic to excited. Valence is the dimension that is related to the positiveness or negativeness of the emotional state. Dominance is related to the level of social status relative to other individuals and range from a weak feeling to an empowered feeling [14]. For the DEAP database the EEG, EMG, EOG, GSR, Respiration pattern, plethysmograph, skin temperature and video signal were recorded for 32 participants watching 40 videos [14]. For the MAHNOB-HCI database the EEG, ECG, respiration amplitude, skin temperature, audio and video signal were recorded while 30 participants were elicited by the stimulus video [26].

2.2 Feature extraction

From the signals included in each database, a linear analysis for feature extraction is developed following certain statistical measures and frequency analysis. This analysis is

commonly used to extract information about the physiological signals [14] as a result from various studies in which it has been concluded that some of these features have a direct relationship with certain emotional states [27]. The features extracted from the EEG and peripheral signals are presented in Table 1. The spectral power in some band frequencies and statistical moments from some signals, such as the average and the standard deviation, are computed [14].

Based on the work presented by Valenza [31], a set of non-linear features are also included in the scheme of multimodal emotion classification. From the results obtained with the use of the non-linear features, an improvement in the performance is reported in terms of percentages of accuracy. These nonlinear features are based on a methodology called Recurrence Plots and are extracted from the GSR, respiratory pattern and the heart rate signals. In recent years, this nonlinear analysis has been applied in several works related with the affective computing, including bipolar patients for emotional response analysis in [9,30]. In these works, the nonlinear analysis shows higher performance in the recognition of affective states than the classical time-frequency analysis of the signals [31].

Recurrence plots are based on a technique for the analysis of complex dynamic systems called embedding procedure [31], where a set of vectors X_i is constructed from the time series representing the behavior of the system. The evolution of the system can be represented by the projection of the vectors on a path between a multidimensional space that is commonly known as a phase space or phase state [31]. Eckmann in [6] introduced a tool that can be used to display states recurrence X_i in phase space. This tool called recurrence plots (RP) allows to investigate the m-dimensional phase space trajectory from a two-dimensional representa-

tion of their recurrences [19]. Following RP calculation, a recurrence quantification analysis (RQA) proposed in [34], is computed to quantify the number and length of recurrences from a dynamic system, represented by its state space trajectory. Then, from the RQA some features are extracted [31]. For further details of the RP technique and the RQA analysis (see [6,31]).

2.3 Feature selection and classification

The embedded methods for feature selection and classification stage using SVM's RFE and MFE are discussed in this section. A brief introduction of the basis of SVM's is presented first, followed by the explanation of how the RFE and MFE methods involve the SVM training into the feature selection task.

2.3.1 Support vector machines

Support vector machines are the state-of-art machine-learning algorithm. The SVM methodology propose that the inputs from the observations $\mathbf{x}$, could be mapped into a higher dimensional space, where a class separation hyperplane could be computed [4]. The computed function then is used to assign a label on the output y [4]. To find an optimum hyperplane that effectively separates the different classes of the data inputs, a small amount of the observations that lies on the edge of separation called support vectors (SV) is used [4].

Let

$$\mathbf{w_0} \cdot \mathbf{z} + b_0 = 0, \tag{1}$$

be an optimal hyperplane in feature space. The weights $\mathbf{w}_0$ for the optimal hyperplane can be written as a linear combination of support vectors [4]

$$\mathbf{w}_0 = \sum_{SV} \alpha_i \mathbf{z}_i.$$

The optimal hyperplane

$$\mathbf{w}_0 \cdot \mathbf{x} + b_0 = 0,$$

is the unique one capable of separate correctly the training data with a maximal margin. It determines the direction $\mathbf{w}/|\mathbf{w}|$, where the distance between the projections of the training vectors of different classes is maximal. If the training data are separable an SVM is a maximum margin classifier. A peculiarity of the SVM 's is that the weights w_i are functions of the support vectors.

The optimal hyperplane (w_0, b_0) are the arguments that maximize the distance in (1) and it is constructed from the

support vectors [4]. Vectors x_i for which $y_i(w \cdot x_i + b) = 1$ will be tagged as support vectors. The vector w_0 that determines the optimal hyperplane can be written as a linear combination of training vectors [4]:

$$\mathbf{w}_0 = \sum_{i=1}^{l} y_i \alpha_i^0 \mathbf{x}_i, \tag{2}$$

where $\alpha_i^0 \geq 0$. Since $\alpha > 0$ only for support vectors, the expression (2) represents a compact form of writing $\mathbf{w}_0$ [4]. To solve the problem of finding the optimal hyperplane (SVM training stage), a constrained optimization problem of maximizing the distance for a given weight vector can be determined by the Lagrangian multiplier method [10]. The SVM training then consist in the implementation of the quadratic problem in (3), minimizing over α_k subject to (4) [11].

$$J = \left(\frac{1}{2}\right) \sum_{hk} y_h y_k \alpha_h \alpha_k \left(x_h \cdot x_k + \lambda \delta_{hk}\right) - \sum_k \alpha_k, \tag{3}$$

$$0 \leqslant \alpha_k \leqslant C \quad \text{and} \quad \sum_k \alpha_k y_k = 0, \tag{4}$$

where $x_k \cdot x_h$ denotes the scalar product, y_k corresponds to the class label, δ_{hk} is the Kronecker symbol and α and C are positive constants (soft margin parameters) that ensure convergence even when the problem is non-linearly separable [11].

2.3.2 Recursive feature elimination (RFE)

Evaluating how one feature contributes to the separation between classes can produce a feature ranking. One of the possible uses for the feature ranking is the design of a classifier on a pre-selected subset of features [11]. Each feature that is correlated with the separation of interest is by itself a class separator. The entries that are associated with larger weights, have a greater influence on the classification decision, therefore if a classifier has a good performance, those entries with the highest weights are the more relevant characteristics [11]. This feature ranking could be obtained during the SVM training stage.

In classification problems the ideal target function is the expected value of the error, this is the error rate calculated on a infinity number of examples, whereas in the training stage this ideal objective function is replaced by a cost function $\mathbf{J}$ estimated only for training patterns. Given this, in [11], the authors introduced the idea of calculating the change in the cost function $\mathbf{DJ}\,(i)$ from removing a single feature or equivalently, from making the weight $\mathbf{w}_i$ zero. Using the change in the cost function when a feature is removed, a feature ranking could be constructed in order to discard the features

with the least ranking value. Nevertheless a good criterion of feature ranking is not necessarily a good criterion for selecting a subset of them. To use the ranking criterion in order to eliminate features, an iterative procedure called recursive feature elimination (RFE) [11] was proposed. The procedure follows as:

1. Train the classifier (optimizing the weights $\mathbf{w}_i$ with respect to $\mathbf{J}$).
2. Compute the ranking criterion for all the features $\mathbf{DJ}(i)$.
3. Remove the feature with the smallest ranking criterion.

This RFE scheme is applied to the SVM algorithm [11]. The feature elimination method can be applied also in non-linear cases [10]. The computations assume no changes in the values of the $\alpha's$ from the SVM training phase, and the cost function to minimize (under the conditions $0 \leq \alpha_k \leq C$ y $\sum_k \alpha_k y_k = 0$)

$$\mathbf{J} = \left(1/2\right) \alpha^T \mathbf{H} \alpha - \alpha^T 1, \tag{5}$$

where $\mathbf{H}$ is the matrix with elements $y_h y_k \mathbf{K}(x_h, x_k)$, $\mathbf{K}$ is a kernel function that measures the similarity between x_h and x_k, and 1 is an l dimensional vector of ones. To calculate the change in the cost function due to the elimination of the component i, the $\alpha's$ remains unmodified and the $\mathbf{H}$ matrix is recalculated. This corresponds to calculating $K(x_h(i), x_k(-i))$, giving the $\mathbf{H}(-i)$ matrix, where the notation $(-i)$ means that the component i has been removed [11]. The resulting ranking coefficient is:

$$DJ(i) = \left(1/2\right) \alpha^T \mathbf{H} \alpha - \left(1/2\right) \alpha^T \mathbf{H}(-i) \alpha. \tag{6}$$

The input with the smallest difference $\mathbf{DJ}(i)$ is eliminated. The elimination of the input with the smallest difference is repeated iteratively producing the recursive feature elimination (RFE) method. The change in the $\mathbf{H}$ matrix must be calculated only for the support vectors [11]. Following the basis of the RFE algorithm, the index m^* of the first feature to remove is $\arg \min_{m \in \{1, \ldots m\}} |\omega_m|$, and generally in the iteration i the same rule of selection is applied to the $M - i$ remaining features.

2.3.3 Margin-maximizing feature elimination

The use of the weights from the trained classifier proposed in RFE algorithm has no consideration of the maximal margin of separation between classes of the SVM. RFE is equivalent to the elimination by maximization of the margin if the following equation is always satisfied [1]:

$$\max_m \min_n \frac{y_n f(x_n) - y_n x_{n,m} w_m}{\sqrt{\|w\|^2 - w_m^2}}$$
$$= \min_n \frac{y_n f(x_n) - y_n x_{n,m^*} w_{m^*}}{\sqrt{\|w\|^2 - w_{m^*}^2}}, \tag{7}$$

where $\mathbf{x}$ are the input examples, $\mathbf{y}$ the corresponding outputs. w is associated to the $DJ(i)$ vector computed by RFE, so w_m corresponds to the ranking coefficient of the m feature.

In order to consider the margin of separation computed from the SV's, a recursive algorithm based in SVM's called margin-maximizing feature elimination (MFE) was proposed in [1]. The authors argue that experimentally they have shown that RFE is not in agreement with margin maximization. RFE is focused on minimally reducing the squared weight vector 2-norm (2), ignoring the margin constraints [1]. The authors in [1] also demonstrate that for the kernel case that the assumption of RFE that the squared weight vector 2-norm is strictly decreasing as features are eliminated is not valid for all the kernels. MFE then propose a feature elimination method based on the recursion over the kernels. For example for the polynomial kernel, $K(u, v) = \exp\left(\|u - v\|^2 / (\gamma^2)\right)$ and denoting $H_{k,n}^{i,m} = H\left(s_k^{i,m}, x_n^{i,m}\right)$ in iteration i the recursion [1]:

$$\mathbf{H}_{k,n}^{i,m} = H_{k,n}^{i-1,m_{i-1}} - \left\|s_{k,m} - x_{n,m}\right\|^2 / (\gamma^2) \forall k, \forall n \tag{8}$$

where s_k corresponds to the support vector k. This recursively calculated kernels are used to evaluate the discriminant function

$$f(x) = \sum_{k \in S} \lambda_{s_k} y_{s_k} K(s_k, x) + b \tag{9}$$

and the weight vector norm through

$$\|\omega\|^2 = \sum_{k \in S} \sum_{l \in S} \lambda_{s_k} y_{s_k} \lambda_{s_l} y_{s_l} K(s_k, s_l) \tag{10}$$

building a MFE-kernel algorithm. The MFE method at each iteration i eliminates the feature m_{MFE} that preserve the maximum positive margin for the training set from the following Eq. [1]:

$$(m_{MFE}, n_{MFE}) = \arg \max_{m \in S = \left\{m' | g_{n'}^{i,m'} > 0, \forall n'\right\}} \min_n \frac{g_n^{i,m}}{\|w\|^{i,m}}, \tag{11}$$

where notation $q^{i,m}$ corresponds to quantity q at feature elimination step i upon elimination of feature m and $g_n^{i,m} = y_n b + \sum_{m=1}^{M} \delta_n^m$ with M being the set of eliminated features and $\delta_n^m = y_n x_{n,m} w_m$.

2.4 Implementation details

For the development of the routines for non-linear feature extraction and the schemes of feature selection based on discriminant methods (SVM's), two toolboxes were used. For the RQA non-linear analysis, the cross recurrence plot toolbox (CRP) [18] is used. The pattern recognition toolbox (PRTools) is used for the routines that allow the training and testing of the SVM's [5]. The implementation of the SVM algorithm is from the PRTOOLS toolbox, and the kernel used for the RFE implementation was the radial basis function (RBF) kernel. An own implementation of the RBF kernel is made for further computing of RFE-SVM and MFE-SVM and combined with PRTOOLS in the "user kernel" mode. The PRTOOLS toolbox uses quadratic programming from the MATLAB Optimization toolbox. The regularization parameter for the SVM, and the parameter of the RBF kernel are estimated using cross-valiadtion over a grid of values for both parameters.

2.5 Validation

To assess the functionality of the discriminant selection algorithms, a validation stage is developed in which we compare the results produced by the algorithms of selection on different data sets against a classical feature space dimensionality reduction scheme such as principal component analysis (PCA) [13]. The PRTools toolbox includes an implementation of the PCA algorithm. PCA uses an orthonormal transformation ton convert possibly correlated variables into linearly uncorrelated variables called principal components. The first principal component should have the largest possible variance, and the succeeding components should have the highest variance. Each component is subject to the constraint that has to be orthogonal with the preceding components [13]. With these reduced spaces, a SVM with a radial basis function is trained to determine the respective classification accuracy.

2.6 Procedure

The signals from the database are processed to obtain the set of features from the linear analysis in Table 1 proposed in [14], and the features from the non-linear analysis RQA for the GSR, HR, and respiration pattern [31].

Using the labels from the database for the arousal and valence dimensions, several datasets are generated for different classification problems. Sets $D1$ and $D2$ correspond to biclass problems for both dimensions with levels from 1–5 to 6–9. In the arousal dimension the range cover the classes of active and passive. For the valence dimension the range cover the classes of pleasant and unpleasant. For the MAHNOB database, datasets are extracted in a equivalent form as the $D1$–$D2$ datasets. In this case the datasets are named as

Table 2 Datasets

Dataset	Description
D1	DEAP database, Arousal dimension, biclass problem, patterns with values 1–5 and 6–9
D2	DEAP database, Valence dimension, biclass problem, patterns with values 1–5 and 6–9
M1	MAHNOB database, Arousal dimension, biclass problem, patterns with values 1–5 and 6–9
M2	MAHNOB database, Valence dimension, biclass problem, patterns with values 1–5 and 6–9

Table 3 Feature indexes for the two databases

DEAP		MAHNOB	
Signal	Index	Signal	Index
GSR	1:11	GSR	1:11
GSR-RQA	12:19	Temp	12:15
Temp	20:23	Resp	16:19
Resp	24:27	HR1	20:29
Resp-RQA	28:35	HR2	30:39
HR	36:45	HR3	40:49
HR-RQA	46:53	EEG	50:273
EEG	54:276		
EOG y EMG	277:288		
Video	289:323		

$M1$–$M2$ for the two spaces of classification. For this database we only take into account the linear features. A summary of the different sets of data generated is presented in Table 2.

For a further quantification analysis, an exact description of the position of each feature into the feature space is presented in Table 3. This information allows a clear understanding of the features that are selected after each feature elimination step, when the algorithms of feature selection are applied.

Based on the different data sets extracted, several experiments are performed to evaluate the performance of RFE and MFE. the RFE-SVM and MFE-SVM algorithms were set to eliminate one feature at each iteration in order to avoid possible elimination of correlated features when a bunch of features is eliminated. A modification of the RFE algorithm proposed in [33] is implemented in order to test possible wrong feature removal from the datasets. This RFE-SVM-CBR implementation is used when RFE-SVM is set to eliminate a bunch of features in each iteration. The size of the final subset of features was computed as the 5 % of the original features set size. This is 16 features for the DEAP database and 14 features for the MAHNOB database.

The classification accuracy (CA) was computed after each iteration of the selection algorithms. Crossvalidation is performed by using 80 % of the patterns for training and 20 %

of the patterns for test. The procedure is repeated 10 times in order to have an statistical validation of the test. The F1-score is a measure of the test accuracy. F1-score is defined as the harmonic mean of precision and recall as the following equation shows:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The F1-score was computed at each iteration of the feature selection algorithms. The vectors with the indexes of the selected features and the CA after each iteration are stored for further analysis. An statistical analysis based on the equal median test is applied to the classification results for each method. The equal median test allows to determine which method has a higher classification accuracy, and if it is statistically different against the other feature selection algorithms [23]. A final test for the RFE and MFE algorithms is made by applying the selection algorithms to EEG features only. The resulting features from the EEG are combined with the other signals for a new selection test.

3 Results

This section presents the results for the different feature selection experiments with RFE and MFE over each dataset, computing the F1-score and the accuracy rate from every feature elimination method performed and presented as the mean and the standard deviation of the ten realizations of each experiment.

3.1 Results for $D1$ dataset

For the $D1$ dataset the Fig. 1a, b show the behavior of the accuracy rate against several feature eliminations using the RFE and MFE algorithms. For the RFE method, the accuracy begins around 72 % for the complete set of features, and only starts to decrease below 65 % when the size of the subset corresponds to less than the 20 % of the original feature space dimension. In the MFE case, the classification accuracy for subsets of 10 % of the original size, reaches a value of around 70 %. The value for the sensibility increases for smaller subsets of features when RFE is applied, in the case of MFE the value of sensibility maintain similar levels for smaller subsets. The specificity tend to decrease at each feature elimination for RFE and for MFE the value of specificity shows variations with each elimination but maintain a similar value compared to the initial value at the original size of features.
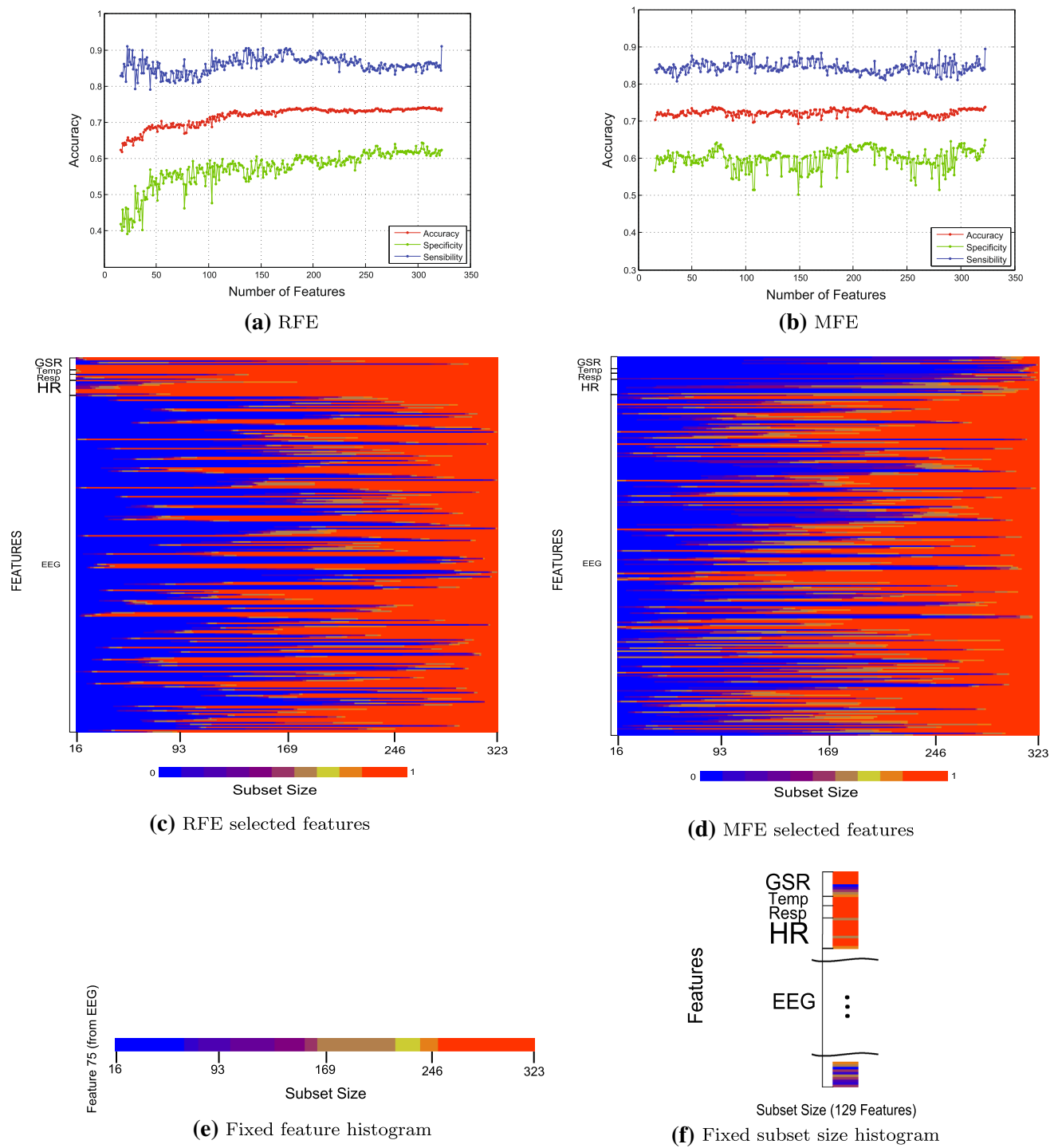
Analysis of the features selected in the different subsets is presented in Fig. 1c, d. These figures show the percentages of occurrence of each feature in different subsets. For example,

when the algorithm is analyzing a subset of features of size 129, and in the ten repetitions of the experiment a particular feature x_h appeared five times, we assign a percentage of 0.5 to the occurrence of that feature. The percentage is represented in a color scale (red for 1 and blue for 0). These histograms allow us to analyze which signals are more relevant at the recognition step, of the different affective states. Notice that the histograms can be understood from two points of view. The first point of view is that given a particular feature x_h, we can see in what percentage that feature appeared when the algorithm analyzed subsets of different size. For example, Fig. 1e shows the frequency of appearance of feature x_{75}, as the size of the features selected changed. The feature corresponds to one of the features extracted from one of the EEG channels. For this single feature, it can be seen how the inclusion of the feature varies for different subsets selected. From the complete size of the set of features (323 features) to 246 features, this particular feature is selected always for all the realizations of the experiment. When the size of the selected subset reaches 169 features, the index of selection is around 0.7 following the color scale. The index of selection continues decreasing as the size of the set of selected features also decreases. When the size of the set is 93, the index of selection is around 0.2, and finally when the elimination of features reaches the smallest size, this feature was not selected in any realization. The second point of view for analyzing the histograms is that given a fixed size of features selected, S, we can see in what percentage each of the available features was included when performing the ten repetitions. For example, Fig. 1f shows the percentage in which each feature was included in the subset of features of size 129. While features 1–5 from GSR show percentages of occurrence of 1, features 6–10 from GSR have percentages of occurrence from 0.1 to 0.4. All temperature features, all respiratory features, and most of the HR features have an index of selection of 1. Due to the large amount of EEG features (222 in total), we only show ten EEG features in Fig. 1f. While some of the EEG features were completely discarded at this stage, others have percentage of occurrence varying from 0.3 to 0.5.

From the index distribution of the features in Table 3, it can be noticed that the features that are discarded in early iterations, when the RFE algorithm is used, are the features from the EEG signal. Some features from the physiological signals are retained despite several feature eliminations. In the case of the MFE algorithm, the features initially discarded are from the physiological signals, while a higher number of features from the EEG are selected in the final subsets.

3.2 Results for $D2$ dataset

Following the same analysis for the $D2$ dataset, corresponding to the biclass problem in the valence dimension, in Fig. 2a,

(a) RFE

(b) MFE

(c) RFE selected features

(d) MFE selected features

(e) Fixed feature histogram
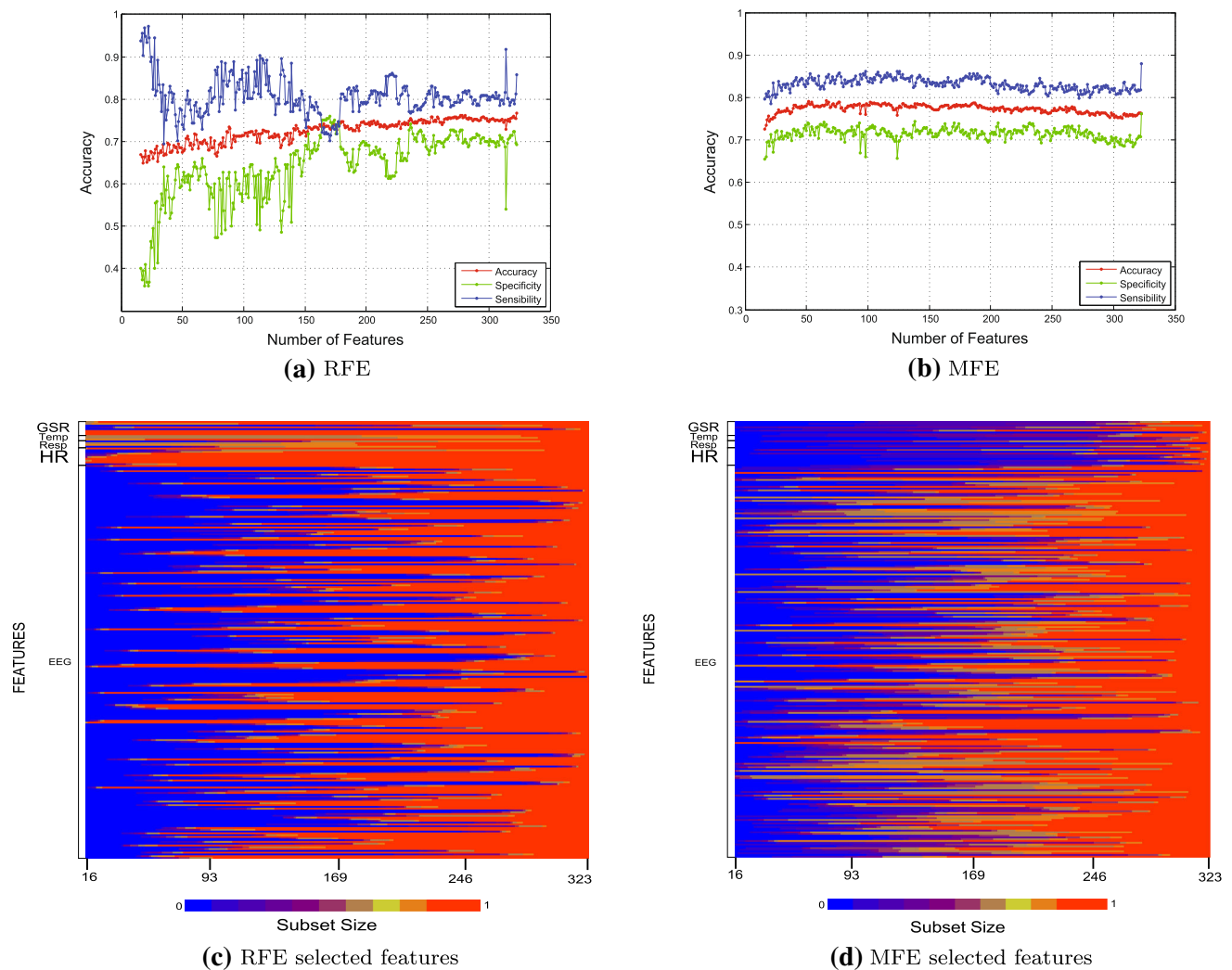

(f) Fixed subset size histogram

Fig. 1 RFE and MFE over $D1$ dataset (mean and standard deviation) and feature apparition in different sized subsets

b the results show a similar behavior compared to the results from $D1$ dataset. An initial accuracy of 73 % is reached using the total set of features in both methodologies, and this percentage remains around 72 % for subsets of less than 30 features when the RFE algorithm is used, also the level of the sensibility show a considerable improvement for the final subset but the specificity level decrease for those smaller subsets of features . In the MFE selected subsets, the initial classification accuracy is maintained even for the final subset as Fig. 2b shows, also the levels of sensibility and specificity maintain similar values for all the selected subsets.

(a) RFE

(b) MFE

(c) RFE selected features

(d) MFE selected features

Fig. 2 RFE and MFE over $D2$ dataset (mean and standard deviation) and feature apparition in different sized subsets

The corresponding percentage of occurrence for the features in different subsets for the $D2$ dataset are presented in Fig. 2c, d. Features from all the physiological signals are selected in smaller subsets using RFE with more influence from all the physiological signals. Some features from the EEG are also selected. With the MFE algorithm the features retained are mostly selected from the EEG as in the $D1$ dataset.

3.3 Results for $M1$ dataset

For the biclass problem in the arousal space using dataset $M1$, the two feature selection methods are employed in a similar manner as in previous experiments. Figure 3 shows the variations in the accuracy rate as several features are eliminated via RFE and MFE. For RFE, Fig. 3a shows an initial success rate of 69 % with the complete set of 273 features, decreasing below rates of 65 % for subsets of less than 50 features. For

the selection using MFE, the initial success rate with all the features is around 69 % and this percentage is maintained despite the different feature eliminations with some slight increase for reduced subsets of less than 50 features with accuracy rates close to 60 %.

Occurrence histograms for the selected features in the different $M1$ subsets are presented in Fig. 3. It can be observed that the features that were not discarded by the RFE algorithm comes from the HR, GSR and EEG, Fig. 3c. When the MFE algorithm is used the features are selected predominantly from the EEG with few coming from the Temperature signal and the respiratory pattern, see Fig. 3d.

3.4 Results for $M2$ dataset

For the valence dimension using $M2$ dataset, both selection algorithms show a similar behavior. From the initial percentage of 65 % using the total set of features, several

(a) RFE

(b) MFE

(c) RFE selected features

(d) MFE selected features
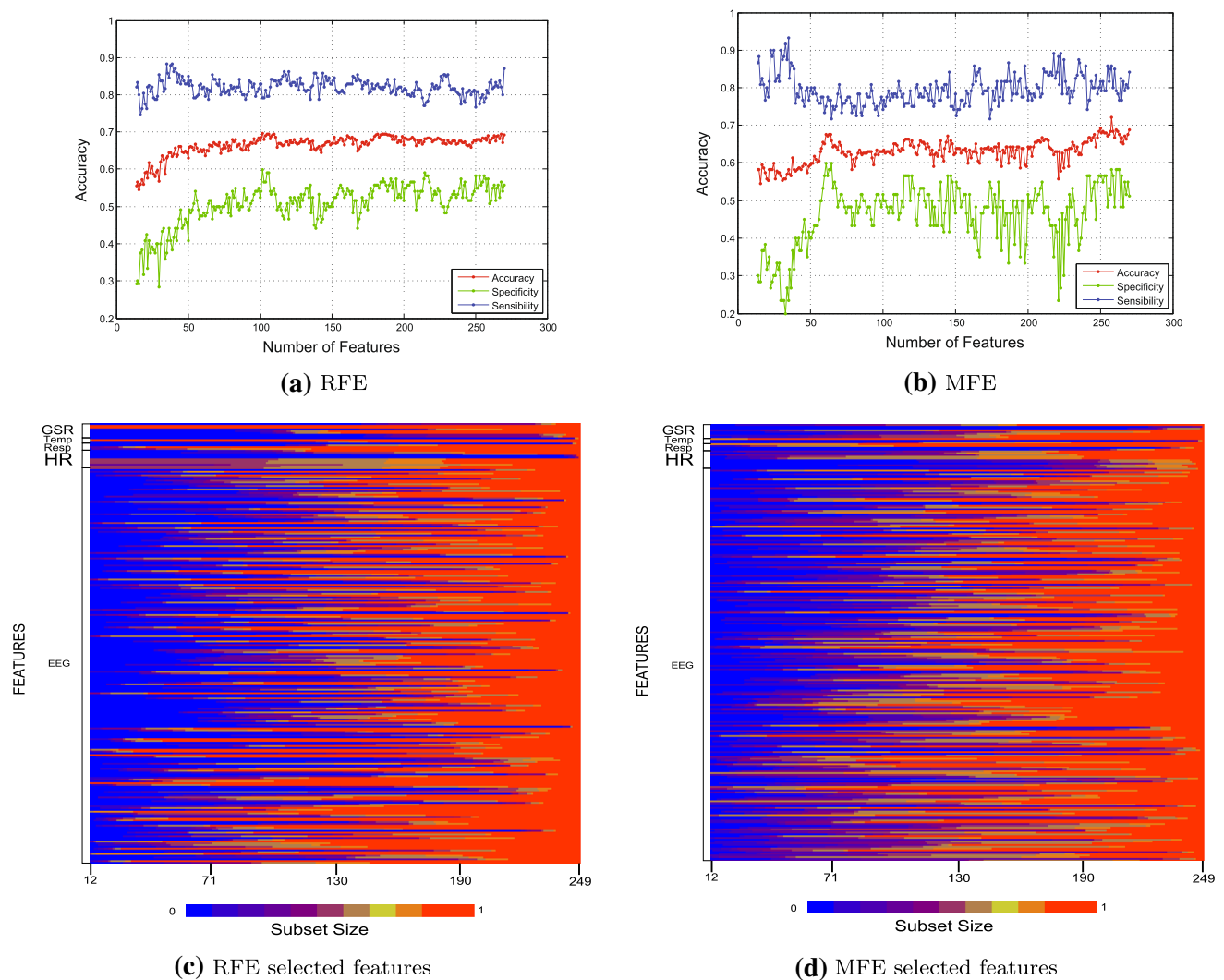
Fig. 3 RFE and MFE over $M1$ dataset (mean and standard deviation) and feature apparition in different sized subsets

feature eliminations are applied and the accuracy is maintained around the initial percentage. The accuracy declines for smaller subsets to percentages around 60 %, as shown in Fig. 4a, b.

From the analysis of the percentage of occurrence, the most selected features come from the HR and the respiratory signal when the RFE algorithm is applied, see Fig. 4c. When MFE is used, the selected features in the smaller subsets come from the EEG signal, respiratory pattern and the temperature signal as well, see Fig. 4d.

A compilation of the classification results from the experiments of feature selection is presented in Table 4. The classification accuracy (CA) and the F1-score are computed in every iteration of the RFE and MFE algorithms for each dataset following the elimination of the less relevant features. The table contains the CA average, maximum CA with the number of features (Nf) where the maximum CA was reached, and the F1-score average for each experiment.

The equal median test is performed over all the results from the two selection algorithms. For all the feature subsets in each experiment, MFE brings better classification results than RFE. The statistical significance analysis based on the equal median test is applied for each dataset selected in each iteration using the RFE and MFE methods. The test allows to determine which method selected the dataset which provides a higher classification accuracy (CA). Over the DEAP database, the equal median test shows that MFE has higher performance than RFE in the selection of 12 datasets in the valence dimension and 22 datasets in the arousal dimension. For the MAHNOB database, the MFE algorithm obtained a higher performance than RFE in the selection of 26 datasets in the arousal dimension and 7 datasets in the valence dimension.

In Table 5, we present a summary of the most selected features in the smallest subset finally obtained by RFE and MFE. We only included the features with a percentage of

(a) RFE

(b) MFE

(c) RFE selected features

(d) MFE selected features
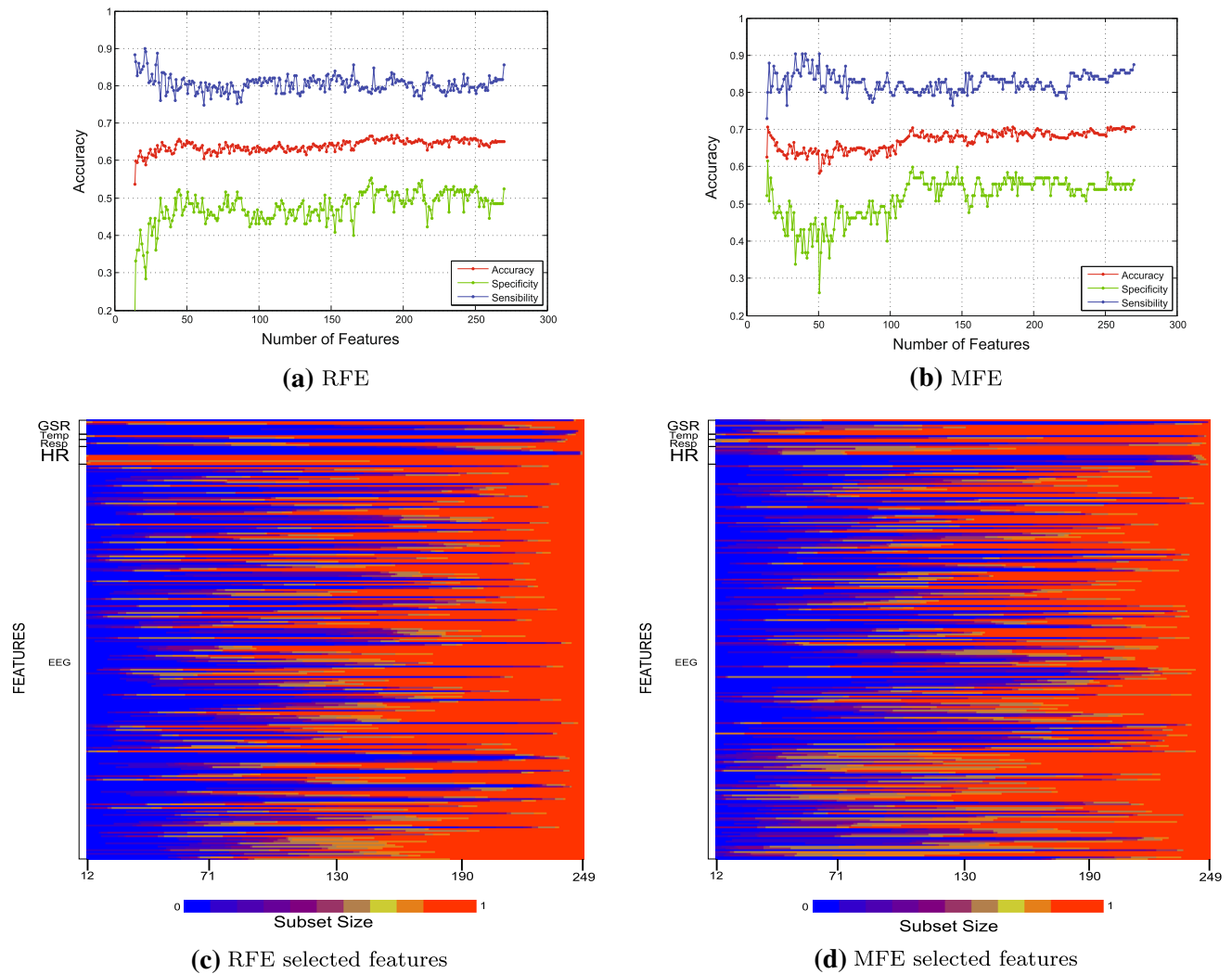
Fig. 4 RFE and MFE over $M2$ dataset (mean and standard deviation) and feature apparition in different sized subsets

Table 4 Classification accuracy and F1-score for RFE and MFE over different datasets

Dataset	Algorithm	Metric		
		CA (average) [%]	CA (max) [%]/ (Nf)	F1-score (average) [%]
D1	RFE	72.09 ± 2.40	74.82 (115)	72.58 ± 3.02
	MFE	72.63 ± 0.82	74.31 (124)	72.65 ± 0.92
D2	RFE	72.35 ± 1.42	74.15 (103)	71.06 ± 1.47
	MFE	73.69 ± 0.80	75.17 (167)	72.07 ± 0.74
M1	RFE	67.01 ± 3.57	69.59 (176)	79.25 ± 1.93
	MFE	64.20 ± 0.87	67.71 (240)	77.92 ± 1.16
M2	RFE	62.58 ± 1.32	66.09 (52)	75.41 ± 1.08
	MFE	63.38 ± 2.20	65.17 (53)	65.17 ± 1.32

occurrence higher than 0.7, both for the arousal and valence dimensions.

As the summary on Table 5 shows, the RFE algorithm selects peripheral signal features in most of the experiments.

Meanwhile, the MFE algorithm gives more relevance to EEG features when choosing the optimal subset. These results are similar in both datasets.

Table 5 Summary of most relevant features

Method	Arousal		Valence	
	Index	Signal	Index	Signal
RFE-DEAP	21, 23	Temp	21, 23	Temp
	26	Resp	24	Resp
	37	HR	37, 41, 44	HR
	52	HR-RQA		
RFE-MAHNOB	2, 3, 10	GSR	2	GSR
	41–46	HR	41, 42, 43, 45, 46	HR
MFE-DEAP	6	GSR	46	HR-RQA
	46	HR-RQA	93, 170	EEG
	65, 74, 80, 156, 217	EEG		
MFE-MAHNOB	11	GSR	9	GSR
	58, 91	EEG	178	EEG

3.5 Feature selection validation

A validation stage for the discriminant feature selection methods is developed. A classical dimensional reduction algorithm is used and the classification accuracy from the reduced subset is obtained for several biclass experiments in each database. In Fig. 5 it can be observed the results for the principal component analysis (PCA) method in comparison against RFE and MFE for the $D1$ and $D2$ datasets. As it can be seen from the results, the performance of the discriminant feature selection algorithms generally have better classification accuracy as the feature space is reduced in comparison to PCA. Note that the performance of PCA in some cases is equal or slightly improves the precision on the classification for some subspaces in comparison with RFE. The selection scheme by MFE has clearly better results in terms of accuracy in all cases, see Fig. 5.

3.6 Additional results

Several additional tests for feature selection were performed in order to asses other relevant aspects of the emotion recognition problem.

– Since the size of the features extracted from all the EEG signals is more than four times the size of the features from the peripheral signals, a feature selection step using RFE and MFE is previously performed only over the EEG features. By doing this, we reduced the number of EEG features from 219 to 50 most relevant features, this is, to a similar size in comparison to the size of the peripheral features (50 for EEG and 49 for the other signals). We then formed a new feature set of 99 features, and we again perform RFE and MFE over this new set. The results obtained for this selection scheme show that MFE continues selecting predominantly EEG features in the final
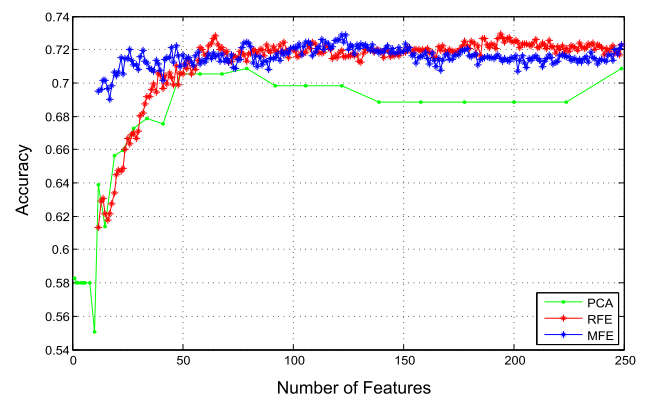
(a) RFE, MFE and PCA over D1 dataset
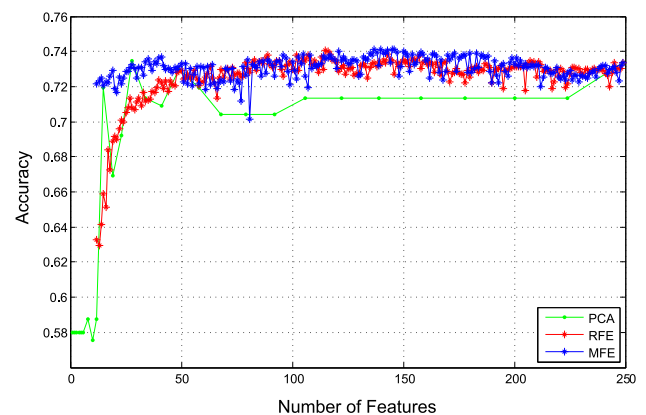
(b) RFE, MFE and PCA over D2 dataset

Fig. 5 Comparison between RFE and MFE against PCA

subset, even when the pre-selection step is performed, achieving similar levels of classification accuracy, around 72 %. RFE keeps selecting heart rate signal features predominantly, reaching classification accuracies around 70 %, for the smallest subsets of features selected. These results allow us to conclude that even in the original test,

when the EEG features outnumbered the other signal features, the RFE and MFE algorithms selected optimal subsets with consistency. Accuracy levels also have similar values in both scenarios: when the number of EEG features is greater than the number of peripheral features, and when the number of EEG features is similar to the number of peripheral features.

– The analysis of correlation proposed in [33] is also performed in an additional scheme. The original selection test using RFE were made by eliminating one feature at each iteration. In this scheme the algorithm is set to remove a bunch of features in each iteration to perform the RFE-SVM-CBR that allows the inclusion of possible misseliminated features into the selected set based on a correlation analysis. The results from this test over the DEAP and MAHNOB database did not show any inclusion of features possibly removed again into the selected subset.

4 Discussion

With all the experiments performed over the different databases using the discriminant feature selection algorithms, the results are consistent with the theory of the recursive feature elimination, where at each iteration a feature (or set of them) is discarded on the basis that it has less relevance for the separation of classes. RFE removes features in each iteration while the classification accuracy remains around the value obtained with the original set of features in both spaces, arousal and valence. Experiments as the one presented in Fig. 1a, show that the classification accuracy for $D1$ dataset, after several eliminations, maintains a constant level. This applies even for subsets with sizes less than half of the original set. This behavior is the same for both databases MAHNOB and DEAP.

Based on the results obtained from the experiments, using the selection algorithm MFE, it can be observed that bunches of features are eliminated retaining a percentage of accuracy of the same magnitude as the total set of features. MFE improves the performance compared to the RFE method in some cases according the statistical analysis. Even with subsets of less than 15 % of the original feature space size, the classification accuracy is close to the value obtained with the whole set of features. The results are similar in all experiments for different classification spaces in biclass problems for both databases. The values of sensibility and specificity that has high relevance in medical studies gives an additional insight about the relevance of the study. In general terms, the sensibility of the classifier improves when less features are used, that is more realizations of the principal class recognized adequately. This behavior has an important relevance

for using a small set of features for an initial detection of the emotional state.

The CA and F1-score metrics presented in Table 4 show a compilation of the classification results. It can be seen that in most of the cases the CA is higher in the MFE experiment and also the F1-score shows that the test accuracy using several subsets of features is also higher in most of the cases for the MFE algorithm. Nevertheless the two selection algorithms prove to effectively reduce the dimension of the feature space without affecting the CA dramatically. Also a statistical test confirms that MFE is superior in the CA metric than RFE, as it was presented in Sect. 3.

Results from the selected features in the different datasets reveals important information of the signals that are more relevant for the emotion recognition problem. For the RFE selection algorithm, it can be seen that the most selected features come from signals as the heart rate and the respiratory pattern for both arousal and valence dimensions for the DEAP database. For the study conducted on the MAHNOB database, the trend in the set of selected features is the greater inclusion of features from the heart rate *HR*. A similar analysis for the MFE algorithm shows from the occurrence histograms that the features from the EEG are selected in both arousal and valence dimensions for the final subset.

The EEG signals and the heart rate signal are the signals from where most of the features were selected. This result would be expected in the sense that the brain and the heart (as part of the central nervous system) are the organs that react more rapidly to an external stimulus [17]. Results show that both RFE and MFE are able to pick on this fact, even when both methods do not select exactly the same features. On the other hand, as it was pointed out in Sect. 2.3.3, while MFE removes features taking into account the margin that separates both classes, and attempts to maximize that margin, RFE only looks for reducing the squared weight vector associated to the SVM, ignoring the margin criterion. For testing the relevance of some signals and features individually, another scheme of selection and classification must be implemented.

Additionally, the experiments for the validation of the discriminant feature selection algorithms showed that RFE and MFE outperforms, in the majority of the experiments, the results obtained with PCA. The results from the MAHNOB database are comparable with the works presented in the state of art for this database. Some of the classification results are even higher than those reported in [15] and [16] that work in a similar framework. In [15], the authors used RFE to select features extracted from the EEG signals, and the video signal. For those features selected, the highest F1-scores were 67.1 % for arousal, and 71.5 % for valence. In our experiments, we obtained an averaged F1-score of 79.25 for arousal, and an averaged accuracy of 75.41 for valence, both results using RFE. In [16], the authors only analyzed the

EEG features. The selection algorithm is based on a sequential search of the best feature subset by an inclusion/exclusion features scheme. The best results reported in this work were 65.1 % in the arousal dimension, and 63.0 % in the valence dimension [16]. Our results show that adding the features from the peripheral signals to the EEG features gives an increment in the F1-score metric in both dimensions, with similar classification accuracy results. Also, the scheme of feature selection from the MFE algorithm outperforms the results obtained with RFE for the smaller subsets of selected features.

5 Conclusions

The discriminant feature selection methods performed successfully in all the experiments by removing several features without affecting the accuracy rate in the classification task. Several experiments show that the emotion classification in the arousal/valence space using the multimodal approach could be improve with an adequate selection feature stage. Nevertheless, any conclusions obtained in terms of the features selected are given in terms of the specific classifier used in this paper. Further studies are needed to assess the performance of other classifiers with the same sets of features selected by our RFE-SVM and MFE-SVM implementations.

For the biclass experiments, the MFE algorithm presented higher classification accuracy than the RFE algorithm for reduced feature subsets in most of the experiments. From the evidence of the results, the more relevant features for emotion classification due to the selection of the features in the smaller subsets are the EEG for both methods. The features from the EEG signal seem to be more relevant in the selection with MFE and the different physiological signals were selected in smaller subsets of features when RFE was applied.

Since there are few works that have made an effort in the feature selection for MER, this work has demonstrated that MER with a stage of feature selection with an embedded methodology based on SVM's could be adapted to this field. Future work could be heading to include audiovisual information and multiclass problems that allows the differentiation of more ranges of Valence and Arousal.

References

1. Aksu Y, Miller DJ, Kesidis G, Yang QX (2010) Margin-maximizing feature elimination methods for linear and nonlinear kernel-based discriminant functions. Trans Neur Netw 21(5):701–717
2. Bouaguel W, Bel Mufti G, Limam M (2013) A fusion approach based on wrapper and filter feature selection methods using majority vote and feature weighting. In: Computer applications technology (ICCAT), 2013 International conference on, pp 1–6
3. Cheng K-S, Chen Y-S, Wang T (2012) Physiological parameters assessment for emotion recognition. In: Biomedical engineering and sciences (IECBES), 2012 IEEE EMBS conference on, pp 995–998
4. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
5. Duin RPW (2000) Prtools version 3.0: a matlab toolbox for pattern recognition. In: Proceedings of SPIE, p 1331
6. Eckmann J-P, Oliffson Kamphorst S, Ruelle D (1987) Recurrence plots of dynamical systems. EPL (Europhys Lett) 4(9):973
7. Ekman P (1972) Universals and cultural differences in facial expressions of emotions. I:n Cole J (ed) Nebraska symposium on motivation, vol 19, pp 207–283. Lincoln University of Nebraska Press
8. García-López F, García-Torres M, Melián-Batista B, Moreno-Pérez J, Moreno-Vega JM (2006) Solving feature subset selection problem by a Parallel Scatter Search. Eur J Op Res 169(2):477–489
9. Greco A, Valenza G, Lanata A, Rota G, Scilingo EP (2014) Electrodermal activity in bipolar patients during affective elicitation. Biomed Health Inform IEEE J 18(6):1865–1873
10. Gu Y, Tan SL, Wong KJ, Ho MHR, Qu L (2009) Using ga-based feature selection for emotion recognition from physiological signals. In: Intelligent signal processing and communications systems, 2008. ISPACS 2008. International symposium on, pp 1–4
11. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1–3):389–422
12. Hanjalic A, Xu L-Q (2005) Affective video content representation and modeling. Multimed IEEE Trans 7(1):143–154
13. Jolliffe Ian (2005) Principal component analysis. Wiley, Oxford
14. 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 Trans Affect Comput 3(1):18–31
15. Koelstra S, Patras I (2013) Fusion of facial expressions and eeg for implicit affective tagging. Image Vis Comput 31(2):164–174
16. Kortelainen J, Seppanen T (2013) Eeg-based recognition of video-induced emotions: Selecting subject-independent feature set. Conf Proc IEEE Eng Med Biol Soc 4287–4290:2013
17. Kreibig SD (2010) Autonomic nervous system activity in emotion: a review. The biopsychology of emotion: current theoretical and empirical perspectives. Biol Psychol 84(3):394–421
18. Marwan N, Wessel N, Meyerfeldt U, Schirdewan A, Kurths J (2002) Recurrence Plot Based Measures of Complexity and its Application to Heart Rate Variability Data. Phys Rev E 66(2):026702
19. Marwan N, Romano MC, Thiel M, Kurths J (2007) Recurrence plots for the analysis of complex systems. Phys Rep 438(5–6):237–329
20. Nicolaou MA, Gunes H, Pantic M (2011) Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. Affect Comput IEEE Trans 2(2):92–105

21. Paleari M, Chellali R, Huet B (2010) Features for multimodal emotion recognition: an extensive study. In: Cybernetics and intelligent systems (CIS), 2010 IEEE conference on, pp 90–95

22. Park BJ, Jang EH, Kim SH, Huh C, Sohn JH (2011) Feature selection on multi-physiological signals for emotion recognition. In: Engineering and industries (ICEI), 2011 international conference on, pp 1–6

23. Pizarro J, Guerrero E, Galindo PL (2002) Multiple comparison procedures applied to model selection. Neurocomputing 48(174):155–173

24. Pun T, Pantic M, Soleymani M (2012) Multimodal emotion recognition in response to videos. IEEE Trans Affect Comput 3(2):211–223

25. Rozgic V, Ananthakrishnan S, Saleem S, Kumar R, Prasad R (2012) Ensemble of svm trees for multimodal emotion recognition. In: Signal information processing association annual summit and conference (APSIPA ASC), 2012 Asia-Pacific, pp 1–4

26. Soleymani M, Lichtenauer J, Pun T, Pantic M (2012) A multimodal database for affect recognition and implicit tagging. IEEE Trans Affect Comput 3(1):42–55

27. Soleymani M, Chanel G, Kierkels JJM, Pun T (2008) Affective characterization of movie scenes based on content analysis and physiological changes. Int J Semant Comput 2:235–254

28. Takahashi K, Namikawa S, Hashimoto M (2012) Computational emotion recognition using multimodal physiological signals: elicited using Japanese Kanji words. In: Telecommunications and signal processing (TSP), 2012 35th international conference on, pp 615–620

29. Tokuno S, Tsumatori G, Shono S, Takei E, Suzuki G, Yamamoto T, Mituyoshi S, Shimura M (2011) Usage of emotion recognition in military health care. Defense science research conference and expo (DSR) 2011, pp 1–5

30. Valenza G, Citi L, Gentili C, Lanata A, Scilingo EP, Barbieri R (2015) Characterization of depressive states in bipolar patients using wearable textile technology and instantaneous heart rate variability assessment. Biomed Health Inform IEEE J 19(1):263–274

31. Valenza G, Lanata A, Scilingo EP (2012) The role of nonlinear dynamics in affective valence and arousal recognition. Affect Comput IEEE Trans 3(2):237–249

32. Wang Y, Mo J (2013) Emotion feature selection from physiological signals using tabu search. In: Control and decision conference (CCDC), 2013 25th Chinese, pp 3148–3150

33. Yan K, Zhang D (2015) Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sensors Actuators B Chem 212:353–363

34. Zbilut Joseph P, Webber Charles L (2006) Recurrence quantification analysis. Wiley encyclopedia of biomedical engineering. Wiley, Oxford

35. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual and spontaneous expressions. In: ICMI '07: Proceedings of the 9th international conference on multimodal interfaces, pp 126–133, New York, NY, USA, 2009. ACM