Clinical Neurophysiology 120 (2009) 264–274

Contents lists available at ScienceDirect

Clinical Neurophysiology

journal homepage: www.elsevier.com/locate/clinph

Model comparison for automatic characterization and classification of average ERPs using visual oddball paradigm

A.C. Merzagora [a,*], M. Butti [b], R. Polikar [c], M. Izzetoglu [a], S. Bunce [d], S. Cerutti [b], A.M. Bianchi [b], B. Onaral [a]

[a] School of Biomedical Engineering, Science and Health Systems, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA
[b] Department of Biomedical Engineering, Polytechnic University of Milan, Italy
[c] Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ 08028, USA
[d] Department of Psychiatry, Drexel University College of Medicine, Philadelphia, PA 19102, USA

ARTICLE INFO

Article history:
Accepted 11 October 2008
Available online 4 December 2008

Keywords:
EEG
ERP
Attention
P300
N200
Oddball
Pattern recognition
Linear discriminant analysis
Support vector machines
Neural networks
Target categorization
SVM
MLP

ABSTRACT

Objective: To determine whether automated classifiers can be used for correctly identifying target categorization responses from averaged event-related potentials (ERPs) along with identifying appropriate features and classification models for computer-assisted investigation of attentional processes.
Methods: ERPs were recorded during a target categorization task. Automated classification of average target ERPs versus average non-target ERPs was performed by extracting different combinations of features from the P300 and N200 components, which were used to train six classifiers: Euclidean classifier (EC), Mahalanobis discriminant (MD), quadratic classifier (QC), Fisher linear discriminant (FLD), multi-layer perceptron neural network (MLP) and support vector machine (SVM).
Results: The best classification performance (accuracy: 91–92%; sensitivity: 85–86%; specificity: 95–99%) was provided by QC, MLP, SVM on feature vectors extracted from P300 recorded at multiple sites. In general, non-linear and non-parametric classifiers (QC, MLP, SVM) performed better than linear classifiers (EC, MD, FLD). The N200 did not explain variance beyond that of P300 recorded at multiple sites.
Conclusions: The results suggest that automatic characterization and classification of average target and non-target ERPs is feasible. Features of P300 recorded at multiple sites used to train non-linear classifiers are recommended for optimal classification performance.
Significance: Automatic characterization of target ERPs can provide an objective approach for detecting and diagnosing abnormalities and evaluating interventions for clinical populations, paving the way for future real-time monitoring of attentional processes.

1. Introduction

Attention is generally recognized as a complex cognitive process: it allows for proper allocation of processing resources based on the relevance of a stimulus and regulates the competition between different information channels (e.g. auditory and visual channels) (Egeth and Yantis, 1997; Pessoa et al., 2003). Attention is an area of intense interest and investigation because it plays a critical role in the execution of everyday tasks and goal-directed behaviors.

The "oddball" paradigm has been widely used for the investigation of attentional processes, their meaning and their neural correlates (Basar-Eroglu et al., 1992, 2001; Polich, 1997; Ravden and Polich, 1999; Struber and Polich, 2002; Sutton et al., 1965). The oddball paradigm is a simple discrimination task in which subjects are presented with two (or more) stimuli or classes of stimuli in a pseudo-random sequence. The probability of occurrence of one type of stimulus is typically less than that of the others, e.g. 20% of the trials might be designated as "target" stimuli, whereas the remaining 80% of trials would be "non-target" or "context" stimuli. The participant's task is to count, identify or respond to the designated target stimulus. If correctly recognized by the subject, target stimuli have been shown to elicit a characteristic brain response, a so called event-related potential (ERP), that is predominant in medial parietocentral regions (Polich and Kok, 1995). Target classification typically elicits a negative deflection (N200) that occurs about 200 ms after the presentation of the infrequent target stimulus, followed by a positive deflection (P300) occurring about 300–500 ms after the target stimulus onset (Naatanen and Picton, 1986; Polich et al., 1985).

The oddball paradigm studies have demonstrated that the average amplitude of the P300 component is sensitive to the probability of task-relevant events, manifesting greater amplitude to the infrequent "target" stimulus. In addition, a target stimulus with a frequency of 10% elicits a P300 with amplitude that is on average

* Corresponding author. Tel.: +1 215 895 1988; fax: +1 215 895 4983.
E-mail address: a.merzagora@drexel.edu (A.C. Merzagora).

1388-2457/$34.00
doi:10.1016/j.clinph.2008.10.016

greater than that elicited by a target stimulus with a frequency of say, 40%, both of which will be greater than the P300 amplitude to the non-target (context) stimulus. Furthermore, it is the subjective probability, rather than the objective probability, that dictates the amplitude of the P300. Stimulus categories can be as varied as the letter "S" versus the letter "H"; male versus female names; the affective valence of a picture, pictures of famous people versus unfamiliar people or, indeed, the absence of a stimulus. Under these circumstances, the P300 has been found to be related to the attention resources allocated to the task (Kok, 2001; Polich, 2003). Anomalies in P300 amplitude and latency have been shown to be present in mental disorders and neurological diseases, such as schizophrenia (Bramon et al., 2004; Coburn et al., 1998; Ford et al., 1992), dementia and Alzheimer's disease (Missonnier et al., 1999; Polich et al., 1985; Polikar et al., 2008; Sumi et al., 2000), attention deficit/hyperactivity disorder (Sangal and Sangal, 2006) and traumatic brain injury (Keren et al., 1998; Lew et al., 2005). Moreover, pharmacological interventions that target the attention domain have also been shown to affect the amplitude and latency of the P300 (Coburn et al., 1998; Sangal and Sangal, 2004, 2006).

The majority of the studies about the aforementioned abnormalities and modulation of P300 have so far been based on direct statistical comparison of the responses to target and non-target stimuli. Still, a computer-assisted characterization of the oddball responses could be a valuable supplement to a standard clinical evaluation. In general, using automated classification allows individual diagnosis/characterization, which is typically not possible with sample/population-based statistical analysis. The main potential advantages of such a characterization would in fact be 2-fold, as suggested by Coburn and colleagues (Coburn et al., 2006). First, it would help in the detection and quantification of abnormalities in brain activity through the comparison of a subject's average target ERP with a normative healthy database. Second, comparing a subject's ERP to a database of ERPs recorded from a variety of clinical populations would provide quantitative information that could be valuable for the diagnosis of psychiatric illnesses and mental disorders. Additionally, it would be possible to use a collection of ERP features for the classification of subjects or patients into groups of clinical interest by means of a multivariate comparison with clinical and normative healthy databases.

In this paper, we present a preliminary assessment of algorithms for the quantitative characterization of average target ERPs obtained from a pool of healthy subjects during a visual target categorization task. In particular we aim at discriminating between average ERP responses to target and non-target stimuli. The non-target ERPs are used as a proxy for the abnormal average target ERP; it is in fact fairly reasonable to assume that methods that reliably discriminate between target and non-target ERPs hold promise for the discrimination between normal and abnormal target ERPs. Additionally, the same algorithms assessed for their ability in differentiating between target and non-target ERPs can be further investigated for their performance in the automated classification of target and non-target ERPs on a single-trial basis. Automatic recognition of single-trial target ERPs would in fact offer a useful tool for the real-time monitoring of the attention level. Recent studies suggest that monitoring the attention level and its fluctuations throughout a task may indeed benefit from a single-trial analysis of the ERP responses, as decreases or fluctuations across the time-on-task may be correlated to disease processes (Holm et al., 2006; Roschke et al., 1996; Tomberg and Desmedt, 1999).

In order to achieve a reliable discrimination between target and non-target ERPs, we evaluate the quantitative characterization of the responses based on multiple features and different combinations thereof. In fact, in addition to the widely-accepted use of the averaged P300 amplitude, we investigated the potential contribution of the N200 component in automated target categorization. Like the P300, the N200 is elicited by target stimuli in target categorization tasks, but its suggested cognitive determinants are different. Neurocognitive studies suggest that N200 amplitude reflects the cognitive resources allocated to conflict monitoring processes (Donkers and van Boxtel, 2004; Nieuwenhuis et al., 2003). This is based on the observation that the N200 peak appears when, in order to respond correctly to an infrequent target, the subjects must override their habituated response to the frequent non-target stimuli (Botvinick et al., 2004).

In recognition of the no-free-lunch theorem, which proves that no classifier (or statistical or probabilistic model) is superior to all other classifiers in the absence of additional information (Wolpert and Macready, 1997) and that different models must be evaluated and compared against each other for any given application, we evaluated six different classification algorithms, each of which employs a different model fitting structure: the Euclidean classifier (EC), the Mahalanobis discriminant (MD), the quadratic classifier (QC), the Fisher linear discriminant (FLD), the multi-layer perceptron neural network (MLP) and the support vector machine (SVM).

2. Methods

2.1. Experimental protocol

A total of 16 healthy adults (4 females) participated in the study. Participants were right-handed non-smokers, with vision correctible to 20/20. Participants denied any history of neurological disorders, psychiatric illness, substance abuse or being on any current medication. The experimental protocol was approved by the Institutional Review Board at Drexel University and all participants gave their written informed consent after a detailed explanation of the procedure. The mean age of the participants was 20.8 years (standard deviation = 4.2 years).

Participants were seated in a dimly-lit, sound attenuated room. ERPs were recorded from two surface Ag/AgCl electrodes placed at International 10–20 System locations Cz and Pz, referenced to linked mastoid leads. The choice of these two electroencephalography (EEG) sites originated from the effort to investigate the feasibility of an attention monitoring tool that could be used in clinical settings, therefore a simplified framework was sought. Furthermore, the P300 peak, arguably among the most widely investigated EEG features in attention studies, is in fact known to be maximal at the midline central and parietal sites (Polich and Kok, 1995).

Vertical and horizontal electrooculograms (VEOG and HEOG) were monitored via electrodes placed above and below the left eye, and at the left and right outer canthi, respectively. ERP signals were collected using a SynAmps amplifier (Neuroscan Inc., El Paso, TX); all impedances were systematically kept below 10 kΩ and the amplification was set to 50 mV/mm. EEG signals were filtered between 0.15 and 100 Hz (−6 dB/octave), using an analog filter, and sampled at 500 Hz.

Participants were asked to perform a visual discrimination task. Visual stimuli were presented on a computer monitor using STIM (Neuroscan, Inc.) software. Stimuli consisted of two strings of white letters (XXXXX and OOOOO) presented against the center of a dark background. A total of 516 stimuli were presented, 480 non-target stimuli (OOOOO; 93.02%) and 36 target stimuli (XXXXX; 6.98%). Stimulus duration was 500 ms, with an interstimulus interval of 1500 ms. Target stimuli were presented randomly with respect to non-target stimuli and a minimum of 12 non-target stimuli were presented between successive targets. However, to prevent the participants from developing expectations about the pattern of target presentation, 4 of the 36 target stimuli were presented more closely together. Participants were required to press

one of two buttons on a response pad after each stimulus, using the index finger of their non-dominant (left) hand for non-target stimuli and the middle finger of the same hand to identify targets. Behavioral accuracy and response times were also recorded through the STIM program.

2.2. Data analysis

2.2.1. Preprocessing

Eyeblink artifacts were minimized using Jung's Independent Component Analysis (ICA) approach (Jung et al., 2000a,b). Stimulus-locked ERPs were extracted in 1000 ms epochs, using a 300 ms pre-stimulus baseline and a 700 ms post-stimulus response window. Epochs were baseline corrected by subtracting the mean of the baseline window from the full epoch. Epochs containing significant movement or muscle artifact were discarded, and only epochs containing correct subject responses were included in the analysis. Mean target responses were calculated by averaging across the remaining target stimuli for each subject and channel. On average, this yielded to 30 available target trials (minimum number: 22; maximum number: 36). To avoid creating a bias in the signal to noise ratio for target and non-target stimuli, a random sample of non-target trials was selected from the 480 non-target trials; the size of the non-target trials subsample matched that of the target trials for each given individual.

2.2.2. Feature extraction and selection

N200 and P300 peaks were automatically identified at each channel and a series of features describing these peaks was extracted. The N200 peak was identified as the largest negative deflection in the 160–330 ms post-stimulus response, whereas the P300 was identified as the largest positive deflection in the 250–480 ms post-stimulus response. The features considered in this study were the amplitudes of both N200 and P300 peaks, as well as the amplitude differences between these peaks. A total of 6 features were extracted from the signals, 3 features each from electrode sites Cz and Pz: amplitude of the N200, amplitude of the P300, and the amplitude difference between N200 and P300. These features were verified through an ANOVA test to show statistical significance between the target and non-target classes.

Three different sets of feature vectors were formed from these individual features extracted from the average ERPs:

- *Feature Set 1:* the feature vector consisted of a single element, the P300 amplitude at Pz.
- *Feature Set 2:* the feature vector consisted of two elements: the P300 amplitudes at Pz and Cz.
- *Feature Set 3:* principal component analysis (PCA) was used to determine which minimum set of independent linear combination of the six features accounted for the most variation in the data. Through PCA, we have retained only the principal components which cumulatively accounted for 98% of the total variance of the data (for which three components were found to be adequate).

Per subject single-trial target ERPs (and randomly selected single-trial non-target ERPs) were averaged to obtain one average target ERP (and one average non-target ERP). Hence, a total of two (averaged) signals were obtained for each of the 16 subjects. Under each of the three feature sets, the pertinent features extracted from the ERPs were combined in feature vectors, one for the target ERP and one for the non-target ERP of each subject. Each of these feature vectors represented an instance (or sample) $\boldsymbol{x}_i$ and was associated with a label y_i that stated if $\boldsymbol{x}_i$ was a feature vector extracted from a target ERP (y_i = "target") or from a non-target ERP (y_i = "non-target"). The total number of feature vectors (i.e., instances $\boldsymbol{x}_i$) in each feature set was 32 and constituted the overall set $\boldsymbol{S}$ of available instances: $\boldsymbol{S} = [\boldsymbol{x}_i, y_i]$.

2.2.3. Classification

We investigated the relative performance of the following six classifiers:

Euclidean (minimum distance) classifier (EC): The instance $\boldsymbol{x}_i$ is assigned to the class whose training data mean is closest to $\boldsymbol{x}_i$, based on the Euclidean distance. For normally distributed data, under the assumption that all variances are identical and all cross-variances are zero for all classes (i.e., the covariance matrix is a constant multiple of the unit matrix), the minimum distance classifier is equivalent to a Bayes classifier and hence is statistically the optimum classifier (Duda et al., 2001).

Mahalanobis discriminant (MD): The instance $\boldsymbol{x}_i$ is assigned to the class whose training data is closest to $\boldsymbol{x}_i$, based on the Mahalanobis distance. This classifier is equivalent to the optimum Bayes classifier if the data is normally distributed with identical (but arbitrary) covariance matrices for all classes (Duda et al., 2001).

Quadratic classifier (QC): Feature vectors are labeled using a Bayesian error minimization approach, under the more general hypothesis that the covariance matrices for all classes can assume any arbitrary value (Duda et al., 2001; Kuncheva, 2004).

Fisher linear discriminant (FLD): It is a linear classifier that projects high-dimensional data onto a smaller dimensional space that maximizes the separability between the groups. A simplified discrimination is then performed in the projected space (Duda et al., 2001) using minimum-error-rate classification, assuming a multivariate normal distribution of the data.

Multi-layer perceptron neural network (MLP): The MLP is a feed-forward neural network that consists of several nodes grouped in an input layer, one or more hidden layers and an output layer. In this architecture, the hidden layer maps the inputs to a non-linear space (where the features are presumably better separable) and the output layer implements a (non-linear) discriminant function in this new space. The mixing weights of the inputs are iteratively adapted to minimize an error criterion function on the training data through a gradient-descent based optimization algorithm, called the backpropagation (Haykin, 1999; Werbos, 1974). The number of input nodes is determined by the number of features in the feature vector. The number of nodes in the hidden layer is typically selected using a k-fold cross-validation approach ($k = 8$ was used in this study). In such an approach, the dataset is partitioned into k blocks; multiple MLPs with different number of hidden nodes are trained on $k - 1$ subsets and tested on the remaining kth block. The number of hidden layer nodes that provides the best performance over k trials is then chosen. In this application, the number of output nodes was one, whose computed value in the $[-1\ 1]$ interval determined the MLP predicted stimulus type as non-target or target.

Support vector machine (SVM): Support vector machines are binary classifiers that use a non-linear mapping kernel function to transform the given data into a higher dimensional space, where the data is believed to be linearly separable. Classification is then performed in the new space by finding the optimal hyperplane that offers the maximum separating margin between the closest samples of the two classes. The performance of a given SVM depends also on a tradeoff parameter C: the C parameter balances the relative importance of minimizing the training error and maximizing the margins between the classes, which directly affect the classifier's generalization ability. In this work, a Gaussian radial basis function was used as the kernel. As we have done for the MLP model selection, a k-fold validation ($k = 8$) was used to choose the standard deviation σ of the kernel and the C parameter for each of the three feature sets.

For training and testing each of the six classifiers, a modified leave-one-out (mLOO) cross-validation was implemented (Fig. 1). One instance $(\boldsymbol{x}_i, y_i)$ from of the available 32 instances in the set $\boldsymbol{S}$ was removed to be used as a test data point. The remaining 31 instances formed the subset $\boldsymbol{S}^{(i)}$. From $\boldsymbol{S}^{(i)}$, 20 instances were randomly selected to serve as training data – 10 representing the target ERP and 10 representing the non-target ERP – forming the training subset $\boldsymbol{TS}^{(i)}_{(r)}$. One classifier (of each of the six types) was trained on this training dataset. This process was repeated 10 times, in each case randomly choosing a different set of 20 (of the 31) instances, creating 10 training sets $\boldsymbol{TS}^{(i)}_{(r)}$ $r = 1, 2, \ldots, 10$, and corresponding 10 classifiers with slightly different decision boundaries $g^{(i)}_{(r)}$ $r = 1, 2, \ldots, 10$. These 10 classifiers were evaluated on the one test data point $(\boldsymbol{x}_i, y_i)$ that was previously left out. This entire process – generating 10 training data subsets of 20 instances and training 10 corresponding classifiers – was repeated a total of 32 times, once for each data point to be used as test data. The pseudo-code in Fig. 1 describes this modified leave-one-out procedure in detail.

The available 32 instances in the set $\boldsymbol{S}$ were classified 10 times using the mLOO, allowing a statistical characterization of the following three performance indices:

1. *accuracy*, defined as the probability of correctly classifying an instance, and computed as the percentage of correctly classified instances (out of the 32 available ones);
2. *sensitivity*, defined as the probability of the test to correctly identify the target class and computed as the ratio of the number of correctly classified targets to the total number of target instances ($n = 16$);
3. *specificity*, defined as the probability of the test to correctly identify the non-target class and computed as the ratio of the number of correctly classified non-targets to the total number of non-target instances (the remaining 16 instances).

The following statistical analyses were conducted on the performance results:

1. a one-way ANOVA to determine whether the differences in classification performances of *different classifiers* are statistically significant, where the classifier type was used as a factor with six levels;
2. a one-way ANOVA to determine whether the differences in classification performances obtained with *different feature sets* are statistically significant, where feature sets are used as a factor with three levels;
3. a one-way ANOVA to determine whether different classes of classifiers, such as "parametric linear" (FLD, MD, EC), "parametric non-linear" (Q) or "non-parametric" (MLP, SVM), have classification performances that are significantly different than those of others;
4. a two-way ANOVA to determine the interaction between the choice of the feature set and the linearity of the classifiers.

In all cases, if a significant difference was found at $\alpha = 0.05$ level, individual factors were compared against each other for pair-wise statistical significance using the multiple comparison (Tukey–Kramer post-hoc) test with a 95% level of significance.

❖ **Input**:
- Set of instances $\boldsymbol{S} = [(\boldsymbol{x}_i, y_i)]$, i=1,2,…,32 , with labels $y_i \in$ {"target", "non-target"}.
- Classification algorithm *CL*.

❖ **Procedure**:

Do for each i=1,2,…,32 :
- Obtain subset $\mathbf{S}^{(i)}$ by removing the instance $(\boldsymbol{x}_i, y_i)$ from set $\mathbf{S}$:

$$\boldsymbol{S}^{(i)} = \overline{\boldsymbol{S} \cap (\boldsymbol{x}_i, y_i)}$$

- **Do** for r = 1,2,…,10 :
  - Obtain training subset $\boldsymbol{TS}^{(i)}_{(r)}$ by drawing 10 target instances and 10 non-target instances from $\boldsymbol{S}^{(i)}$.
  - Train classifier *CL* on $\boldsymbol{TS}^{(i)}_{(r)}$, obtaining decision boundary $g^{(i)}_{(r)}$.
  - Based on $g^{(i)}_{(r)}$, determine predicted label $\tilde{y}^{(i)}_{(r)}$ for instance $(\boldsymbol{x}_i, y_i)$.

Create matrix of all predicted labels for all instances in $\mathbf{S}$:

$$\tilde{\boldsymbol{Y}} = \begin{bmatrix} \tilde{y}^{(1)}_{(1)} & \cdots & \tilde{y}^{(i)}_{(1)} & \cdots & \tilde{y}^{(32)}_{(1)} \\ \vdots & \ddots & & & \vdots \\ \tilde{y}^{(1)}_{(r)} & & \tilde{y}^{(i)}_{(r)} & & \tilde{y}^{(32)}_{(r)} \\ \vdots & & & \ddots & \vdots \\ \tilde{y}^{(1)}_{(10)} & \cdots & \tilde{y}^{(i)}_{(10)} & \cdots & \tilde{y}^{(10)}_{(32)} \end{bmatrix}$$

Compute accuracy $Ac_{(r)}$, sensitivity $Se_{(r)}$ and specificity $Sp_{(r)}$ for each row r of $\tilde{\boldsymbol{Y}}$.

❖ **Output**: Statistical characterization of *Ac*, *Se* and *Sp* for the classification algorithm *CL*.

Fig. 1. Pseudo-code description of the modified leave-one-out (mLOO) procedure used for cross-validation of each of the six classifiers.

3. Results

3.1. Behavioral results

The average stimulus–response accuracy achieved by the study participants was 90.6% (standard deviation: 0.07%). A paired *t*-test ($t(15) = -7.84$, $p < 0.001$) revealed that response times differed for non-target (272 ± 39 ms) versus target (407 ± 57ms). This finding is a reasonable consequence of the ratio of target to non-target responses.

3.2. Feature extraction

We first computed the grand averages of the ERPs obtained from the Cz and Pz electrodes in response to the two different stimuli (targets and non-targets). These grand averages are shown in Fig. 2 for channels Cz (Fig. 2A) and Pz (Fig. 2B), where the thick solid line is the average response to the target stimuli and the thin dashed line is the average response to the non-target stimuli.

As described above, N200 and P300 amplitudes, as well as amplitude differences between N200 and P300, were extracted from the individual average target and non-target ERPs to be used as features. Table 1 reports the amplitudes of the N200 and P300 peaks and their differences for Cz and Pz channels, respectively, obtained as grand averages over all subjects. From these six features, we derived the previously described three feature sets on which classification was performed.

3.3. Results for Feature Set 1

Feature Set 1 consisted of the average amplitude of the P300 peak as recorded at Pz. Fig. 3A shows the distribution of this fea-

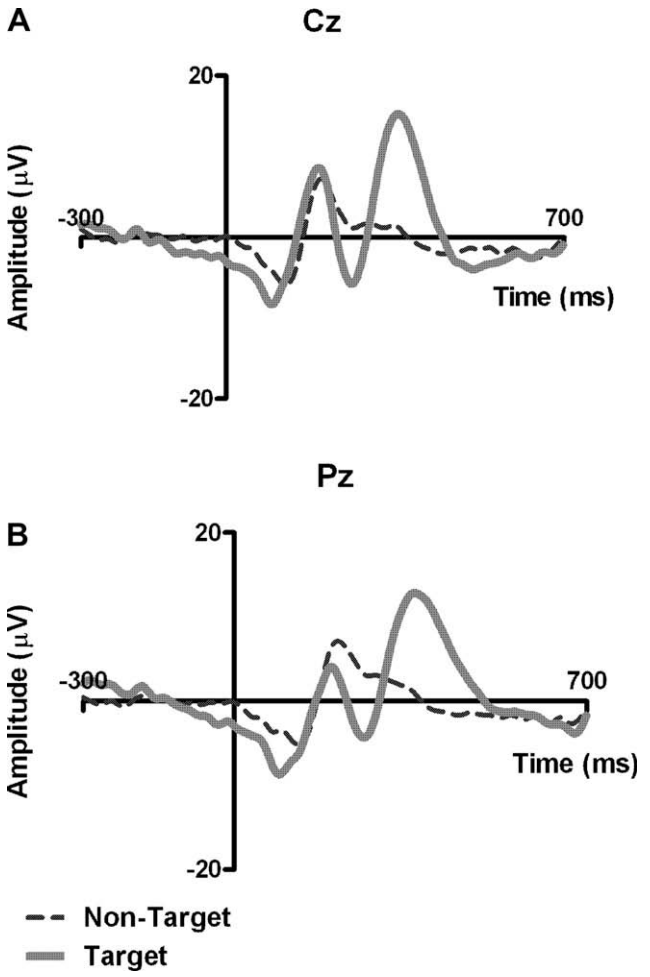

Fig. 2. Grand averages of ERPs recorded at Cz (A) and at Pz (B). Thick lines represent average ERPs elicited by infrequent targets; average ERPs elicited by frequent non-targets are represented by the thin dashed line. At both channels, the amplitudes of N200 and P300 are visibly larger for target stimuli.

ture for the two classes, indicating a substantial overlap (of P300 amplitudes plotted on the horizontal axis) between the two classes of target and non-target stimuli.

Two of the six classifiers evaluated in this study, the MLP and SVM, have free parameters that need to be selected: the number of hidden layer nodes for MLP, and the kernel and margin-error tradeoff parameters for SVM. The number of hidden layer nodes for the MLP was optimized based on accuracy, sensitivity and specificity; these performance indexes were calculated using a *k*-fold cross-validation approach. Based on this preliminary information, the number of hidden layer nodes was set to 2: a higher number of nodes in fact would not improve the overall performance of the MLP, but would increase its complexity (see Fig. 4A). Similar cross-validation based optimization for SVM revealed the Gaussian kernel width σ equal to 3 and the margin-error tradeoff parameter C equal to 1 as the optimal values.

For Feature Set 1, the mean accuracy, sensitivity and specificity were 83.65%, 77.29% and 90.0%, respectively, averaged among all classifiers (Fig. 5A). Significant differences in performance were determined across classifiers in terms of accuracy ($F(5,54) = 2.89$, $p = 0.022$), sensitivity ($F(5,54) = 35.19$, $p < 0.001$) and specificity ($F(5,54) = 44.63$, $p < 0.001$): the MD and MLP classifiers offered

Table 1
Mean and standard deviation of the six features extracted from ERPs obtained at Cz and Pz.

Channel	Physiological measure	Target (μV)	Non-target (μV)
Cz	N200 amplitude	−8.15 ± 5.77	−1.56 ± 2.97
	P300 amplitude	17.02 ± 8.94	3.62 ± 3.03
	ΔAmplitude	25.17 ± 10.25	5.17 ± 3.44
Pz	N200 amplitude	−6.85 ± 5.09	−0.76 ± 2.39
	P300 amplitude	14.62 ± 7.28	4.24 ± 2.84
	ΔAmplitude	21.47 ± 7.03	5.00 ± 3.36

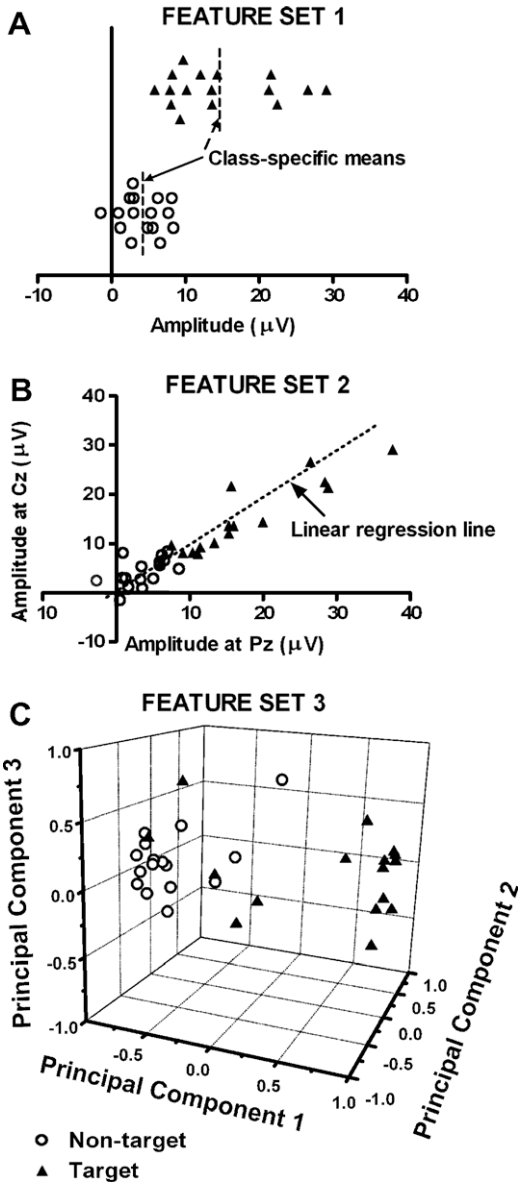

Fig. 3. Features used for classification in the three feature sets; empty circles are for non-target samples and filled triangles are for target samples. In Feature Set 1 (A), only the P300 peak amplitude recorded at Pz is used; the two classes have different means (represented by dashed lines) but the distributions of the samples partially overlap. In Feature Set 2 (B), classification is based on the P300 peak amplitude recorded at both Pz and Cz; the regression line obtained using the amplitude at Pz as predictor has been drawn in order to highlight the strong correlation between the two variables. In Feature Set 3 (C), the variables that significantly differed between targets and non-targets were transformed into principal components; only the first three were retained, since they sufficed to explain more than 98% of the data variability. In the 3D scatter plot obtained mapping the data onto the space defined by the three retained principal components the two classes still present a certain degree of overlap.

higher sensitivities but lower specificities, whereas the ED and SVM provided lower sensitivities but higher specificities.

3.4. Results for Feature Set 2

Feature Set 2 consisted of average amplitudes of the P300 peak as recorded at both Pz and Cz locations. The scatter plot of these two features in Fig. 3B indicates that there is overlap between the feature spaces of these two features. Linear regression analysis was performed, with the P300 amplitude recorded at Pz serving as the independent variable. The estimated slope was 1.17, with a 95% confidence interval of 1.01–1.33. The Pearson correlation coefficient between the P300 amplitude recorded at Pz and Cz locations was 0.93, with an R^2 statistic of 0.88, indicating that the variation in the amplitude at Pz can explain 88% of the variation in the amplitude recorded at Cz. Again, based on the model selection results, the number of hidden layer nodes for the MLP was set to 2, as in Feature Set 1 (see Fig. 4B). The optimal values for the SVM parameters were found as $\sigma = 1$ and $C = 0.01$.

Similar to Feature Set 1, the performance differed significantly across classifiers (accuracy: $F(5,54) = 20.36$, $p < 0.001$; sensitivity: $F(5,54) = 46.71$, $p < 0.001$; specificity: $F(5,54) = 23.39$, $p < 0.001$): the QC and SVM performed significantly better than the other classifiers. Similar to Feature Set 1, MD classifier significantly outperformed the other classifiers in terms of sensitivity; the EC, QC and SVM offered instead a significant improvement in terms of specificity (Fig. 5B).

3.5. Results for Feature Set 3

In the third feature set, the classification was performed on the principal components that explained 98% of the variability in the six considered features. The principal components analysis revealed that the first component alone explained 72.1% of the data variance in the data, whereas the first two components together explained 96.2% of the data variance. We retained the first 3 principal components (associated with the highest eigenvalues), which cumulatively explained 98.7% of the data variance. We should note that it is not three features that are chosen, but rather three specific linear combinations of all six features. That is, all six features contributed – to various degrees – to the principal components. Fig. 6 depicts the weights (as a percentage) of the different features in forming the first three principal components. We observe that the first component, accounting for the 72.1% of the global variance, is composed of approximately equal amounts of all six features that carry information about the P300 and N200 peaks; in the second component the contribution of the two features related to the difference in amplitude between N200 and P300 partially decreases. The third component is largely comprised of features related to the N200 amplitude, and the contribution of the P300-related features is marginal. The features mapped in the scaled principal component space revealed partial overlap between the target and non-target instances, as shown in Fig. 3C. Based on the *k*-fold validation results (Fig. 4C), the number of hidden layer nodes for the MLP was set to 5, whereas the SVM parameters, σ and C, were, respectively, set to 3 and 1.

Statistically significant differences in performance across classifiers were found (accuracy: $F(5,54) = 46.23$, $p < 0.001$; sensitivity: $F(5,54) = 17.63$, $p < 0.001$; specificity: $F(5,54) = 21.23$, $p < 0.001$); in particular, the FLD offered an overall performance significantly *lower* than that of the other classifiers (Fig. 5C).

3.6. Comparison between feature sets and type of classifiers

Classification performances were evaluated by directly comparing the overall accuracy, sensitivity and specificity obtained using the three different feature sets, independently of the chosen classifier. A one-way ANOVA with feature set as a factor was performed: the analysis revealed significant differences between the feature sets in terms of accuracy and sensitivity (accuracy: $F(2,177) = 5.46$, $p = 0.0049$; sensitivity: $F(2,177) = 12.99$, $p < 0.001$). A Tukey–Kramer test with a 95% level of significance showed that Feature

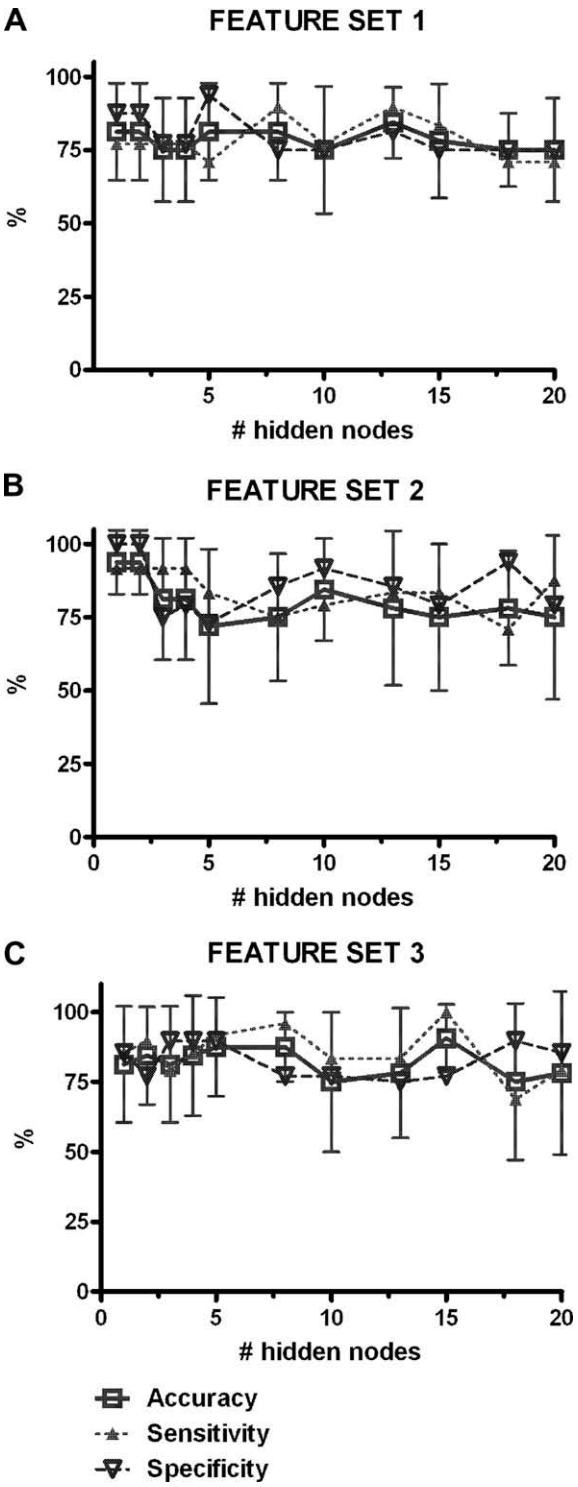

Fig. 4. Performance indexes of the MLP as a function of the number of hidden nodes for the three different feature sets. The standard deviation of the accuracy index is represented by the whiskers. In the first two feature sets, 2 hidden nodes have been chosen as the optimal number; in Feature Set 3 the selected number of hidden nodes was 5.

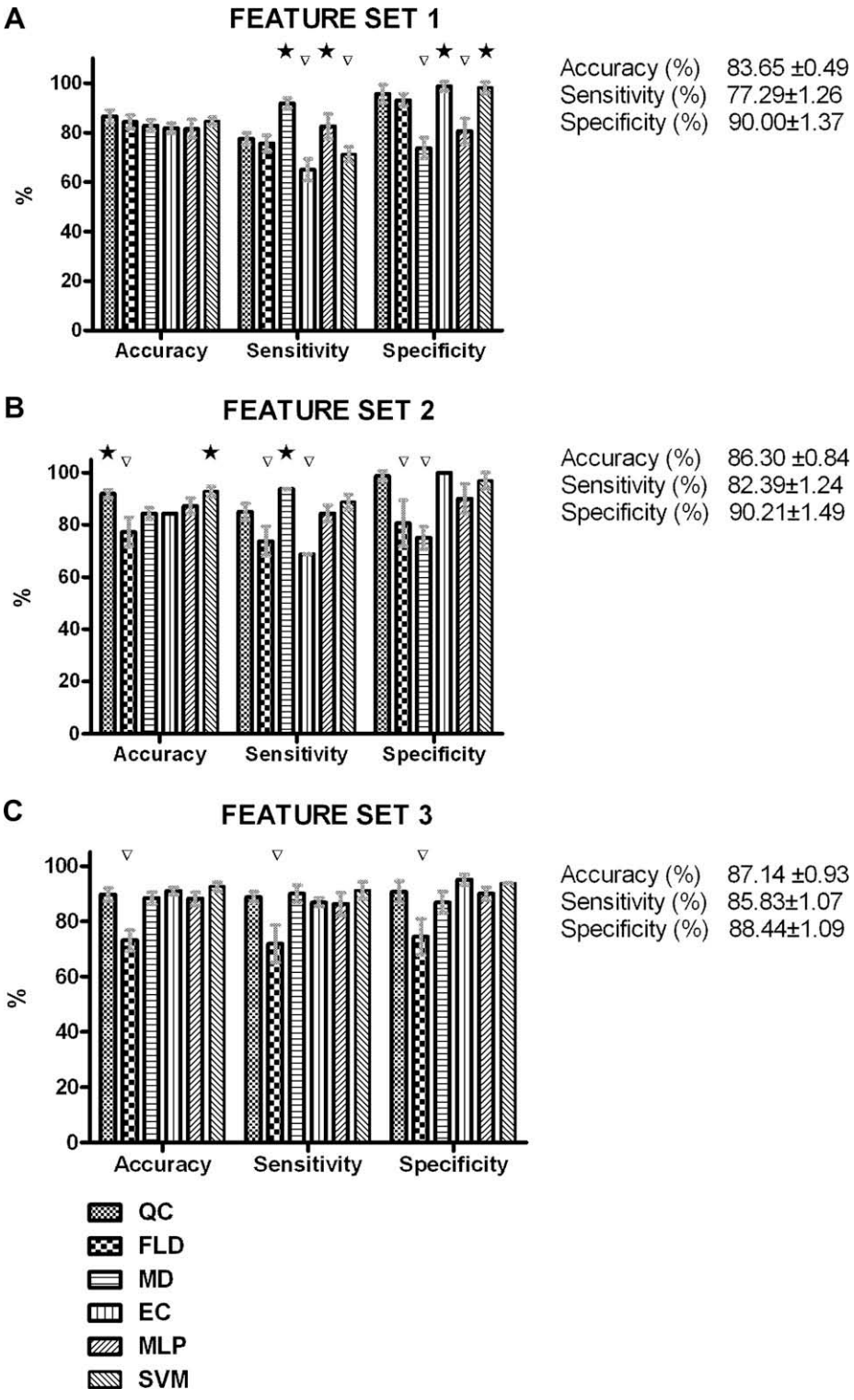

Fig. 5. Classifiers performances evaluated in terms of accuracy (percentage of correct classifications), sensitivity (percentage of correctly classified targets) and specificity (percentage of correctly classified non-targets). The bar height represents the mean and whiskers represent the 95% confidence intervals for the mean. On the right-hand side, tables with the mean and standard error of the performance indeces are reported for each feature set. Feature Set 2 (B) and Feature Set 3 (C) produced similar results in terms of accuracy and both performed generally better than Feature Set 1 (A). Solid stars and empty triangles offer a graphical representation of the statistical differences between classifiers. These differences were determined using a one-way ANOVA (followed by a Tukey–Kramer post-hoc test) separately conducted on each performance index for each feature set. The solid stars indicate a performance that is significantly higher than that of the other groups considered in the analysis, whereas the empty triangles indicate a performance that is significantly lower.

Sets 2 and 3 achieved higher accuracy and sensitivity, while the lowest values were obtained from Feature Set 1.

Furthermore, Fig. 7 reports the differences in performance connected only with the linearity degree of the classifiers, independently of the chosen feature set. The effect of the classifiers' linearity was tested with a one-way ANOVA, showing statistically significant differences (accuracy: $F(2,177) = 20.76$, $p < 0.001$; sensitivity: $F(2,177) = 4.34$, $p = 0.0145$; and specificity: $F(2,177) = 10.72$,

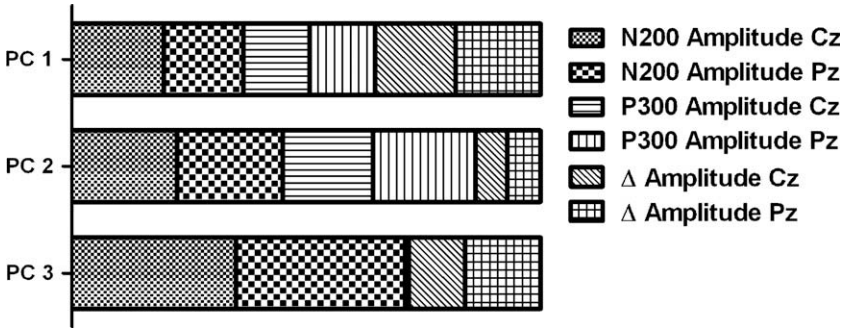

Fig. 6. Percentages of the six different features in forming the first three principal components (PCs).

$p < 0.001$). In the case of accuracy and specificity, the "non-parametric" (SVM and MLP) and "parametric non-linear" (QC) groups performed better than the "parametric linear" group (EC, MC, and FLD); in the case of sensitivity, only the "non-parametric" group outperformed the "parametric linear" group.

Finally, the interaction between the feature sets and the degree of linearity of the classifiers was evaluated using a two-way ANOVA (Fig. 7). The effect of the interaction between the feature set and the degree of linearity was found to be significant only for accuracy (accuracy: $F(4,171) = 4.35$, $p = 0.0022$; sensitivity: $F(4,171) = 1.61$, $p = 0.173$; specificity: $F(4,171) = 1.60$, $p = 0.177$). The classification performed by the "parametric non-linear" and "non-parametric" groups on Feature Sets 2 and 3 yielded statistically higher accuracy.

4. Discussion

In this study, we investigated whether automated classifiers can be used to correctly identify target categorization responses from average event-related potentials (ERPs). To do so, we compared the performances of six classifiers on three different sets of ERP features with regard to their ability to accurately characterize and discriminate such ERP responses as target and non-target stimuli.

In general, the results obtained in this study support the use of the parametric non-linear (QC) and non-parametric classifiers (MLP, SVM) instead of linear classifiers (EC, MD, FLP). Furthermore, the additional use of features describing the N200 peak for the classification did not increase the performance obtained from the use of only features describing the P300 peak; this result suggests that, for this application, the N200 did not explain variance beyond that of P300 recorded at multiple sites. Overall, the best classification performance was provided by the parametric non-linear classifiers (accuracy: 92%; sensitivity: 83%; specificity: 99%) and by the non-parametric classifiers (accuracy: 91%; sensitivity: 86%; specificity: 95%) on feature vectors extracted from P300 recorded at multiple sites, specifically the Pz and Cz locations.

More in detail, the six features that were evaluated included the peak amplitudes of the N200 and P300 at Pz and Cz, as well as the difference in amplitude between N200 and P300 at Pz and Cz. All features were found to differentiate between responses to target and non-target stimuli. This result confirms our expectations because N200 and P300 are known to be prominent in the responses to the target stimuli in target categorization tasks (Botvinick et al., 2004; Polich and Kok, 1995). In particular, the visual discrimination task used in this study to elicit the ERPs required the participants to respond to both target and non-target stimuli. This specific design presumably enhanced the inhibition that participants were required to exert during target trials in order to overrule the habituated response to the non-target stimuli. Therefore, since the N200 component is linked to the monitoring of conflicts, it is reasonable to expect a prominent N200 component in the responses to target stimuli.

Six different classifiers were considered (QC, FLD, MD, EC, MLP, SVM) and their performances were evaluated on three different sets of features: P300 amplitude recorded at Pz (Feature Set 1); P300 amplitude recorded at Pz and Cz (Feature Set 2); and the first three principal components of the six features (Feature Set 3). In Feature Set 3, the original six variables (P300 and N200 amplitudes and the difference between these two, calculated at Cz and Pz) were reduced to three principal components, used then as classification features. The contribution of the original 6 variables to each of these final 3 principal components revealed that these original variables were not independent of each other (Fig. 6). PCA was therefore chosen as a suitable tool for the analysis and manipulation of such inter-dependent data, given its extensive applications to EEG in the literature (Brown et al., 1979; Casarotto et al., 2004; Lange and Inbar, 1996; Liberati et al., 1992). Moreover, given the small number of available samples, data reduction was a necessary step, since the number of samples needed for proper training and discrimination increases exponentially with the number of considered features, a problem known as the "curse of dimensionality" (Friedman, 1997; Jain et al., 2000).

Several trends were observed across the first two feature sets: the EC showed high specificity but low sensitivity, the MD and SVM classifiers achieved high sensitivity but suffered from relatively low specificity. These observations indicate that some classifiers were biased towards one of the two classes: they tended to classify either targets better than non-targets (sensitivity > specificity) or non-targets better than targets (specificity > sensitivity). On the other hand, the MLP showed a balance between sensitivity and specificity (i.e., a balance in its ability to recognize the two classes). This observation is in line with the behavior of the network for the chosen number of hidden nodes (Figs. 4 and 5).

Feature Set 1 offered an overall performance that, though acceptable, was lower than that obtained using the other two feature sets. This result illustrates why it is standard practice in brain–computer interface (BCI) applications to perform classification using a variety of features collected at multiple recording sites (Babiloni et al., 2001; Krusienski et al., 2008; Serby et al., 2005), rather than a single site.

No significant differences were observed between the performances of Feature Sets 2 and 3, though Feature Set 2 had higher specificity and Feature Set 3 had higher sensitivity. In addition to the information about the amplitude of the P300 component, Feature Set 3 also included knowledge about the N200 component. The overall equivalence, in terms of classification ability, between Feature Sets 2 and 3 suggests that the additional information does not contribute substantially to the discrimination between responses to the two classes of stimuli. Furthermore, many oddball

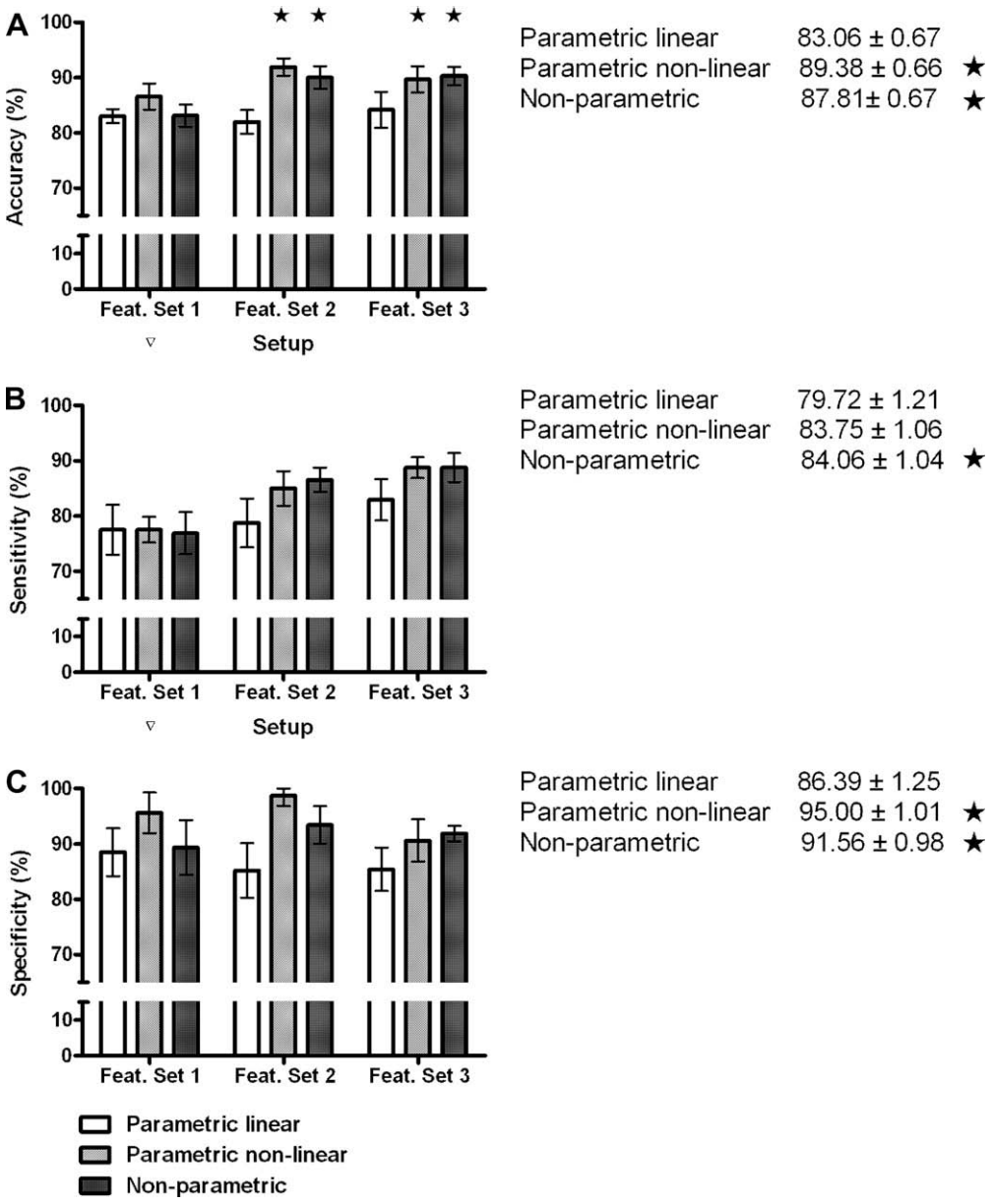

Fig. 7. Comparison of performances divided by feature sets and by linearity degree of the classifiers. The bar height represents the mean and whiskers represent the 95% confidence intervals for the mean. On the right-hand side, tables with the mean and standard error of the performance indeces are provided for each group of classifiers. Solid stars and empty triangles offer a graphical representation of statistical differences. The solid stars indicate a performance that is significantly higher than that of the other groups considered in the analysis, whereas the empty triangles indicate a performance that is significantly lower. The three feature sets were statistically compared within each performance index and the existing significant differences are graphically presented at the bottom of each graph. Similarly, the three groups of classifiers were statistically compared within each performance index and the existing significant differences are graphically presented at right of the data tables. Lastly, the interaction between the feature set and the linearity degree of the classifiers was investigated within each performance index and the existing significant differences are graphically presented on top of the bar graphs. Overall, the classifiers in the "parametric non-linear" and "non-parametric" groups performed generally better than the "parametric linear" group. Feature Set 1 was not able to achieve accuracy and sensitivity comparable to those of Feature Sets 2 and 3.

paradigms require the participants to respond only to target stimuli: the N200 component in the ERPs is thus minimized. Therefore, the equivalence between Feature Sets 2 and 3 seems to indicate that the conflict resolution processes raised by the specific task design and reflected in the N200 component did not affect the discriminability of the ERPs.

Furthermore, different classification algorithms providing different performance characteristics on different feature sets confirm the notion of the no-free-lunch theory (Wolpert and Macready, 1997) and justify the need to evaluate a series of classification algorithms with a broad spectrum of learning characteristics. Statistical analyses also revealed significant differences in performance with respect to the degree of linearity of the classifiers. The "parametric non-linear" and "non-parametric" groups outperformed the "parametric linear" group, regardless of the chosen feature set. This result is in line with the observation that the two classes partially overlap in the feature spaces (Fig. 3), thus it is a case of linearly non-separable classes. Krusienski and colleagues (Krusienski et al., 2006), however, obtained different results when comparing classifiers for their ability to detect responses to target stimuli in a BCI application. Based on their experiments, they suggested that no advantage was gained from the use of non-linear classifiers. The discrepancy between their results and the results reported here may be due to differences in the experiment design:

the participants in their study were given online feedback and were tested in multiple sessions. On the contrary, our results are based on responses collected from a population of naïve subjects who were not provided any feedback or training, which represent instead a key component in most BCI setups (Elbert et al., 1980; Guger et al., 2003).

In conclusion, this study suggests that automated characterization and classification of average ERP responses to infrequent targets is feasible with good accuracy. In particular, the use of features characterizing the P300 amplitude at multiple recording sites is recommended. Furthermore, the performed analyses seem to discourage the use of parametric linear classifiers for the specific application.

One aspect worth consideration is that the conclusions drawn from this work are based on the analysis of average ERPs. Specifically, the averaging approach is common practice in ERP research because it ensures some improvement in the signal to noise ratio (SNR). In the case of lower SNR conditions, i.e., single-trial recordings that are susceptible to the influence of background EEG, the features extracted from the ERPs and used for classification will show a higher variability. This increased dispersion of the two classes in the feature space is likely to degrade the performance of the classification algorithms overall. Parametric linear classifiers would likely show a larger decrease in performance, given the higher degree of overlap between the two classes. Classification algorithms with decision boundaries other than hyperplanes, and hence more flexible, would probably be less affected and show better generalization of the results. Furthermore, classification based on features from multiple recording sites would probably still be recommended. It is reasonable to assume that the relative information provided by multiple recording sites would provide some level of redundancy that would prove useful in classification.

5. Future work

A direct extension of the study presented in this paper will examine how these findings extend to a variety of clinical populations that have attentional impairments. The aim would be to create a tool for the assessment of such impairments and the evaluation of possible treatments.

Future work will also include the investigation of automatic recognition of single-trial target ERPs. Such a study could pave the way for real-time monitoring of the attentional level and its fluctuations under different conditions. These variations in attention may be due to everyday factors such as fatigue (Boonstra et al., 2007; Davidson et al., 2007; Pilcher et al., 2007), or they may result from more insidious processes such as mental or neurological disorder (Castellanos et al., 2005; Liu et al., 2002; Swaab-Barneveld et al., 2000). The vast majority of P300-based investigations of attention and the effects of time-on-task have employed the typical ERP methodology of averaging over multiple trials to increase the signal to noise ratio (Maatta et al., 2005; Slater et al., 1994). However, recent research has suggested that monitoring the level of attention and its fluctuations throughout a task may provide additional insight to disease processes (Holm et al., 2006; Roschke et al., 1996; Tomberg and Desmedt, 1999). Several algorithms for the single-trial classification of oddball responses have been developed in the BCI literature (Birbaumer, 2006; Donchin et al., 2000; Farwell and Donchin, 1988; Krusienski et al., 2006; Piccione et al., 2006; Sellers and Donchin, 2006; Sellers et al., 2006; Serby et al., 2005; Thulasidas et al., 2006; Wang et al., 2005; Wolpaw et al., 2002). Most of the methods used in BCIs, however, require several trials before their parameters can be optimally adjusted to a given individual. In addition, subjects usually undertake training sessions in order to maximize their efficiency in communicating using a specific BCI setup. These requirements are therefore suboptimal if the goal is to examine the modulation of the oddball effect as a consequence of neurological disorders or when training and feedback would interfere with the analysis of interest. Therefore, given the growing literature suggesting that real-time measures of the neural correlates of attention yield important information not found in averaged data, improved methods with which to assess single-trial responses are worth further investigation and this study can be considered a first step in this direction.

Acknowledgments

The authors would like to thank Dr. Patricia Shewokis for her kind support and valuable advice.

This work was in part supported by the funds from the Defense Advanced Research Projects Agency (DARPA) Augmented Cognition Program, the Office of Naval Research (ONR) and Homeland Security, under Agreements N00014-02-1-0524, N00014-01-10986, and N00014-04-1-0119. This work was also partially supported by the HINT@Lecco project.

References

Babiloni F, Bianchi L, Semeraro F, del R Millan J, Mourino J, Cattini A, et al. Mahalanobis distance-based classifiers are able to recognize EEG patterns by using few EEG electrodes. In: 23th Annual International Conference of IEEE EMBS; 2001. p. 651-4.

Basar-Eroglu C, Basar E, Demiralp T, Schurmann M. P300-response: possible psychophysiological correlates in delta and theta frequency channels. A review. Int J Psychophysiol 1992;13:161–79.

Basar-Eroglu C, Demiralp T, Schurmann M, Basar E. Topological distribution of oddball 'P300' responses. Int J Psychophysiol 2001;39:213–20.

Birbaumer N. Breaking the silence: brain–computer interfaces (BCI) for communication and motor control. Psychophysiology 2006;43:517–32.

Boonstra TW, Stins JF, Daffertshofer A, Beek PJ. Effects of sleep deprivation on neural functioning: an integrative review. Cell Mol Life Sci 2007;64:934–46.

Botvinick MM, Cohen JD, Carter CS. Conflict monitoring and anterior cingulate cortex: an update. Trends Cogn Sci 2004;8:539–46.

Bramon E, Rabe-Hesketh S, Sham P, Murray RM, Frangou S. Meta-analysis of the P300 and P50 waveforms in schizophrenia. Schizophr Res 2004;70:315–29.

Brown WS, Marsh JT, Smith JC. Principal component analysis of ERP differences related to the meaning of an ambiguous word. Electroencephalogr Clin Neurophysiol 1979;46:709–14.

Casarotto S, Bianchi AM, Cerutti S, Chiarenza GA. Principal component analysis for reduction of ocular artefacts in event-related potentials of normal and dyslexic children. Clin Neurophysiol 2004;115:609–19.

Castellanos FX, Sonuga-Barke EJ, Scheres A, Di Martino A, Hyde C, Walters JR. Varieties of attention-deficit/hyperactivity disorder-related intra-individual variability. Biol Psychiatry 2005;57:1416–23.

Coburn KL, Shillcutt SD, Tucker KA, Estes KM, Brin FB, Merai P, et al. P300 delay and attenuation in schizophrenia: reversal by neuroleptic medication. Biol Psychiatry 1998;44:466–74.

Coburn KL, Lauterbach EC, Boutros NN, Black KJ, Arciniegas DB, Coffey CE. The value of quantitative electroencephalography in clinical psychiatry: a report by the Committee on Research of the American Neuropsychiatric Association. J Neuropsychiatry Clin Neurosci 2006;18:460–500.

Davidson PR, Jones RD, Peiris MT. EEG-based lapse detection with high temporal resolution. IEEE Trans Biomed Eng 2007;54:832–9.

Donchin E, Spencer KM, Wijesinghe R. The mental prosthesis: assessing the speed of a P300-based brain–computer interface. IEEE Trans Rehabil Eng 2000;8:174–9.

Donkers FC, van Boxtel GJ. The N2 in go/no-go tasks reflects conflict monitoring not response inhibition. Brain Cogn 2004;56:165–76.

Duda RO, Hart PE, Stork DG. Pattern classification. New York, NY: Wiley Interscience; 2001.

Egeth HE, Yantis S. Visual attention: control, representation, and time course. Annu Rev Psychol 1997;48:269–97.

Elbert T, Rockstroh B, Lutzenberger W, Birbaumer N. Biofeedback of slow cortical potentials. I. Electroencephalogr Clin Neurophysiol 1980;48:293–301.

Farwell LA, Donchin E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalogr Clin Neurophysiol 1988;70:510–23.

Ford JM, Pfefferbaum A, Roth W. P3 and schizophrenia. Ann NY Acad Sci 1992;658:146–62.

Friedman JH. Loss, and the curse-of-dimensionality. Data Min Knowledge Dis 1997:55–77.

Guger C, Edlinger G, Harkam W, Niedermayer I, Pfurtscheller G. How many people are able to operate an EEG-based brain–computer interface (BCI)? IEEE Trans Neural Syst Rehabil Eng 2003;11:145–7.

Haykin S. Neural networks: a comprehensive foundation. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 1999.
Holm A, Ranta-aho PO, Sallinen M, Karjalainen PA, Muller K. Relationship of P300 single-trial responses with reaction time and preceding stimulus sequence. Int J Psychophysiol 2006;61:244–52.
Jain AK, Duin RPW, Jianchang M. Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 2000;22:4–37.
Jung T-P, Makeig S, Westerfield M, Townsend J, Courchesne E, Sejnowski TJ. Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. Clin Neurophysiol 2000a;111:1745–58.
Jung TP, Makeig S, Humphries C, Lee TW, McKeown MJ, Iragui V, et al. Removing electroencephalographic artifacts by blind source separation. Psychophysiology 2000b;37:163–78.
Keren O, Ben-Dror S, Stern MJ, Goldberg G, Groswasser Z. Event-related potentials as an index of cognitive function during recovery from severe closed head injury. J Head Trauma Rehabil 1998;13:15–30.
Kok A. On the utility of P3 amplitude as a measure of processing capacity. Psychophysiology 2001;38:557–77.
Krusienski DJ, Sellers EW, Cabestaing F, Bayoudh S, McFarland DJ, Vaughan TM, et al. A comparison of classification techniques for the P300 speller. J Neural Eng 2006;3:299–305.
Krusienski DJ, Sellers EW, McFarland DJ, Vaughan TM, Wolpaw JR. Toward enhanced P300 speller performance. J Neurosci Methods 2008;167:15–21.
Kuncheva LI. Combining pattern classifiers: methods and algorithms. New York, NY: Wiley Interscience; 2004.
Lange DH, Inbar GF. Variable single-trial evoked potential estimation via principal component identification. In: 18th annual international conference of the IEEE engineering in medicine and biology society, vol. 3; 1996. p. 954–5.
Lew HL, Poole JH, Chiang JY, Lee EH, Date ES, Warden D. Event-related potential in facial affect recognition: potential clinical utility in patients with traumatic brain injury. J Rehabil Res Dev 2005;42:29–34.
Liberati D, Brandazza P, Casagrande L, Cerini A, Kaufman B. Detection of transient single-sweep somatosensory evoked potential changes via principal components analysis of the autoregressive-with-exogenous-input parameters. In: Annual international conference of the IEEE engineering in medicine and biology society, vol. 6; 1992. p. 2454–5.
Liu SK, Chiu CH, Chang CJ, Hwang TJ, Hwu HG, Chen WJ. Deficits in sustained attention in schizophrenia and affective disorders: stable versus state-dependent markers. Am J Psychiatry 2002;159:975–82.
Maatta S, Herrgard E, Saavalainen P, Paakkonen A, Kononen M, Luoma L, et al. P3 amplitude and time-on-task effects in distractible adolescents. Clin Neurophysiol 2005;116:2175–83.
Missonnier P, Ragot R, Derouesne C, Guez D, Renault B. Automatic attentional shifts induced by a noradrenergic drug in Alzheimer's disease: evidence from evoked potentials. Int J Psychophysiol 1999;33:243–51.
Naatanen R, Picton TW. N2 and automatic versus controlled processes. Electroencephalogr Clin Neurophysiol Suppl 1986;38:169–86.
Nieuwenhuis S, Yeung N, van den Wildenberg W, Ridderinkhof KR. Electrophysiological correlates of anterior cingulate function in a go/no-go task: effects of response conflict and trial type frequency. Cogn Affect Behav Neurosci 2003;3:17–26.
Pessoa L, Kastner S, Ungerleider LG. Neuroimaging studies of attention: from modulation of sensory processing to top-down control. J Neurosci 2003;23:3990–8.
Piccione F, Giorgi F, Tonin P, Priftis K, Giove S, Silvoni S, et al. P300-based brain computer interface: reliability and performance in healthy and paralysed participants. Clin Neurophysiol 2006;117:531–7.
Pilcher JJ, Band D, Odle-Dusseau HN, Muth ER. Human performance under sustained operations and acute sleep deprivation conditions: toward a model of controlled attention. Aviat Space Environ Med 2007;78:B15–24.
Polich J. On the relationship between EEG and P300: individual differences, aging, and ultradian rhythms. Int J Psychophysiol 1997;26:299–317.
Polich J. Theoretical overview of P3a and P3b. In: Polich J, editor. Detection of change: event-related potential and fMRI findings. Boston, MA: Kluwer Academic; 2003. p. 83–98.
Polich J, Kok A. Cognitive and biological determinants of P300: an integrative review. Biol Psychol 1995;41:103–46.
Polich J, Howard L, Starr A. Effects of age on the P300 component of the event-related potential from auditory stimuli: peak definition, variation, and measurement. J Gerontol 1985;40:721–6.
Polikar R, Topalis A, Parikh D, Green D, Frymiare J, Kounios J, et al. An ensemble based data fusion approach for early diagnosis of Alzheimer's disease. Inf Fusion 2008;9:83–95.
Ravden D, Polich J. On P300 measurement stability: habituation, intra-trial block variation, and ultradian rhythms. Biol Psychol 1999;51:59–76.
Roschke J, Wagner P, Mann K, Fell J, Grozinger M, Frank C. Single trial analysis of event related potentials: a comparison between schizophrenics and depressives. Biol Psychiatry 1996;40:844–52.
Sangal RB, Sangal JM. Attention-deficit/hyperactivity disorder: cognitive evoked potential (P300) topography predicts treatment response to methylphenidate. Clin Neurophysiol 2004;115:188–93.
Sangal RB, Sangal JM. Attention-deficit/hyperactivity disorder: use of cognitive evoked potential (P300) to predict treatment response. Clin Neurophysiol 2006;117:1996–2006.
Sellers EW, Donchin E. A P300-based brain–computer interface: initial tests by ALS patients. Clin Neurophysiol 2006;117:538–48.
Sellers EW, Krusienski DJ, McFarland DJ, Vaughan TM, Wolpaw JR. A P300 event-related potential brain–computer interface (BCI): the effects of matrix size and inter stimulus interval on performance. Biol Psychol 2006;73:242–52.
Serby H, Yom-Tov E, Inbar GF. An improved P300-based brain–computer interface. IEEE Trans Neural Syst Rehabil Eng 2005;13:89–98.
Slater JD, Wu FY, Honig LS, Ramsay RE, Morgan R. Neural network analysis of the P300 event-related potential in multiple sclerosis. Electroencephalogr Clin Neurophysiol 1994;90:114–22.
Struber D, Polich J. P300 and slow wave from oddball and single-stimulus visual tasks: inter-stimulus interval effects. Int J Psychophysiol 2002;45:187–96.
Sumi N, Nan'no H, Fujimoto O, Ohta Y, Takeda M. Interpeak latency of auditory event-related potentials (P300) in senile depression and dementia of the Alzheimer type. Psychiatry Clin Neurosci 2000;54:679–84.
Sutton S, Braren M, Zubin J, John ER. Evoked-potential correlates of stimulus uncertainty. Science 1965;150:1187–8.
Swaab-Barneveld H, de Sonneville L, Cohen-Kettenis P, Gielen A, Buitelaar J, Van Engeland H. Visual sustained attention in a child psychiatric population. J Am Acad Child Adolesc Psychiatry 2000;39:651–9.
Thulasidas M, Guan C, Wu J. Robust classification of EEG signal for brain–computer interface. IEEE Trans Neural Syst Rehabil Eng 2006;14:24–9.
Tomberg C, Desmedt JE. The challenge of non-invasive cognitive physiology of the human brain: how to negotiate the irrelevant background noise without spoiling the recorded data through electronic averaging. Philos Trans R Soc Lond B Biol Sci 1999;354:1295–305.
Wang C, Guan C, Zhang H. P300 brain–computer interface design for communication and control applications. In: 27th annual international conference of IEEE EMBS, vol. 5. Shangai, China; 2005. p. 5400–3.
Werbos PJ. Beyond regression: New tools for prediction and analysis in the behavioral sciences. Ph.D. Thesis. Cambridge, MA: Harvard University; 1974.
Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain–computer interfaces for communication and control. Clin Neurophysiol 2002;113:767–91.
Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Trans Evol Comput 1997;1:67–82.