ORIGINAL ARTICLE

Efficient automatic classifiers for the detection of A phases of the cyclic alternating pattern in sleep

Sara Mariani · Elena Manfredini · Valentina Rosso · Andrea Grassi · Martin O. Mendez · Alfonso Alba · Matteo Matteucci · Liborio Parrino · Mario G. Terzano · Sergio Cerutti · Anna M. Bianchi

Received: 20 April 2011 / Accepted: 24 February 2012 / Published online: 20 March 2012

Abstract This study aims to develop an automatic detector of the A phases of the cyclic alternating pattern, periodic activity that generally occurs during non-REM (NREM) sleep. Eight polysomnographic recordings from healthy subjects were examined. From EEG recordings, five band descriptors, an activity descriptor and a variance descriptor were extracted and used to train different machine-learning algorithms. A visual scoring provided by an expert clinician was used as golden standard. Four alternative mathematical machine-learning techniques were implemented: (1) discriminant classifier, (2) support vector machines, (3) adaptive boosting, and (4) supervised artificial neural network. The results of the classification, compared with the visual analysis, showed average accuracies equal to 84.9 and 81.5% for the linear discriminant and the neural network, respectively, while AdaBoost had a slightly lower accuracy, equal to 79.4%. The SVM leads to accuracy of 81.9%. The performance achieved by the automatic classification is encouraging, since an efficient automatic classifier would benefit the practice in everyday clinics, preventing the physician from the time-consuming activity of the visually scoring of the sleep microstructure over whole 8-h sleep recordings. Finally, the classification based on learning algorithms would provide an objective criterion, overcoming the problems of inter-scorer disagreement.

Keywords Sleep · Cyclic alternating pattern · Neural networks · Support vector machines · Machine learning

1 Introduction

The cyclic alternating pattern (CAP) is a periodic activity that occurs on the electroencephalographic (EEG) signal during sleep and is characterized by an activation phase, called phase A, which is very different from the background, and a second phase, called B, in which only the background is visible. Both A and B phases may have a duration between 2 and 60 s. Phases A are classified into three subtypes: A1 is characterized by strong delta waves (0.5–4 Hz); A2 has rapid EEG activities that occur for 20–50% of the total activation time and A3 is characterized by rapid activities, especially beta (16–30 Hz), that occupy more than the 50% of the total activation time. Phase A and the following phase B shape a CAP cycle and at least two consecutive CAP cycles are needed to define a CAP sequence. A typical example of the EEG signal during CAP and non-CAP sleep is shown in Fig. 1 [26].

In normal physiological conditions, CAP only occurs during NREM sleep, although it can appear in REM sleep in pathological conditions. Being connected with sleep instability, CAP sleep contains information that is relevant

Electronic supplementary material The online version of this article (doi:10.1007/s11517-012-0881-0) contains supplementary material, which is available to authorized users.

S. Mariani (✉) · E. Manfredini · S. Cerutti · A. M. Bianchi
Department of Biomedical Engineering, Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milan, Italy
e-mail: sara1.mariani@mail.polimi.it

V. Rosso · A. Grassi · L. Parrino · M. G. Terzano
Department of Neurology, Sleep Disorders Center, University of Parma, via Gramsci 14, 43126 Parma, Italy

M. O. Mendez · A. Alba
Department of Electronics, Universidad Autonoma de San Luis Potosi, Salvador Nava S/N, San Luis Potosì, Mexico

M. Matteucci
Department of Information Engineering, Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milan, Italy

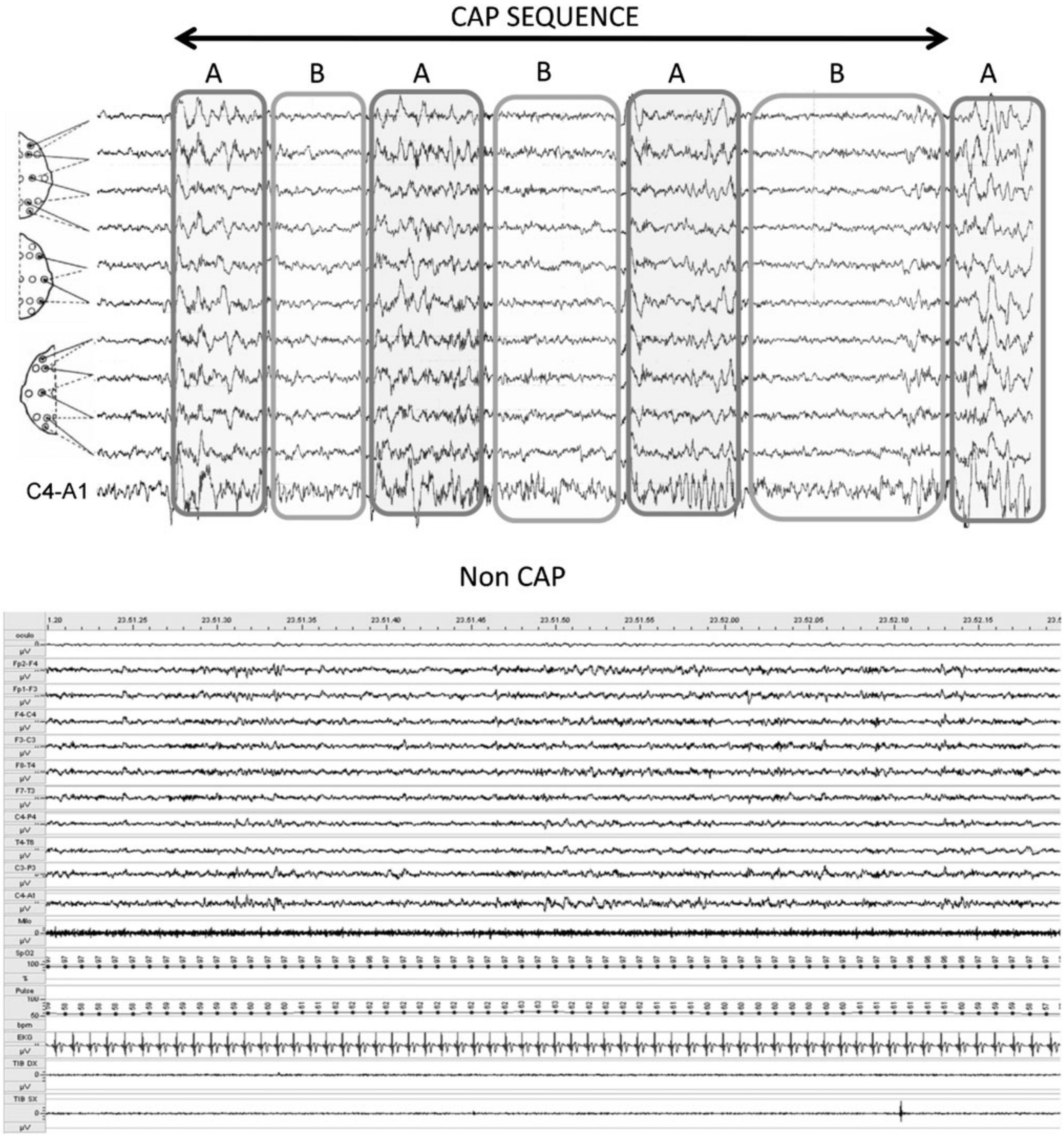


Fig. 1 An example of non-CAP and CAP sleep in sleep stage 2. Phase A is characterized by amplitude/frequency content that stands out against the background, while Phase B shows a return to the background itself. Both Phase A and B may last between 2 and 60 s. We have a CAP sequence when there are at least two consecutive CAP cycles, terminated by a phase A. The remaining NREM sleep is classified as non-CAP

in clinics for the evaluation of the quality of a subject's sleep. Increased amounts of CAP are a regular finding in obstructive sleep apnea syndrome (OSAS) [27] and in the upper airway resistance syndrome (UARS) as a reaction of the sleeping brain to a repetitive breathing disturbance. Primary insomnia shows increased amounts of CAP, compared with sound sleepers [25]. Furthermore, CAP A phase has been interpreted in several studies as a kind of gate through which pathologic events occur more easily. The gating effect has been demonstrated among several sleep disturbances such as periodic leg movements (PLM) [6, 11, 20], sleep bruxism [15], and epilepsy [5, 12, 28].

In the light of this, the ratio between NREM CAP sleep and total NREM sleep (*CAP rate*), and the different distribution of CAP A phases through the sleep stages can be measured in sleep centers to characterize sleep pathologies. Nowadays, the neurologists in sleep centers analyze the EEG of whole night sleep recordings and visually score each activation to compute the parameters used in diagnosis [3]. This method has two drawbacks: first of all, it is extremely time-consuming for the neurologist, limiting the use and study of sleep microstructure in routine clinics; secondly, it is subject to a certain inter-scorer variability in the classification: a fairly recent study estimated that the average repeatability between the classifications of single EEG traces by two different clinicians ranges from 69 to 77% [24].

This work is aimed at creating an automatic method to detect activations that may constitute CAP A phases, where the goal is not only to accelerate and optimize the physician's time, but also to provide a more precise and objective detection based on the EEG spectral parameters.

Only a few studies do exist in literature where similar algorithms were developed:

- Largo et al. [16] implemented an automatic detection algorithm using wavelets to filter the EEG signal in several bands, two moving average windows to compute the band descriptors, and a two-threshold criterion to detect the beginning and the end of each A phase. The two moving average windows and the thresholds were established through a Genetic Algorithm. They applied a minimum overlapping criterion of 0.25 s for counting a true positive (TP), i.e., an automatically scored activation that has also been visually scored, and thus obtained 84.9% of sensitivity auto-visual (ratio between the number of true positives and the total number of visually recognized activations), 77.6% of sensitivity visual-auto (ratio between the number of true positives and the total number of automatically recognized activations) and 81.1% of concordance (a combination of the previous indexes).
- Barcaro et al. [1] used band descriptors computed on the F4-C4 trace and two-threshold criteria (a detection threshold and a length threshold) in order to recognize activations from the background. They achieved a maximum concordance equal to 83.5%, although the criterion for counting a TP and the other statistics are not reported.
- Ferri et al. [7] implemented a human-supervised automatic approach for the detection of CAP A phases, letting the clinician choose the threshold values for each subject by examining part of the sleep recording (C3-A2 or C4-A1 derivation), and then using threshold criteria applied to a delta and a beta band descriptor. The statistic reported in the article are computed on the average number of A phases automatically and visually detected over 10 subjects' recordings, and show values of Kendall's *W* coefficient greater than 0.8.
- Navona et al. [19] developed an automatic method for A phases detection based on the computation of five descriptors, each of them corresponding to a different EEG frequency band. The computation of these descriptors, followed by the superimposition of two thresholds and the application of logical criteria, provided a 77% correctness, a 90% sensitivity auto-visual and a 84% sensitivity visual-auto. The analysis was carried out on some selected segments of the F4-C4 EEG trace.

Although all these methods achieved remarkable results, most of them require additional intervention from the clinician, either by tuning parameters and thresholds for each subject, or by a prior selection of trace segments at specific sleep stages; therefore, further effort has to be put in the research of an automatic method that could be reliable enough to be effectively introduced in everyday clinics.

The present study wishes to implement such an automatic system able to detect the CAP dynamic based on features extracted from a single EEG lead, lasting approximately 8 h, independent of any a priori information provided by the clinicians or any other type of human intervention. For the automatic classification, four different classifiers were compared: support vector machines (SVM), linear discriminant (LD), neural network (NN) and AdaBoost.

2 Methods

2.1 Clinical protocol

The recordings employed in this study belong to the all-night polysomnographic database of the Parma Sleep Disorders Center. Eight healthy subjects, four males and four females, aged between 29 and 42 years (mean 34.25 ± 4.86), were selected after the accomplishment of an entrance investigation in order to obtain a homogeneous group free from psychiatric, neurological and medical disorders. All subjects gave their informed consent to the study. Sleep/wake schedule was investigated for 14 days before the recordings with a sleep log. Inclusive criteria were the absence of sleep disorders and daytime napping. A personal interview integrated by a structured questionnaire confirmed good vigilance level during daytime, normal sleep habits without any difficulties in falling or remaining asleep at night. All participants were requested to avoid any drug intake and excessive alcohol or coffee consumption in the previous 3 weeks. All subjects slept at least two consecutive nights in a video-monitored, temperature-controlled and soundproof ($Leq < 35$ dB) laboratory. The first night was used for adaptation to the recording environment and for screening respiratory or other sleep-related disorders. Exclusion criteria were: apnea–hypopnea index ≥ 5 and/or PLM index ≥ 15 [20, 27]. Only the PSG recordings following the adaptation night were analyzed.

Each signal was recorded using the commercial PSG-system Siesta (Compumedics$^{\text{TM}}$); sleep scoring was manually performed on the commercial software Somnologica Studio (Embla Systems$^{\text{TM}}$).

Sleep stages were scored by an expert on a monopolar derivation (C3-A2 or C4-A1) using standard criteria [14] CAP detection [26] was based on a bipolar montage from the left or right hemisphere according to the 10–20 international system (Fp1-F3; F3-C3; C3-P3; P3-O1 or Fp2-F4; F4-C4; C4-P4; P4-O2).

A central trace (C3-A2 or C4-A1, containing equivalent information due to their symmetry) was used for the automatic analysis. This choice is due to the requirements of conventional sleep scoring [23] that state that monopolar derivations are preferred when performing a single-lead analysis. In the recordings used in the present study, EEG signals were originally digitized with frequencies between 128 and 256 Hz. The acquired EEG signals were exported from Somnologica to the programming environment Matlab (The Mathworks Inc.) at a sampling frequency equal to 100 Hz. Portions of the signal relative to wake and REM sleep were removed from the analysis. The A phases scoring provided by the clinicians was used as the golden standard for the automatic classification.

Sleep parameters about the subjects used in the study—according to the visual scoring performed by the sleep experts—such as time in bed, total sleep time, sleep efficiency, wake, NREM and REM time, number of phases A of each subtype, number of phases B and CAP rate, are reported in Table 1.

2.2 Feature set

Several features were extracted from the NREM sleep EEG:

- *Band descriptors* The EEG signal was filtered with a low-pass anti-aliasing filter at 30 Hz. Then, it was separated in the following bands: $0.5 < \text{delta} \leq 4$ Hz, $4 < \text{theta} \leq 8$ Hz, $8 < \text{alpha} \leq 12$ Hz, $12 < \text{sigma} \leq 15$ Hz, $15 < \text{beta} \leq 30$ Hz. A FIR filter with 30 coefficients and a Kaiser window was used for this purpose.

For each band, the resulting signal was squared and normalized between 0 and 1 (with respect to the maximum power in the band), and, for each of the five bands under study, a set of descriptors was implemented in the form:

$$d_{\mathrm{b}}(t) = \frac{p_{\mathrm{bs}}(t) - p_{\mathrm{bl}}(t)}{p_{\mathrm{bl}}(t)} \tag{1}$$

where p_{bl} and p_{bl} are the mean power in the considered band on a window of 2 and 64 s, respectively, centered on the second t. For example, the delta descriptor was computed as follows: the EEG signal $x(t)$was filtered between 0.5 and 4 Hz $x_{\mathrm{delta}}(t)$, then squared and normalized $x_{\mathrm{deltan}}(t)$, and for each second t its average value on a 64 s window p_{deltal} and its average value on a 2 s window p_{deltas} centered on the 1 s time window t were computed. Thus, the delta descriptor, resulting in a new signal sampled at 1 Hz, is calculated as:

$$d_{\mathrm{delta}}(t) = \frac{p_{\mathrm{deltas}}(t) - p_{\mathrm{deltal}}(t)}{p_{\mathrm{deltal}}(t)} \tag{2}$$

The windows were chosen according to previous studies [1, 16].

- *Hjorth activity* [13] It was applied to the EEG signal filtered in the delta band. It was computed over overlapped 3-s windows, each centered on the second of interest. This descriptor captures the overall increase of the delta power occurring during the activations over a longer time span. It is calculated as the simple variance σ^2 of the signal segment.

Table 1 Sleep data of the eight recorded subjects: the table reports time in bed, total sleep time (TST), sleep efficiency (SE), wake, NREM and REM time, number of A1, A2 and A3 phases, both belonging to CAP sequences and isolated, number of B phases and CAP rate for each subject

Subject	Time in bed (min)	TST (min)	SE (%)	Wake time (min)	NREM time (min)	REM time (min)	A1 phases (CAP)	A2 phases (CAP)	A3 phases (CAP)	A1 phases (isolated)	A2 phases (isolated)	A3 phases (isolated)	B phases	CAP rate (%)
1	513.5	495	96.40	18.5	371	124	274	111	61	48	26	10	446	57.35
2	526	491	93.35	35	359	132	296	84	76	31	30	17	456	61.31
3	492.5	459	93.20	33.5	335.5	123.5	285	120	37	53	29	6	442	55.67
4	500.5	437.5	87.410	63	338.5	99	152	58	115	53	21	45	325	45.84
5	515	455	88.35	60	342	113	163	39	31	55	19	17	233	31.54
6	490	405.5	82.76	84.5	323	82.5	135	93	65	61	35	9	293	39.29
7	495	479.5	96.87	15.5	330.5	149	183	28	19	35	16	21	230	31.74
8	492.5	469	95.23	23.5	370	99	228	90	66	21	38	29	384	45.31

All the parameters derive by scorings performed by sleep experts

- *EEG variance* It was computed from the raw EEG signal on 1-s windows. The variance difference between adjacent 1-s windows was calculated and the result normalized by its maximum value. This descriptor is expected to account for the abrupt frequency shifts occurring in correspondence with the activations.

Figure 2 shows an example of the trend of the descriptors.

Since each feature refers to a resolution of 1 s, for an approximately 8-h-long sleep and 8 subjects, we had a total of 240,429 samples, 26,305 of which accounted for activations and the remaining 214,124 accounting for the background.

In order to homogenize the values of the descriptors around the activations, a moving window *f*(*x*) was applied to all the band descriptors besides the delta, and to the activity descriptor in the delta band, which is given by:

$$f(x_t) = \max_{k \in [-r,+r]} x_{t+k} \qquad (3)$$

where x_t is the center of the window, and r is equal to 2 for the band descriptors and to 1 for the activity.

For the differential variance of the EEG, the mere absolute value was computed.

All these seven features were used for training the four presented machine-learning methods.

A detailed study of the features here employed has been shown in a previous study of the same authors, where the features have been evaluated by means of ROC curves, and redundancies in the feature set eliminated by correlation analyses [18]. All these features have been shown to have relevant information content in discriminating CAP A phases from the background, especially the delta descriptor, the delta activity and the differential EEG variance.

2.3 Statistical measurements

We define as a true positive (TP) every second recognized as belonging to an activation both by the human and the automatic classifiers, as a false positive (FP) every second belonging to the background but recognized as belonging to an activation by the automatic classifier, as a false negative (FN) every second belonging to an activation but recognized as belonging to the background by the automatic classifier, as a true negative (TN) every second recognized as belonging to the background both by the human and the automatic classifiers.

The statistics sensitivity [TP/(TP + FN)], specificity [TN/(TN + FP)], and accuracy [(TP + TN)/TP + TN + FP + FN] were calculated through a simple second-by-second comparison of the automatic classification vectors with the visual classification vector.

Furthermore, the statistical parameter Cohen's kappa was computed. This parameter is able to provide a measure of the agreement between different scorers (or different scoring methods), statistically discarding the cases in which the agreement is due to chance. The equation for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \qquad (4)$$

where Pr(*a*) is the relative observed agreement among raters (accuracy), and Pr(*e*) is the hypothetical probability

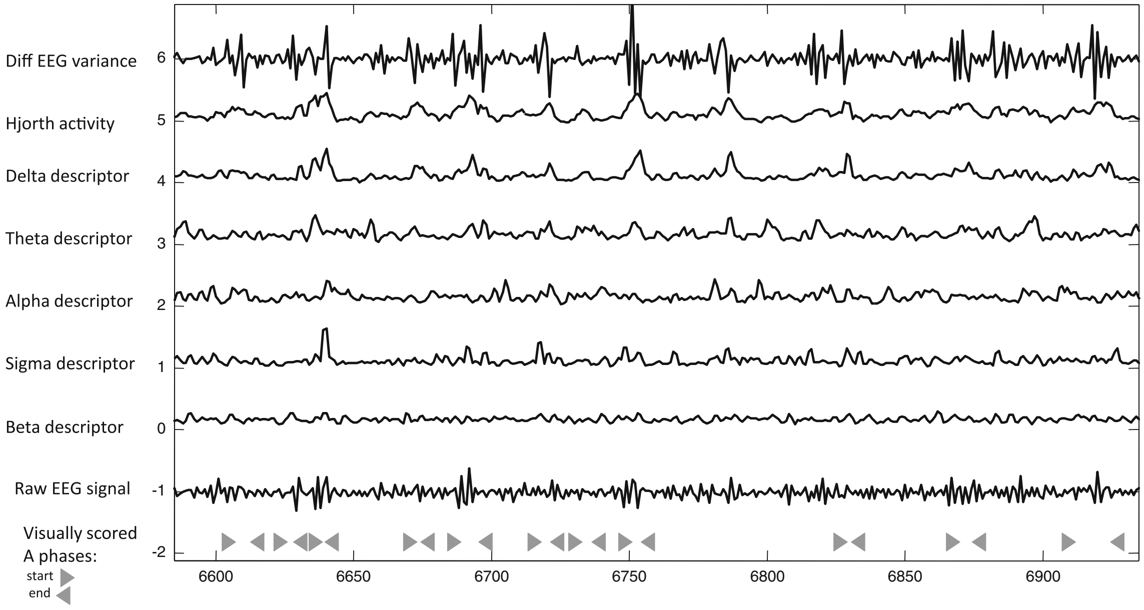


Fig. 2 Example of the trend of the descriptors in correspondence of visually scored A phases

of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$ [10].

2.4 Training stage

The previously described features were used to train the classifiers. Four different classifier types were taken into account and their performances were compared: the discriminant function, the support vector machines, the adaptive boosting and the neural networks. A detailed mathematical description of these methods is reported in the Appendix (ESM).

The data were classified by using the leave one out procedure: one subject at the time was classified with the algorithm previously trained over the remaining seven subjects' data.

The desired output data consisted of binary vectors of the same length of the features (one sample per second), where each sample was assigned value 1 if belonging to a visually scored activation, 0 if belonging to the background. Since there is more background than activation, there were more zeros than ones. Thus, in order to avoid biasing the classifier, a re-sampled training set was created that included an equal number of samples indicative of an A phase and of those indicative of the background. This was done simply by taking into account only a fraction of the samples corresponding to the background, uniformly distributed through the original descriptors.

2.4.1 Discriminant

Three types of discriminant function were tried in order to select the one providing the best classification:

- *Linear* Dividing the feature space by a hyperplane decision surface that maximizes the ratio of between-class variance to within-class variance.
- *Quadratic* Dividing the feature space by a hypersphere, hyperellipsoid or hyperhyperboloids decision surface that maximizes the ratio of between-class variance to within-class variance.
- *Mahalanobis* Exploits the Mahalanobis distance

$$D^2 = (\mathbf{x} - \mu)'\Sigma^{-1}(x - \mu) \qquad (5)$$

where $\mathbf{x}$ is the vector of the data, μ is the centroid of a certain class, and Σ is the covariance matrix of the data distribution, and assigns each datum $\mathbf{x}$ to the class μ that minimizes D^2 (see Appendix) [4].

The statistics in Table 2 compare the performance of the three discriminant functions. A t test was applied to compare the average performances of the classifiers, requiring $p < 0.05$ as the statistical significance level.

As shown by the table, the discriminant function offering the best performance (highest accuracy and Cohen's kappa) is the linear, thus, it was chosen for the classification. The statistical significance is verified for the k index with a p value equal to 0.0198 for the linear discriminant versus the quadratic, equal to 5.26×10^{-8} for the linear versus the Mahalanobis discriminant.

2.4.2 Support vector machines

Support vector machines are motivated by many of the same considerations as the linear discriminant, but rely on preprocessing the data to represent patterns in a high dimension—typically much higher than the original feature space. With an appropriate nonlinear mapping Φ to a sufficiently high dimension, data from two categories can always be separated by a hyperplane. The mapping function Φ is called kernel. A soft-margin support vector machine was implemented, that means introducing variables that measure the degree of misclassification of each datum and assigning a penalty C to such error (see Appendix) [2, 8].

The first step consisted in finding the best SVM parameters to classify the data. Two separate studies were carried out for the two types of SVM kernel: *polynomial* and *Gaussian* [29].

For what concerns the polynomial kernel, the varying parameters were the error penalty C, and the polynomial order o. C varied from 2^{-30} to 2^{12} while o varied from 1 (therefore including also the linear kernel case) to 6.

For what concerns the Gaussian kernel, the varying parameters were the error penalty C, and the Gaussian standard deviation σ. C varied from 2^{-5} to 2^{12}, σ varied from 2^{-12} to 2^{5}.

Table 2 Comparison of the performances of the three discriminant functions in classifying the data

Discriminant function	Sensitivity (%)	Specificity (%)	Accuracy (%)	Cohen's kappa
Linear	72.5 ± 10.9	86.6 ± 6.3	84.9 ± 4.8	0.45 ± 0.05
Quadratic	69.9 ± 8.0	88.0 ± 3.8	84.1 ± 5.9	0.41 ± 0.07
Mahalanobis	95.8 ± 7.5	44.8 ± 15.2	50.4 ± 8.7	0.18 ± 0.05

The mean classification statistics and their respective standard deviations over the eight subjects are reported

The leave one out cross-validation method was applied. One subject at the time was taken out of the dataset. The data of the remaining seven subjects were used for the determination of the optimal kernel parameters, in the following way: for each combination of the parameters, one of the seven subjects in turn was used as testing set, while the remaining six were used as training set. Then, the Cohen's kappa was computed. In order to pick the parameters that led to the best classification, the statistics were averaged on the seven patients' classifications. The best parameters were identified as those maximizing the Cohen's kappa. Such parameters were used to classify the remaining subject's data.

The procedure is synthesized in Fig. 3.

Figure 4 reports an example of the Cohen's kappa values, averaged over seven subjects, while varying the values of the parameters.

The optimal parameters concerning the polynomial kernel ranges are: $C = 2^{11}–2^{12}$ and polynomial order = 3–4; while for the Gaussian kernel, the optimal parameters ranges are: $C = 2^{9}–2^{12}$ and Gaussian standard deviation = $2^{-2}–2^{0}$.

The mean values for the classification statistics obtained using the two kernels with the parameters determined with this procedure were compared, applying the t test and requiring $p < 0.05$, as the statistical significance level.

The polynomial kernel offers a slightly better performance, measured with the k index, with respect to the Gaussian, with a p value equal to 0.0081. The average results over the eight subjects are reported in Table 3.

2.4.3 AdaBoost

AdaBoost, from "Adaptive Boosting", is an algorithm that combines a certain number of classifiers, called "weak learners", to form an ensemble whose joint decision rule has arbitrarily high accuracy on the training set. The final classification decision is composed by the weighted sum of the outputs of all the classifiers (see Appendix) [4].

Similar to what was done for the SVM, the leave one out technique was applied for the selection of the parameter of interest, in this case, the number of weak learners to employ.

Each analysis was carried out increasing the number of weak learners used from 1 to 20. The data from one subject at the time were used as testing set, and for each number of weak learners the testing error was computed as follows:

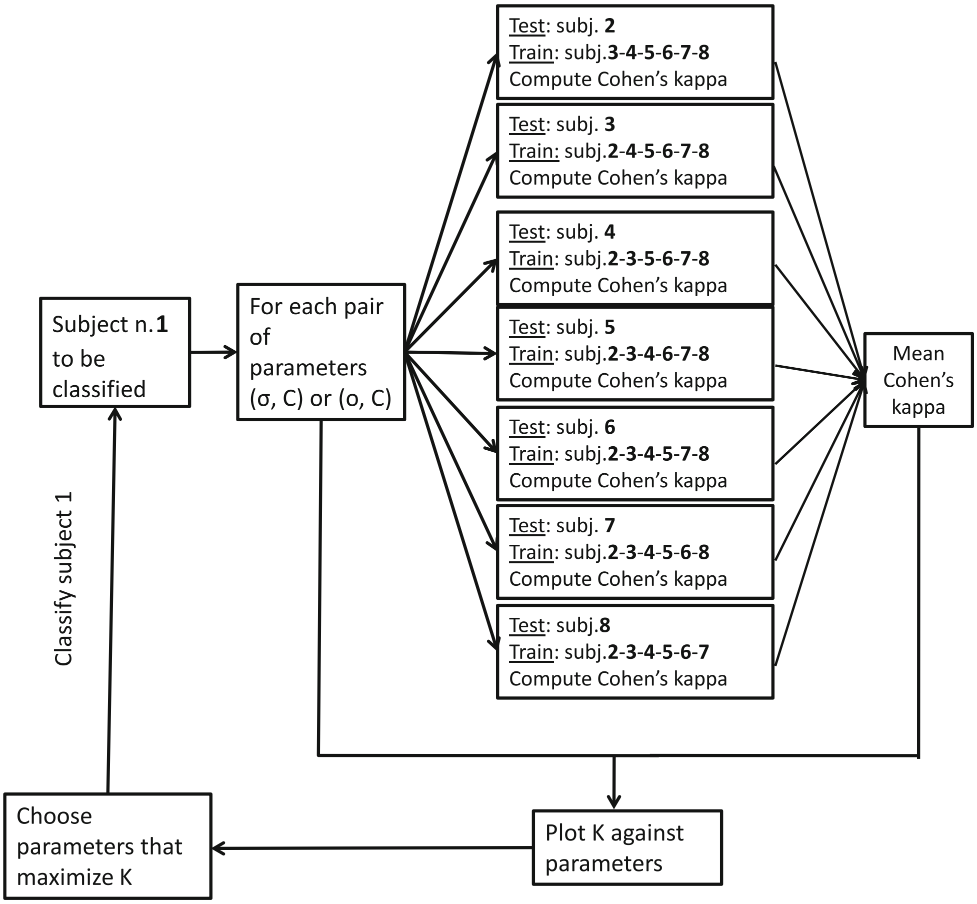


Fig. 3 LOO procedure for the estimate of the optimal classification parameters (σ and C for Gaussian kernel, o and C for the polynomial kernel) and classification of the data

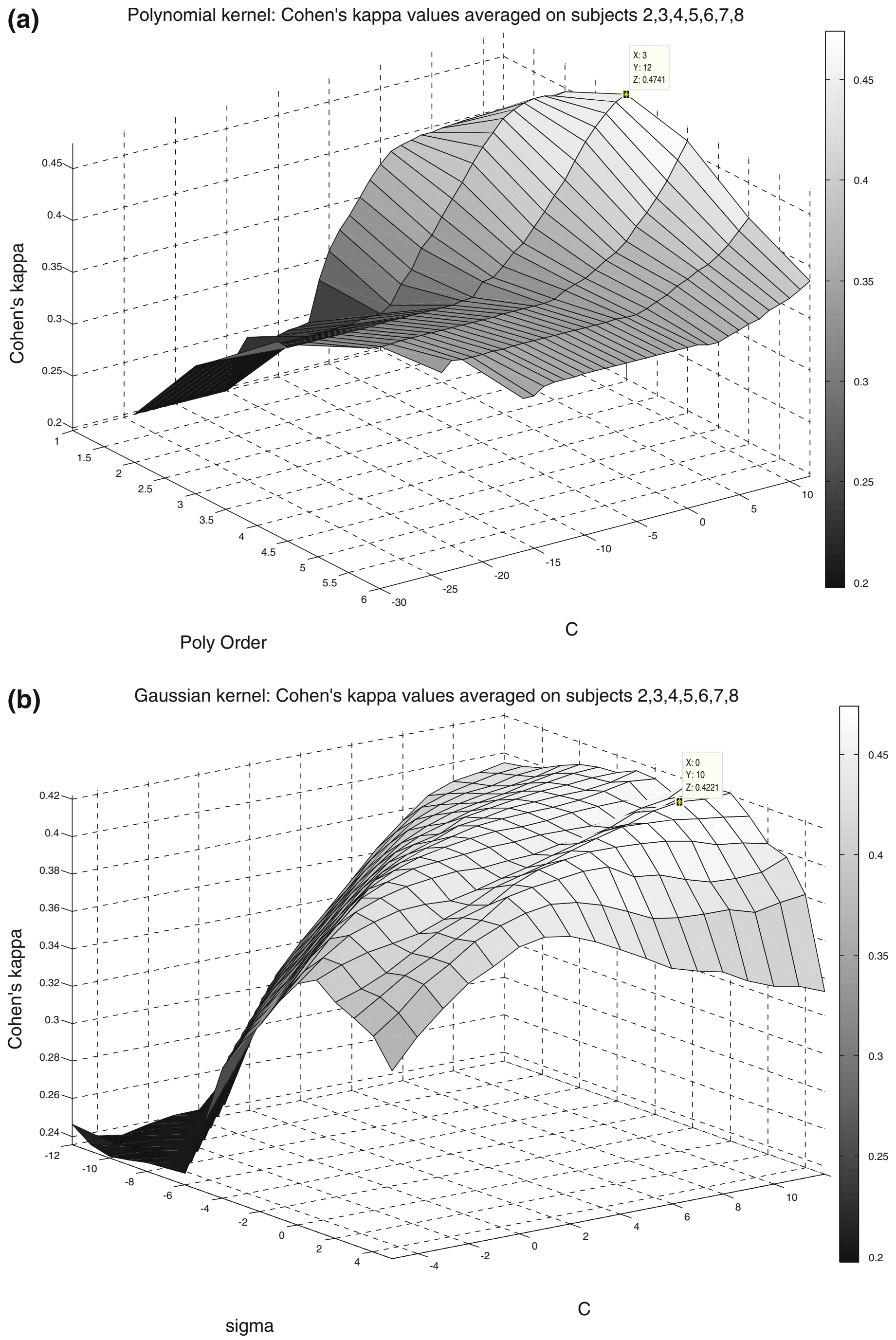


Fig. 4 **a** Polynomial kernel: mean Cohen's kappa values averaged over seven subjects for varying values of the penalization parameter, *C* and of the polynomial order, *o*. **b** Gaussian kernel: mean Cohen's kappa values averaged over seven subjects for varying values of the penalization parameter, *C* and the Gaussian standard deviation, σ. The *boxes* report the optimal couples chosen for the parameters in this example case

Table 3 Results obtained with a SVM using a polynomial kernel or a Gaussian kernel

Kernel	Sensitivity (%)	Specificity (%)	Accuracy (%)	Cohen's kappa
Polynomial	70.1 ± 8.6	84.0 ± 11.1	81.9 ± 7.8	0.44 ± 0.08
Gaussian	75.0 ± 7.5	78.0 ± 11.2	77.7 ± 8.2	0.39 ± 0.10

Mean and standard deviation over the eight subjects' recordings

$$\text{Testing}_{\text{error}} = \frac{n - (\text{TP} + \text{TN})}{n} = 1 - \text{Accuracy} \qquad (6)$$

The testing errors for the eight analyses were averaged and plotted with respect to the number of weak learners.

Figure 5 illustrates an example of the trend of the training and the testing error. It can be noticed how they both decrease with respect to the number of weak learners used. This was true for all the LOO reiterations. Thus, a

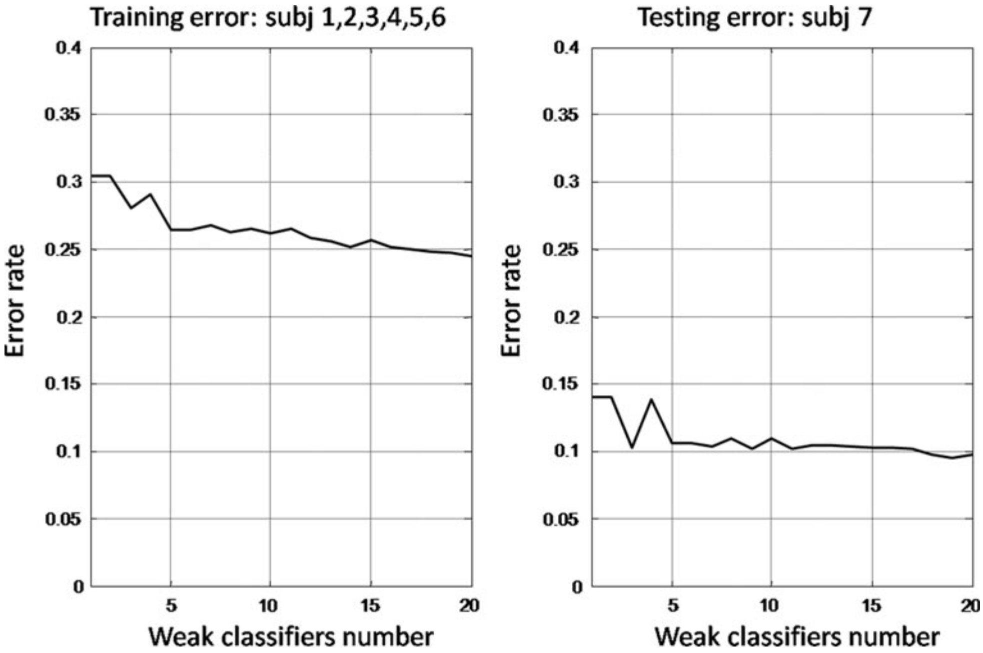


Fig. 5 Training (*left*) and testing (*right*) error evaluated with respect to the number of weak learners employed

number of weak learners equal to 20 was considered optimal for this type of classification.

2.4.4 Neural network

An artificial neural network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of several processing elements (neurons) that are able of adapting their parameters to learn from a specific training set (*inductive learning*), in a systematic fashion (*training mode*). If the examples in the training set are accompanied by labels, we talk about *supervised learning*, otherwise, the learning is *unsupervised* (see Appendix) [21, 22].

A three-layer supervised neural network was chosen, with a 7-neuron input layer, a x-neuron hidden layer, and a 1-neuron output layer. The number x of the hidden layer neurons varied from 2 to 30.

The chosen activation function was *logsig* for the hidden and output layers. The training mode was the backpropagation with the Levenberg–Marquardt algorithm

In order to achieve an adequate classification performance for each subject, we proceeded in the following way: for each test subject, the neural networks were trained using the remaining seven subjects' data, and partitioning them using the leave one out technique: the data of one of the seven subjects at the time were used as the *testing set*, while the remaining six subjects' data were equally divided into *training set* and *validation set*. With these data, and for each value of x, $2 < x < 30$, the neural networks were trained and restarted 10 times, and the one with the best performance was chosen, in order to avoid local minima problems. The network with the best performance, i.e., the one with the lowest testing error, was then chosen among those obtained with different values of x.

This *modus operandi* led to 7 "best" neural networks for each of the 8 subjects, for a total of 56 networks.

The seven networks were then simulated on the corresponding subject's data, in order to obtain seven classification vectors. The seven vectors were rounded to 0 or 1 by setting a threshold at 0.5 and the final classification vector was computed second by second thanks to a majority voting system: thanks to the odd number of voters (the seven output vectors), ties were avoided. An example of the procedure for the classification of subject number 1 is synthesized in Fig. 6.

2.5 Post-processing stage

According to Terzano's rules, [26] CAP A phases cannot last more than 60 s or less than 2 s, thus the automatically recognized activations lasting longer than this limit had to be re-classified. This was necessary for only an approximately 1% of the automatically recognized A phases. The re-classification was performed using a competitive neural network for clustering, with 2 neurons and 500 epochs, that received in input only the delta and the beta band descriptors and split the original activations into shorter-lasting ones, as shown in Fig. 7a). An alternative clustering method such as K-means [17] was tried for this re-classification, but it was eventually discarded because it offered a poorer classification performance with respect to the competitive neural network.

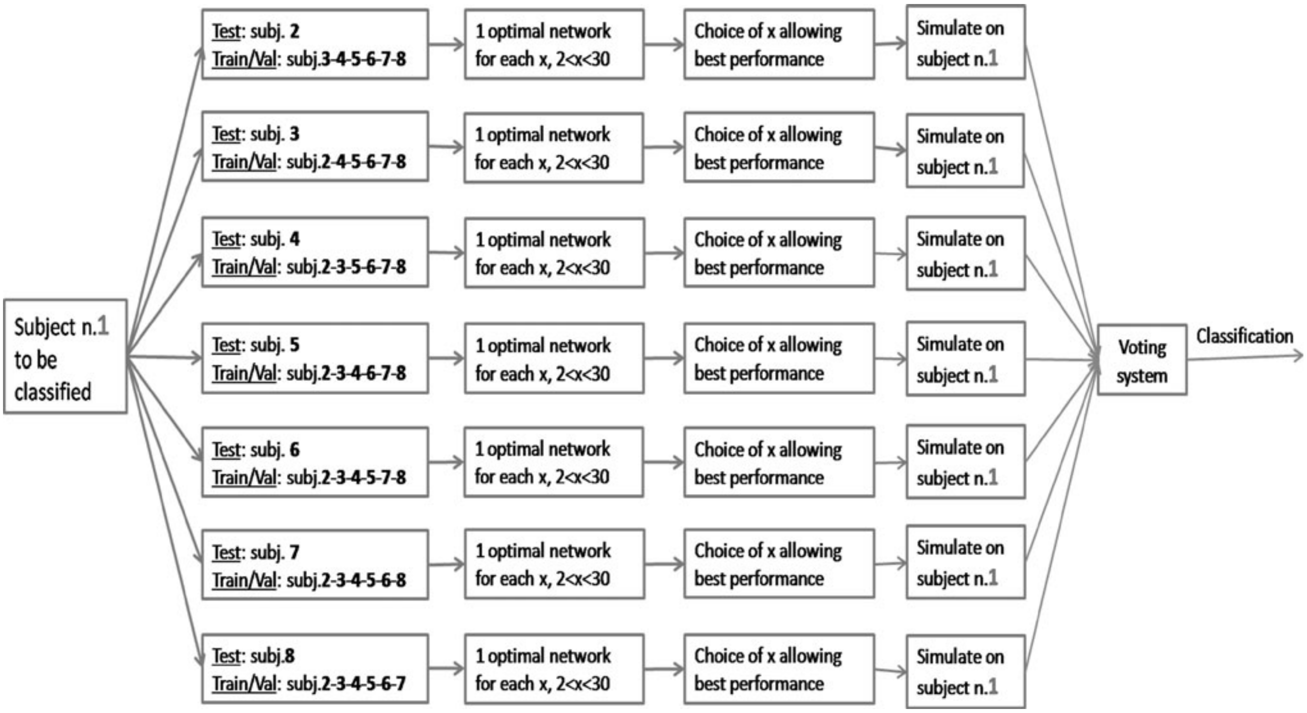

Subject n.1 to be classified
Test: subj. 2
Train/Val: subj.3-4-5-6-7-8
Test: subj. 3
Train/Val: subj.2-4-5-6-7-8
Test: subj. 4
Train/Val: subj.2-3-5-6-7-8
Test: subj. 5
Train/Val: subj.2-3-4-6-7-8
Test: subj. 6
Train/Val: subj.2-3-4-5-7-8
Test: subj. 7
Train/Val: subj.2-3-4-5-6-8
Test: subj.8
Train/Val: subj.2-3-4-5-6-7
1 optimal network for each x, 2<x<30
Choice of x allowing best performance
Simulate on subject n.1
Voting system
Classification

Fig. 6 Scheme of the procedure for the choice of the seven optimal neural networks for a generic subject. Subject 1 is taken out from the set, one of the remaining subjects at the time is used as the testing set, while the remaining six subjects are divided in half and used as training and validation sets. For each *x* (number of hidden layer neurons) the best network is chosen after ten restarts. The optimal network is eventually chosen among those with different values of *x*. The seven resulting networks are simulated on Subject 1 data obtaining seven classification vectors. For each second, a voting system among the seven classifiers provides the final classification

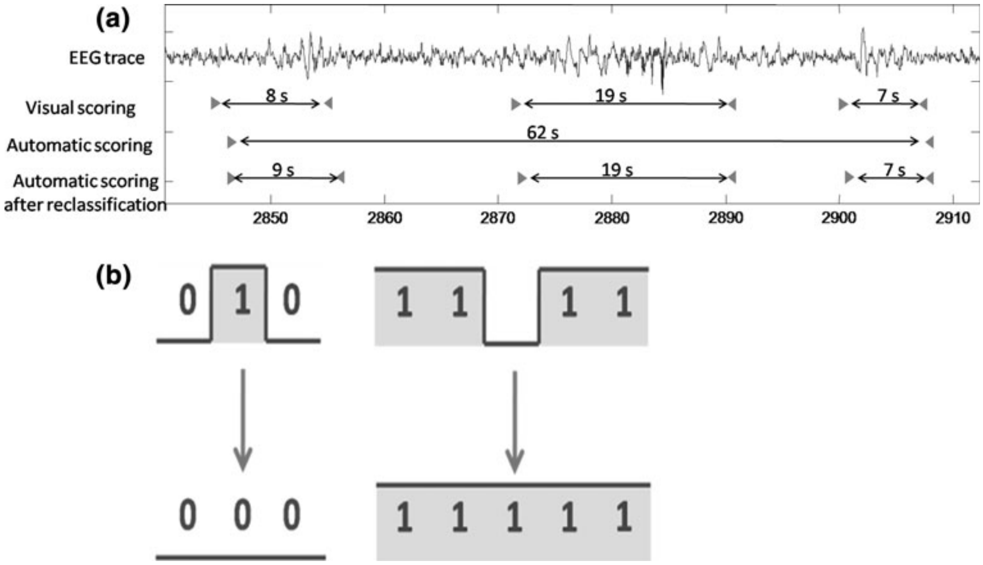


Fig. 7 **a** Example of re-classification of a fragment of activation lasting longer than 60 s. The *first set of arrows* shows the classification computed by the expert clinician. The second shows an example of classification performed by the automatic algorithm that exceeds the 60 s limit. The *bottom line* reports the result of the re-classification performed via the competitive NN. **b** Scheme of the post-processing procedure

Moreover, a post-processing was applied to eliminate automatically detected A phases and B phases shorter than 2 s, as shown in Fig. 7b): isolated ones and isolated zeros, indicating A phases and B phases lasting approximately 1 s, respectively, were replaced by a 0 and a 1, respectively. It must be highlighted that the three operations happen in sequence: (1) re-classification step, (2) replacement of isolated zeros, (3) replacement of isolated ones. A double-check was applied to avoid eventual merging into segments longer than 60 s due to the post-processing.

3 Results

The statistics averaged over the eight subjects, are reported in Table 4.

Table 4 Results obtained with each classifier

Method	Sensitivity (%)	Specificity (%)	Accuracy (%)	Cohen's kappa
Linear discriminant	72.5 ± 10.9	86.6 ± 6.3	84.9 ± 4.8	0.45 ± 0.05
SVM	70.1 ± 8.6	84.0 ± 11.1	81.9 ± 7.8	0.44 ± 0.08
AdaBoost	68.5 ± 6.7	79.3 ± 9.4	79.4 ± 5.5	0.41 ± 0.11
Neural network	72.9 ± 7.5	82.3 ± 7.1	81.5 ± 6.4	0.45 ± 0.20

Mean and standard deviation over the eight subjects' recordings

Table 5 Comparison among the performances of the four classifiers obtained by applying the paired t test to the classification statistics, and requiring $p < 0.05$ as the statistical significance level

	Sensitivity	Specificity	Accuracy	Cohen's kappa
LD vs. SVM				
LD vs. AdaBoost		Higher for LD		Higher for LD
LD vs. NN			Higher for LD	
SVM vs. AdaBoost		Higher for SVM		
SVM vs. NN	Higher for NN			
NN vs. AdaBoost	Higher for NN			

Filled cells indicate statistical significance according to the t test with $p < 0.05$ while empty cells indicate no statistical significance

A comparison among the classifiers' performance was conducted applying the paired t test to the classification statistics, and requiring $p < 0.05$ as the statistical significance level and is reported in Table 5. Empty cells indicate no statistical significance. Filled cells indicate statistical significance according to the t test with $p < 0.05$.

The SVM leads to a high accuracy, equal to 81.9% and specificity (84.0%) but to a lower sensitivity with respect to the NN. AdaBoost leads to good accuracy and specificity (79.4%, 79.3%), but to a lower sensitivity (68.5%). The main drawback of AdaBoost is the long computational time required for its training, especially when a large number of classifiers is employed.

The linear discriminant and the neural network seem to be the better-performing methods, showing high accuracy values, equal to 84.9 and 81.5%, and Cohen's kappa values, equal to 0.45. Examples of classification performed by these two algorithms are shown in Fig. 8.

All the correctly recognized phases were circled with a thin black box:

- Solid lines are used for automatically recognized A phases that coincide with visually recognized A phases.
- Dashed lines are used for visually recognized A phases that have been automatically recognized as more than one A phase.
- Dotted lines are used for automatically recognized A phases that are longer than visually recognized A phases.

All the incorrectly recognized phases have been circled with a thicker gray box. Some of the false positive A phases show an evident increase of more than one descriptor so, according to the criteria used by the discriminant, they should be classified as true positives. Maybe, this kind of misclassification can be due to the variability of classification between different human scorers [24]. A possible solution could be having more than one visual classification of A phases in order to better understand this kind of error. Though the performances of the algorithms are comparable, the neural network (Fig. 8b), seems to better follow short-lasting activations.

4 Discussion

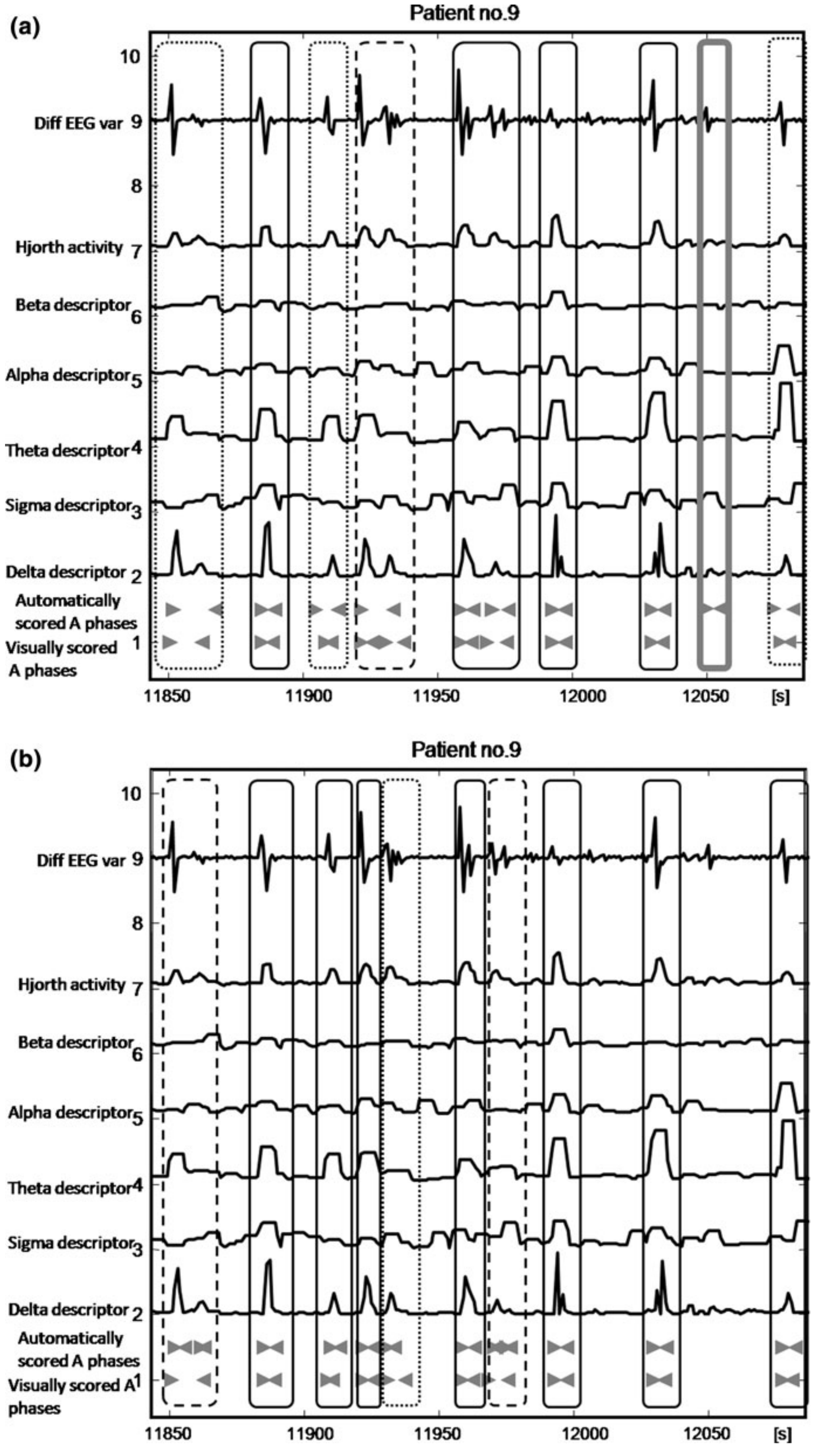
A statistical comparison among different classifiers to automatically distinguish the EEG activations that characterize the cyclic alternating pattern was presented.

Our main achievements are:

(a) The sleep microstructure can be automatically detected using a single EEG lead with good accuracy.
(b) The proposed methods are entirely automatic, without any need of human intervention.
(c) With a reduced number of descriptors, a complete frame of the EEG variations that occur during CAP A phases could be captured.

We can observe that, due to the complexity of the data, inductive machine-learning methods constitute much more accurate classifiers than a simple threshold method, employed in many studies in literature [1, 7, 16, 19].

Fig. 8 Example of classification via: **a** linear discriminant, **b** neural network. In both figures, the descriptors (*from the top*) differential variance of the EEG signal, Hjorth activity, beta, alpha, theta, sigma and delta descriptors are represented. The CAP phases A are shown at the *bottom* (beginning and end marked by the *gray arrows*), where the automatic scoring is compared to the visual one. The *boxes* highlight the correctness of classification: *solid black lines* highlight correct classification, *solid gray lines* highlight incorrect classification, *dotted lines* indicate correct recognition, but wrong duration, *dashed lines* highlight more than one phase A recognized by the automatic scorer where the visual scorer had selected a single event



The linear discriminant and the artificial neural network seem to be the better-performing algorithms in the classification, although the performance of the four methods is rather similar. This suggests that a peculiar role in a good classification is played not only by the choice of the specific machine-learning classifier, but by the selection of appropriate descriptors.

A possible idea for a future improvement of the method is that of splitting the EEG signal into windows of different lengths. In fact, here, for ease of computation, all the descriptors have been computed on 1 s-long windows, although, being the EEG a non-stationary signal, it would perhaps be more appropriate to employ segmentation techniques such as that proposed in [9] in order to split it into windows in which it maintains uniform statistical properties, and to compute the features on these new windows.

The use of a single EEG trace makes the algorithms easy to implement and reduces the computational burden of the methods: the introduction of a second trace, perhaps a frontal derivation, where the delta components are better represented, could somehow improve the classification at the expense of some computational load.

Among the automatic methods used to detect the sleep microstructure, some published techniques require some sort of intervention by the clinician [7], obtaining moderate classification results and reducing the time for the detection of the microstructure. The advantage of the proposed methods is their total independence from any a priori information besides the mere REM/NREM distinction.

There are a few studies [1, 16] that have employed band descriptors used as training features similar to those used in this paper. However, the introduction of new features, other than the band descriptors, improves the classification: the Hjorth activity descriptor is able to better account for the average increase of delta power during activations, whereas the differential variance of the raw EEG signal captures the abrupt frequency variations occurring during CAP A phases.

Moreover, differently from previous studies [1, 16, 19], that also report high accuracy values, ranging around 77–84%, here all the statistics were computed not only by applying a mere overlap criterion between visually and automatically scored activations, but considering each 1-s window as an observation, leading to a much more precise statistic.

As it can be seen from the high standard deviations in the statistics (see Table 4), the results are strongly dependent on the subject, even after the normalization of the descriptors.

This is probably due to some inter-subject variability of the CAP A phase rhythms. The subject-dependency could also be due to the scoring performed by different clinicians: in fact, as we mentioned before, the repeatability between classifications of the same EEG trace performed by two different experts ranges between 69 and 77% [24]. Thus, an automatic method based on training data could distinguish the activations with a criterion that is similar to that of a certain human scorer, but dissimilar to that of another.

In spite of the limited number of subjects, the statistics obtained are encouraging, and suggest that better results could be obtained increasing the dimensions of the dataset and thus the size of the training set for the automatic algorithms.

The limited number of subjects available for studies on CAP is indeed a critical issue and the main reason why CAP scoring still remains a debated topic in sleep medicine. The development of an efficient automatic classifier, on the other hand, could allow for the quick scoring of a large number of sleep recordings, that could then be double-checked by experts, leading to significant advancements in the field.

The intrinsic characteristics of these methods increase the potential discrepancy between visual and automatic definition of phase A duration. In fact, while the clinician performs her scoring by visualizing portions of the EEG traces of duration approximately equal to 30 s, the automatic recognition methods are based on 1-s moving windows, and classifies each second as a separate entity. In order to overcome this problem, before identifying CAP sequences based on the recognition of A phases and the application of the duration criteria, we suggest having an expert perform an a posteriori validation and control of the duration of each potential A phase as scored by the chosen automatic algorithm.

As described in the Sect. 1, CAP A phases can belong to three subtypes (A1, A2 and A3), with different frequency–amplitude characteristics and distribution through the night. As a final idea for a further development, an algorithm capable of distinguishing among the three A phase subtypes could be implemented, that exploits filters in low and high frequency bands to attribute the activation to subtype 1, 2 or 3, depending on the de-synchronization time/total time rate.

In conclusion, the present study could constitute a good starting point for the development of an efficient CAP detection tool for the use in clinics, allowing to speed up the study of sleep microstructure and avoid the difficulties due to human rater disagreement, together with providing a means to shed light on the physiological mechanisms that are at the basis of CAP.

Acknowledgments This work was supported in part by PSYCHE under Project FP7-247777 of the European Community.

References

1. Barcaro U, Bonanni E, Maestri M, Murri L, Parrino L, Terzano MG (2004) A general automatic method for the analysis of NREM sleep microstructure. Sleep Med 5:567–576
2. Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2:121
3. Chervin RD (2011) Engineering better sleep. Med Biol Eng Comput 49:623. doi:10.1007/s11517-011-0777-4
4. Duda RO, Hart PE, Stork DG (2000) Pattern classification. Wiley, New York
5. Eisensehr I, Parrino L, Noachtar S, Smerieri A, Terzano MG (2001) Sleep in Lennox–Gastaut syndrome: the role of the cyclic

alternating pattern (CAP) in the gate control of clinical seizures and generalized polyspikes. Epilepsy Res 46:241–250

6. El-Ad B, Chervin RD (2000) The case of a missing PLM. Sleep 23:450–451
7. Ferri R, Bruni O, Miano S, Smerieri A, Spruyt K, Terzano MG (2005) Inter-rater reliability of sleep cyclic alternating pattern (CAP) scoring and validation of a new computer-assisted CAP scoring method. Clin Neurophysiol 116:696–707
8. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16:906–914. doi:10.1093/bioinformatics/16.10.906
9. Gharieb RR (2001) Segmentation and tracking of the electroencephalogram signal using an adaptive recursive bandpass filter. Med Biol Eng Comput 39:237. doi:10.1007/BF02344808
10. Gwet K (2002) Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. Stat Methods Inter Rater Reliab Assess 2
11. Haba-Rubio J, Staner L, Macher JP (2002) Periodic arousals or periodic limb movements during sleep? Sleep Med 3:517–520
12. Halász P, Terzano MG, Parrino L (2002) Spike-wave discharge and the microstructure of sleep–wake continuum in idiopathic generalised epilepsy. Neurophysiol Clin 32:38–53
13. Hjorth B (1970) EEG analysis based on time domain properties. Electroencephalogr Clin Neurophysiol 29:306–310. doi:10.1016/0013-4694(70)90143-4
14. Iber C, Ancoli-Israel S, Chesson A, et al. (2007) The AASM manual for scoring of sleep and associated events: rules, terminology and technical specifications, 1st edn. American Academy of Sleep Medicine, Westchester, IL
15. Kato T, Montplaisir JY, Guitard F, Sessle BJ, Lund JP, Lavigne GJ (2003) Evidence that experimentally induced sleep bruxism is a consequence of transient arousal. J Dent Res 82:284–288
16. Largo R, Munteanu C, Rosa A (2005) CAP event detection by wavelets and GA tuning. In: Proceedings of 2005 IEEE international workshop on intelligent signal processing, pp 44–48
17. Macqueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol 1, pp 281–297
18. Mariani S, Manfredini E, Rosso V, Mendez MO, Bianchi AM, Matteucci M, Terzano MG, Cerutti S, Parrino L (2011) Characterization of A phases during the cyclic alternating pattern of sleep. Clin Neurophysiol 122(10):2016–2024
19. Navona C, Barcaro U, Bonanni E, Di Martino F, Maestri M, Murri L (2002) An automatic method for the recognition and classification of the A-phases of the cyclic alternating pattern. Clin Neurophysiol 113:1826–1831
20. Parrino L, Boselli M, Buccino GP, Spaggiari MC, Di Giovanni G, Terzano MG (1996) The cyclic alternating pattern plays a gate-control on periodic limb movements during non-rapid eye movement sleep. J Clin Neurophysiol 13:314–323
21. Pearlmutter BA (1990) Dynamic recurrent neural networks. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA
22. Principe JC (2010) Information theoretic learning: Renyis entropy and kernel perspectives. In: Principe JC (ed) Information science and statistics. Springer, Berlin
23. Rechtscahffen A, Kales A (1968) A manual of standardized terminology, techniques and scoring system for sleep stages in human subjects. National Institutes of Health Publications 204
24. Rosa A, Alves GR, Brito M, Lopes MC, Tufik S (2006) Visual and automatic cyclic alternating pattern (CAP) scoring: inter-rater reliability study. Arq Neuropsiquiatr 64
25. Terzano MG, Parrino L, Spaggiari MC, Palomba V, Rossi M, Smerieri A (2003) CAP variables and arousals as sleep electroencephalogram markers for primary insomnia. Clin Neurophysiol 114:1715–1723
26. Terzano MG, Parrino L, Sherieri A, Chervin R, Chokroverty S, Guilleminault C, Hirshkowitz M, Mahowald M, Moldofsky H, Rosa A, Thomas R, Walters A (2001) Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. Sleep Med 2:537–553
27. Terzano MG, Parrino L, Boselli M, Spaggiari MC, Di Giovanni G (1996) Polysomnographic analysis of arousal responses in obstructive sleep apnea syndrome by means of the cyclic alternating pattern. J Clin Neurophysiol 13:145–155
28. Terzano MG, Parrino L, Anelli S, Halasz P (1989) Modulation of generalized spike-and-wave discharges during sleep by cyclic alternating pattern. Epilepsia 30:772–781
29. Tong S, Koller D (2002) Support vector machine active learning with applications to text classification. J Mach Learn Res 2:45–66