J Med Syst (2012) 36:2675–2688
DOI 10.1007/s10916-011-9742-x

ORIGINAL PAPER

Decision Support Algorithm for Diagnosis of ADHD Using Electroencephalograms

Berdakh Abibullaev · Jinung An

Received: 22 February 2011 / Accepted: 23 May 2011 / Published online: 15 June 2011

Abstract Attention deficit hyperactivity disorder is a complex brain disorder which is usually difficult to diagnose. As a result many literature reports about the increasing rate of misdiagnosis of ADHD disorder with other types of brain disorder. There is also a risk of normal children to be associated with ADHD if practical diagnostic criteria are not supported. To this end we propose a decision support system in diagnosing of ADHD disorder through brain electroencephalographic signals. Subjects of 10 children participated in this study, 7 of them were diagnosed with ADHD disorder and remaining 3 children are normal group. Our main goal of this sthudy is to present a supporting diagnostic tool that uses signal processing for feature selection and machine learning algorithms for diagnosis.Particularly, for a feature selection we propose information theoretic which is based on entropy and mutual information measure. We propose a maximal discrepancy criterion for selecting distinct (most distinguishing) features of two groups as well as a semi-supervised formulation for efficiently updating the training set. Further, support vector machine classifier trained and tested for identification of robust marker of EEG patterns for accurate diagnosis of ADHD group. We demonstrate that the applicability of the proposed approach provides higher accuracy in diagnostic process of ADHD disorder than the few currently available methods.

Keywords EEG · ADHD · Mutual information · Shannon theory · SVM · ADHD diagnosis

Introduction

Attention-deficit hyperactivity disorder (ADHD) is one of the most common neurological disorders that affect 3–7(%) percent of school-age children, characterized by developmentally inappropriate levels of inattention, impulsivity, and/or hyperactivity. It is considered to be a chronic condition with 3 to 5(%) of individuals diagnosed continuing to have symptoms into adulthood [1, 2]. It is also considered a developmental disorder accompanied by learning disabilities, depression, anxiety and conduct disorder. The etiology of ADHD is still unknown, and the disorder may have several different causes. Scientists have studied, for example, the relation of ADHD to the abnormal brain function, morphologic brain differences, and electroencephalograph (EEG) patterns [3–5]. The diagnosis of ADHD is based on the presence of particular behavioral symptoms that are judged to causes significant impairment in an individual's functioning and not on the results of a specific task. Unfortunately, there's no objective laboratory tests (urine, blood, x-ray or psychological analysis) that can support the diagnosis of children as ADHD or not [6]. Therefore, most often the behavioral symptoms of ADHD can be easily confused, even by mental professionals with the routine actions of children. diagnosed as ADHD. Recently, one study reported that approximately one million children

B. Abibullaev · J. An (✉)
Daegu Gyeongbuk Institute of Science and Technology,
Sangri 50-1 Hyonpung Dalseon-Gun, Daegu, 711-873, Korea
e-mail: robot@dgist.ac.kr

B. Abibullaev
e-mail: berdakho@dgist.ac.kr

in the USA are misdiagnosed when they receive an ADHD diagnosis [7]. Quantitative EEG studies have been conducted to address the problems mentioned and support the diagnosis of ADHD. The studies often analyze EEG abnormalities such as slow wave activity, epileptiform or specific EEG frequency bands, event related potentials and coherence measures. It is mainly based on finding that individuals with ADHD have distinctive pattern of brain electrical activity that is characterized by and change of low/high frequency waves of EEG. The changes in EEG frequency is later associated with ADHD diagnosis. Accordingly, many studies have been published; for example recent research findings reported that ADHD children's EEG show fairly consistent difference in their brain electrical activity when compared to normal children, particularly regarding frontal and central theta activity, which is associated with underarousal and indicative of decreased cortical activity [3, 5]. Most studies find excess slow brain activity (theta) and a decreased fast brain activity (beta). Theta EEG activity is often associated with an "inattentive" or a dreamy state, and beta activity is often seen when the brain is very busy with for instance solving a cognitive task [8–13]. It was found that children with ADHD showed increased theta power, slight elevations in frontal alpha power, and diffuse decreases in beta mean frequency. Increased power is the most consistent findings in this ADHD EEG literature, indicating that cortical hypo-arousal is a common neuropathological mechanism in ADHD. All of these studies require long term analysis of EEG signals (e.g. visual inspection, analysis) and subject to variations. Although the use of machine learning methods have been used in the analysis of EEG signals for decades, there are only few studies reported to support the diagnosis of ADHD with practical application. For instance, Mueller, A. et al. analyzed the event related potentials (ERP) of EEG signal in the automatic discrimination of ADHD group from normal group subjects. The authors proposed an Independent component analysis based feature extraction method and further support vector classification algorithm for discrimination. The study reported the classification accuracy of their approach reaches in average 92%, with 90%—sensitivity and 94%—specificity [14]. Another study conducted by Anduradha, J. et al. also reported of applicability of SVM method in automatic diagnosis of ADHD. They used the same classification scheme with different preprocessing and feature extraction stage [15]. Recently, Ahmadlou, M. et al. proposed a new approach by using wavelet neural networks approach.The study demonstrates the efficiency of using wavelet transforms for feature extraction and artificial neural networks for classification between ADHD and normal group [16]. Most of these studies use fixed number of features of limited number of ADHD or normal subjects. As you have noticed the studies in general, employs four stages in research: (1) data acquisition (2) preprocessing (denoising, filtering, etc) (3) feature extraction and finally (4) classification. The third step is perhaps the most significant, since it determines in a high degree the overall performance of the classifier algorithm. A feature extraction method can be considered successful if the resulted features describe the object uniquely in the analyzed signal. As a result one can achieve efficiency in classification accuracy. It is usually hard to find most informative or discriminative features of ADHD children due to the variability of the disorder and the limitation of the available subjects in the study. Besides one should decide the type of features to seek in the research (such as, EEG abnormal waves, ERP, Frequency) which is usually done by an expert in the domain. Sometimes, even a human expert may not be able to construct the features that describe ADHD different from normal group. There's a need for an adaptive algorithm which partially is informed with the available information (feature) about ADHD and further tries to find the good choice for feature selection based on the information available. In this paper we propose a novel adaptive algorithm using information theoretic and a statistical learning theory in order to detect the robust EEG features and classify accordingly to ADHD or normal group. Particularly we use mutual information measure to extract the dominant features of ADHD and normal groups further we extend our algorithm for semi-supervised feature selection method. In semi-supervised algorithm the previously selected training set can be updated provided if new useful EEG characteristics are available (which is not included in the training set). The other advantage is that we try to reduce the redundant features of EEG that is associated with both groups and minimize relevance of ADHD and normal group. In the final step support vector machines are implemented for classification of two groups. To our knowledge this our first study to report in providing such an intelligent algorithm. All of the EEG recordings were obtained from 10 children 7 of them were diagnosed with ADHD according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV; APA, 1994), and 3 normal children. Our ultimate goal is to explore the robust predictive features of EEG that will minimize the misdiagnosis of ADHD with other types of disorder of normal children. Specifically, we report good sensitivity and specificity measures. This is our preliminary attempt to use EEG to support the diagnosis of ADHD.

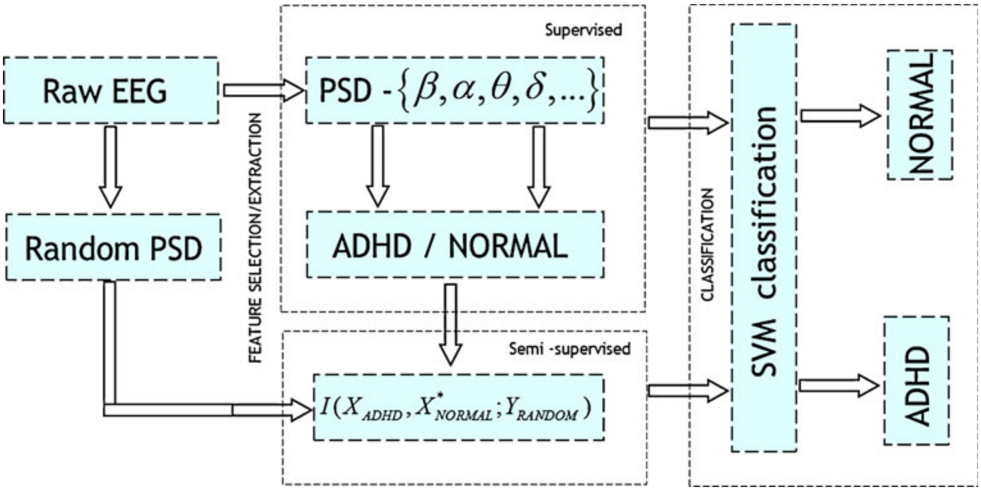
Supervised
Raw EEG
PSD - $\{\beta,\alpha,\theta,\delta,...\}$
Random PSD
ADHD / NORMAL
FEATURE SELECTION/EXTRACTION
Semi -supervised
$I(X_{ADHD}, X^{*}_{NORMAL}; Y_{RANDOM})$
CLASSIFICATION
SVM classification
NORMAL
ADHD

Fig. 1 Flowchart of the methodology

Methods and materials

The methodology proposed in this work is summarized in Fig. 1. It consists of selecting features and defining maximal discrepancy subsets of two class data set; namely ADHD and Normal. Further, a semi-supervised feature selection is performed from the data introduced to the training set. In semi-supervised feature selected we propose a new algorithm which updates the previously selected training set with new unknown data set. In the final step, a support vector machine classifier is implemented to classify the data into Normal and ADHD groups.

Participants

In this study subjects of 10 children, ages of 7 and 12, participated. Participants had a full-scale WISC-III IQ score of 85 or higher. Decision to include children in the ADHD groups was based on a clinical evaluation by pediatricians and psychologists and their mutual agreement on the diagnosis. Clinical interviews incorporated information from as many sources as were available. These included a history given by a parent or school reports for the past 12 months, reports from any other health professionals, and behavioral observations during the assessment. Children were excluded from the ADHD groups if they had a history of a problematic prenatal, prenatal or neonatal period, a disorder of consciousness, a head injury resulting in cognitive deficits, a history of central nervous system diseases, convulsions or a history of convulsive disorders, paroxysmal headaches or tics, or an anxiety or depressive disorder. All children were required to meet the diagnostic criteria for ADHD of the Diagnostic and Statistical Manual of Mental Disorders [17], including current symptoms and a retrospective diagnosis of childhood ADHD. In addition we used Korean version of the Child Behavior Checklist (K-CBCL)[18], WISC-III evaluation methods [19]. Participants took K-CBCL for checking externalized behavior problems and internalized behavior problem. We set cut off score to '70T' in all subtests and we exclude the participants had over 70T in delinquent rule breaking behavior, somatic complaints, withdrawn, anxious/depressed subtitles in K-CBCL. The control group consisted of children from local schools and community groups. Control participants took part in a clinical interview and completed a self-report rating on ADHD and interview with a parent and teacher similar to that of ADHD group.

Data acquisition

EEG measurements of participants were obtained using multichannel G-tec. EEG acquisition system. The sampling rate of the acquired data is 256 Hz and the data is digitized using the 16-bit A/D converter. The 9-mm thin disk Ag/AgCl electrodes were mounted inside the cap with bipolar references behind the ears. Impedance levels were kept less than 5 kOhm. Each signal was amplified and filtered using a 1–40 Hz band-pass filter. A four-pole Butterworth filter was used as a low-pass filter and as an anti-aliasing scheme. The recordings were obtained from the frontal region of the prefrontal cortex according to the 10–20 international system. The frontal regions were the left frontal (Fp1, F3, F7), midline frontal (Fpz, Fz), right frontal (Fp2, F4, F8), midline central (Cz).

Experimental tasks consisted of performing a cognitive task to evaluate the focused attention. In particular, each child performed a focused attention to select one object among multiple choices. Our tasks were designed to evaluate the ability of the participants to discriminate relevant from irrelevant information (i.e., the ability to focus attention). The task is considered as a variant of a widely used CPT. During the task, four objects are presented on the screen. The objects are the pictures of fruits, such as *apple*, *watermelon*, *strawberry* and *banana*. The tasks are presented in three different forms as follows: (1) the voice is activated randomly providing the names of the fruit (while fruits appear on the screen) and the participant should select it accordingly. This task is designed to evaluate the visual focused attention; (2) the same procedure is performed but the difference is the figures appear on the screen after the name of the fruit is asked. (3) This cognitive task is designed to evaluate the focused selective attention of participants. Participants hear the names of the fruit and in four blocks the names of the fruits again pronounced without providing any particular figure. Then, a participant should select the correct block of the given task. The duration of the tasks consists of 12 seconds of pre-task and the task (stimulus) duration was 1200 ms. Each task was performed 10 times at different days. This is actually the part of our long term research which lasted during 6 months. Our cognitive tasks include other types of tasks such as, focused, selective and sustained tasks. However, in this research we have selected the data from focused attention tasks only in this study.

Feature extraction

For the feature selection, our problem can be formalized as follows. Let $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, be a set of training patterns, where x is the supervised selected input feature vector and y—is the random pattern $x, y \in X$ both taken from the nonempty set. Our goal is to find the maximal relevant (predictive) features of y-patterns with x-patterns and assign to the labeled class accordingly. For example, EEG features patterns of predefined ADHD/NORMAL children with randomly recorded EEG patterns. This will result in the more classification accuracy of the used classifier. Our approach for feature selection is based on power spectrum features since the EEG power based features are the most informative and common in use in clinical ADHD diagnosis. Specifically, for the training set, we defined power spectrum density features of EEG each channel obtained by Fast Fourier Transform (FFT) with a *Han* window using 256 data points with 50(%) overlap. The EEG signal length is 2 minute epochs from each channel. The resulting power spectrum density function P(f) is then divided into the four frequency bands: delta (0.5–3.5 Hz), theta (3.5–7.5 Hz), alpha (7.5–13 Hz) and beta (13–30 Hz), for both absolute and relative power, as well as the total power of the EEG (1.5–30 Hz). Ratios were also calculated between frequency bands by dividing the power of the slower frequency band by the power of the faster frequency band. These were calculated for theta/alpha and theta/beta frequencies. According to the studies reported in literature the following Table 1 was extracted. We have selected each frequency band and the goal is to find the most dominant signal features in ADHD children or vice versa. Therefore, our analysis should employ a hybrid method that would extract only most informative and relevant features with minimum redundancy with respect to the known dataset.

In the next section, we introduce the proposed mutual information technique to extract important features of the EEG signal.

Mutual information feature selection

Mutual information measures the statistical dependency between two random variables [20]. The dependency can be related to the information measure of relevance of random variables. Here we apply mutual information method for efficient feature selection.Particularly we aim to build a semi-supervised algorithm based on the information theoretic approach that measures the information exchange, dependency or relevance between EEG patterns.

Shanon's entropy provides a powerful formalization of uncertainty of a random variable. Let X be a random variable and its entropy $H(X)$ is defined as,

$$H(X) \triangleq -\int_X p(x)log(x)dx, \tag{1}$$

whereas the conditional entropy $H(X|Y)$ is defined as

$$\begin{aligned} H(X|Y) &\triangleq -\int_X \int_Y p(x, y)log(x|y)dxdy \\ &= -\int_Y p(y)\left(\int_X p(x|y)logp(x|y)dx\right)dy. \end{aligned} \tag{2}$$

Note that here Y is a discrete binary random variable representing the class labels and X is a particular feature x_j corresponding to a dimention of the input vector. As a result, the conditional entropy $H(X|Y)$ can be expressed as

$$H(x_j|y) = -\sum_{y\in\{0,1\}} p(y)\int_{z\in\text{range}(x_j)} p(z|y)logp(z|y)dz \tag{3}$$

where P_0 and P_1 are the class priors. From equation above we can measure the dependency of two random

Table 1 Selected predictive power features based on literature and the major differences found between ADHD and normal group

Frequency band (Hz)	Delta (0.5–3.5 Hz)	Theta (3.5–7.5 Hz)	Alpha (7.5–13 Hz)	Beta (13–30 Hz)	Theta/alpha	Theta/beta
Associated cognitive states	Deep sleep	Unfocused drowsiness	Eyes closed, alert restfulness	Mental activity, concentration	Good indicator	Good indicator
Findings in ADHD	Increased and reduced delta [4]	Increase in absolute and relative theta [3]	Reduced relative alpha [8]	Lower relative beta [3] & increased or reduced [4]	Barry et al. [3]	Increased [9]

variables X and Y. If X depends on Y, the uncertainty on X is reduced when Y is known and this formalization is obtained through conditional entropy. Then, the mutual information can be both defined in terms of the joint and conditional entropy (see also Fig. 2):

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{4}$$

Note in order to compute the entropy and mutual information we need to evaluate the probability distributions. One can compute the entropy of a random variable in closed form. In particular, for a Gaussian random variable $X\ N(X;\mu,\sigma^2$ the entropy is $H(X) = 1/2(1+\log 2\pi\sigma^2)$. In more general case, even for a mixture of Gaussians, we can no longer compute the entropy in closed form. If the $p(x)$ can be calculated in closed form e.g., if we make a certain assumptions about the parametric form of the $p(x)$ and estimate the parameters, the we can still evaluate $H(X)$ fairly easily, by numerically approximating the integral in equation of the entropy. If no parametric assumption are made regarding the form of $p(x)$ but we have set of observation drawn from it, we can estimate the density at any given value of X using the kernel density estimate as described in class:

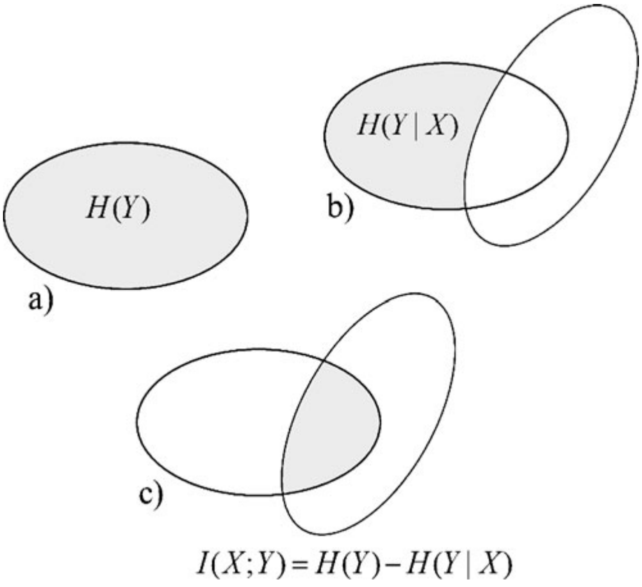

Fig. 2 Mutual information and entropy relationships in Venn Diagram

$$p(x) = 1/N \sum_{i=1}^{N} K(x, x_i) \tag{5}$$

where $K(x, x_i)$ is the kernel function which itself is a valid probability density function and x_i are observations drawn from $P(X)$. In this problem, rather than assuming a parametric form for the class-conditions, we apply a Gaussian kernel density estimator to model the distributions directly from the observations, which is given by,

$$p(x) = \frac{1}{N\sqrt{2\pi}\sigma} \sum_{i=1}^{N} \exp\left(\frac{-(x-x_i)^2}{2\sigma^2}\right) \tag{6}$$

here σ is user defined standard deviation of the Gaussian kernel function in our case $\sigma = 0.4$.

At this point lets recall that the mutual information between random variables is given by $I(X;Y)$ and formulate it to our feature selection purpose. The mutual information measures the reduction in uncertainty on X resulting from the knowledge of Y. The mutual information satisfies the bound $0 \leq I(X,Y) \leq H(X)$. The lower bound is reached if and only if X and Y are independent hence $H(X|Y) = H(X)$. The upper bound is achieved when X and Y are fully dependent, or $P(X|Y) = 1$, $H(X|Y) = 0$ and $I(X,Y) = H(X)$ which means we can get all information of X from Y. Therefore we can consider the mutual information measure as a relevance or dependency measure between two random variables. Moreover, $I(X,Y)$ can be selected as a similarity measure or a likelihood function of two or more random variables, based on the dependency. We therefore intuitively formalize the applicability of entropy and mutual information in our feature selection problem. Let $(x_i, y_i)_{i=1}^{n}$, be a two random variables which can be described with $p(x,y)$ and $H(x,y)$ and where x_i is random EEG patterns and y_i is known features classes for ADHD and NORMAL children. If x and y have relevant patterns then mutual information between them tends to be larger value of $0 \leq I(X, Y) \leq H(X)$ and if they are irrelevant (or independent)then mutual information tend to $I(X, Y) \approx 0$. The mutual information between given random variable and known variable then can be assumed as an estimation of a similarity measure or relevance. Therefore our formulation will be as follows, given $(x_i)_{i=1}^{n}$ a random EEG signal and $(y_i)_{i\in 0,1}$ where $y = 0$ is known NORMAL and $y = 1$ is ADHD signal patterns. The task of MI feature selection is to find the (x_i) which is most relevant to $(y_i)_{i\in 0,1}$ in other words find $k_j = \text{argmax}\{I(x_i, y_i)\}$.

Maximal discrepancy criterion

We introduce another formulation of mutual information in our work where we consider to work on the training dataset (previously defined set). Given dataset x_i with m features and n instances, where $F_{0,1} = (f_1, ..., f_m)$ and $D = (d_1, ...d_n)$ are sets of features and instances, respectively. The F here corresponds to selected features given in Table 1, while D is every instance in a single feature. We have two classes of dataset $(y_i)_{i\in 0,1}$ as mentioned already. It is necessary to find subsets of these two class dataset with maximum discrepancy in order to achieve higher decision accuracy. As noticed earlier maximizing the mutual information maximized the relevance between variables, here we do the reverse of it. We try to extract subset of F with minimum information measure between F_0 and F_1 for maximal discrepancy. The set F_0 can be considered as having two parts F_0' and F_0^* where F_0' is the subset relevant to F_1 and F_0^* subset that is irrelevant to F_1 (non-overlapping part, see Figs. 2c and 3) then $F_0 = F_0' + F_0^*$. We minimize relevant part (F_0') as follows:

$$
\begin{aligned}
I(F_0, F_1) &= H(F_0) + H(F_1) - H(F_0, F_1) \\
&= H\left(F_0', F_0^*\right) + H(F_1) - H\left(F_0', F_0^*, F_1\right) \\
&= H\left(F_0'\right) + H\left(F_0^*\right) + H(F_1) - H\left(F_0'\right) \\
&\quad - H\left(F_0^*, F_1\right) \\
&= H(F_0) + H(F_1) - H(F_0, F_1) \\
&= I\left(F_0^*, F_1\right) \approx 0; \qquad (7)
\end{aligned}
$$

By elimination of F_0^* we minimize dependence of $X \cap Y \approx null$ as the result we reduce the number of feature instances that are most likely to degrade the classifiers accuracy.

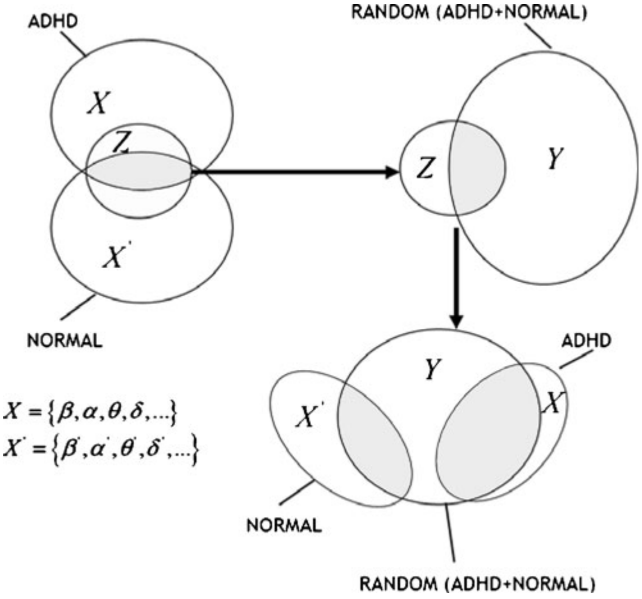
ADHD
RANDOM (ADHD+NORMAL)
X
Z
X'
NORMAL
Z
Y
Y
ADHD
X'
X
$X = \{\beta, \alpha, \theta, \delta, ...\}$
$X' = \{\beta', \alpha', \theta', \delta', ...\}$
NORMAL
RANDOM (ADHD+NORMAL)

Fig. 3 Feature selection problem: we observe two random variables trying to maximize the discrepancy of the variables finding the minimal mutual information measure and in third stage a semisupervised algorithm updates the training set

Semi-supervised feature selection

After selection set of features with maximum discrepancy we further proceed to semi-supervised feature selection. The selected features from training dataset may not have the ideal (marker) informative feature sets of ADHD group. It is usually the case when the size of EEG data number of subjects are limited and in the case when the there is a subject variability constraints. In order to make our approach more adaptive to out of the scope of the available training dataset (EEG data, subjects 10) we propose the following algorithm. Generally, we provided in our training dataset initially defined as $F_{0,1}$ for NORMAL and ADHD case. However,when a new subjects (probably with different symptoms) are tested with the proposed method, the algorithm may fail (or output low accuracy) to analyze whether the subject belong to ADHD or NORMAL group. Then the only thing should be done is to update the training set $F_{0,1}$ by providing new information $(N_{1,0})$ related to both group where $N \in (y_i)_{i\in 0,1}$. When new information is available there's a possibility that subset of N are already in F and in the other case it is completely new feature set (information) which is not included in F. Hence we propose the following formulation which measures the interactions between previously selected training set and the new input information. We have now three feature set (variables) as $(F_0, F_1, N_{0,1})$ and we can ask how we learn about N a random input feature by observing the related class $y_{0,1}$ ADHD/NORMAL when we already have the feature set $F_{0,1}$, then the answer can be explained by conditional mutual information. If the new input feature N is independent of F_1 and y_1 we have $I(y_1, N|F) = I(y_1, F)$. In other words, we can ask how much information N contains about group $y_{0,1}$ after we condition on F:

$$I(y_1, N|F) \equiv H(y_1, F) - H(y_1|N, F) \tag{8}$$

Notice that this is the average of

$$H(y_1, F = f) - H(y_1|N, F = f) \equiv I(y_1, N|F = f) \tag{9}$$

over possible feature values of f. For each f, we have an ordinary mutual information, which is nonnegative, so the average is also non-negative. In fact, $I(y_1, N|F) = 0$ if and only if y and N are conditionally independent given F. Again, the output of $I(y_1, N|F)$ is nonnegative, it can be bigger than, smaller than or equal to $I(y_1|N)$. When it is not equal, we can say there's a interaction (relevance) between F and N—as far as their information about y. It is a positive interaction if $I(y_1, N|F) > I(y_1, |N)$, and negative when the inequality goes the other way. If the interaction is negative ,then we say that some of the information in N about y is redundant given F. We can see how this connects to our feature selection: we will select feature sets containing non-reduandant information about y.

We conclude this section with the following semi-supervised feature selection formulation:

$$n(i) = \text{argmax} I\left(y_{(0,1)}, N_i\right) \tag{10}$$

$$k(i) = \text{argmax} H\left(y_{(0,1)}, F_i\right) - H\left(y_{(0,1)}|N_i, F_i\right) \tag{11}$$

Select those features which satisfies the eqality $I(y_1, N|F) > I(y_1, |N)$, non redundant and relevant with respect to F and y.

Classification

The final step of our research is to use supervised learning method to classify ADHD or Normal group, based on the EEG patterns extracted in the previous sections. In particular we use a Support vector machine (SVM) classifier a learning machine that can be used for classification problems as well as for regression and novelty detection. Important features of SVMs are the absence of local minima, the well-controlled capacity of the solution, and the ability to handle high-dimensional input data efficiently. It is conceptually quite simple, but also very powerful: in its infancy, it has performed well against other popular classifiers and has been applied to problems in several fields. The EEG patterns can be thought of as points in an n-dimensional space. SVM is then trained to classify the data points of several classes. Particularly, SVM chooses the hyperplane that provides maximum margin between the plane surface and the positive and negative points. The separating hyperplane becomes optimal when the distance, from the closest data points, is maximized. These data points are called the support vectors. We briefly review SVM algorithm here for more detailed information, one can find it in other literatures such as [21, 22] or elsewhere. For non-separable case the SVM is constructed by solving a following dual optimization problem,

$$\text{argmax}_{\{\alpha\}} \left(\sum_{i=1}^{N} \alpha_i - 1/2 \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K\left(x_i, x_j\right) \right) \tag{12}$$

subject to

$$\sum_{i=1}^{N} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C \forall i = 1, ...N.$$

where the Lagrangian multiplier are given by $\alpha = (\alpha_1, ..., \alpha_N)^T$,, the training samples are (x_i, y_i) and their respective labels given by $y = (y_1, ..., y_N)$, and C is the penalty parameter for slack variable that should be minimized. In equation above $K(x_i, x_j)$ is the kernel function that is used to embed the training samples into n-dimensional space. The accuracy of SVM classifier also strongly depends on the type of the Kernel function used. For example, there are several available kernel function for non-linear mapping of input patterns. In this research we use radial basis Gaussian kernel function to train and test the SVM. It is given as

$$K\left(x_i, x_j\right) = exp\left(1/2\sigma^2\left(-|x_i - x_j|^2\right)\right) \tag{13}$$

There are free parameters namely *sigma*, of the SVM kernel function and margin-loss trade-off C, which should be determined to find the optimal solution. The objective is to obtain best C and so that the classifier can accurately predict unknown data (testing data).In our case the optimum values of the parameters are obtained with the $5 - fold$ cross-validation using grid search algorithm. Here the data is partitioned into 5 equally sized subsets. Subsequently 5 iterations of training and validation are performed such that within each iteration a different subset of the data is held-out for validation while the remaining 4 subsets are used for learning. We rearranged the data to ensure that each subset is good representative of the whole data. In other words, for a classification of ADHD and Normal each subset contains 50% of feature patterns where class comprises around half the instances.

After finding the optimal parameters, the classification performance of SVMs was measured using standard criteria as follows: where a true positive (TP) outcome was registered when both SVM and physicians classified a EEG pattern as ADHD. False positive (FP), true negative (TP) and false negative (FN) outcomes were similarly defined [Ref]. Using these definitions, True Positive Rate ($TPrate$) and False Positive Rates ($FPrate$) were calculated using formulas below,

1. True positive rate (also referred to as sensitivity) is the percentage of positive examples which are correctly classified

$$TPrate = \frac{TP}{TP + FN} \tag{14}$$

2. False positive rate—is the percentage of negative examples which are misclassified

$$FPrate = \frac{FP}{TN + FP} \tag{15}$$

We analyzed the classification performance using a Receiver Operating Characteristic (ROC) curves, which plots as sensitivity (true positive rate) versus one minus specificity (false positive rate). The area under the ROC curve (AUC) as a measure of the discriminatory power of a classifier, which is insensitive to class distributions and the costs of misclassifications; for instance AUC = 1 indicates perfect classification, while AUC = 0.5 means that the classifier does not perform better then random guessing.

Experimental results

For the feature selection we have tried various combinations of feature sets mainly given in Table 1 and compared the robustness of each feature type for our problem. As stated earlier our feature selection steps are given by: (1) finding maximal mutual information (see Section "Mutual information feature selection"); (2) maximizing discrepancy of two group (Section "Maximal discrepancy criterion") and finally (3) semi-supervised feature selection (in Section "Semi-supervised feature selection"). The goal

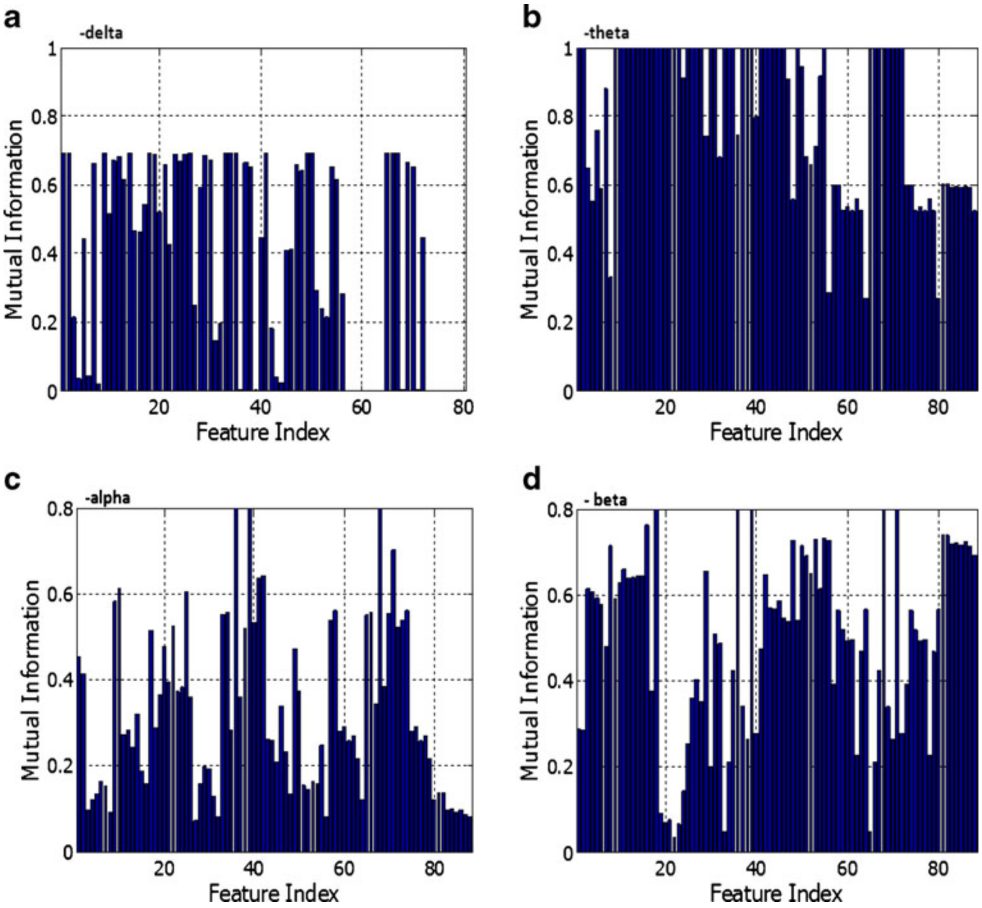

Fig. 4 Comparison of mutual information measure of the whole data from all participants for the following specific frequency bands, **a** delta, **b** theta, **c** alpha and **d** beta. Theta and beta frequency bands are found to provide more predicting features for ADHD or normal

is to find the most predictive or the most distingushing and robust features that is used to recognize ADHD group. We organized the data of 10 children from all 8 channels and computed its respective power spectrum densities. Further the power spectrum were divided to specific frequency band. We applied proposed mutual information to the each frequency band to explore how good predictor is each band for all children. Figure 4 demonstrates the obtained results and the comparative view of the *Delta*, *Theta*, *Alpha* and *Beta* frequency bands. It was noticed that *Theta* features were more consistent among all children (see Fig. 4b), which means the more mutual information the more better feature for the specific problem. In contrast, *Delta* and *Alpha* frequency bands are found to be less informative than the remaining bands. One can notice that *Delta* features show absence in some children which achieves the minimum of mutual information value. Similarly *Alpha* features have less mutual information measure in most cases. However, in average *Theta* feature sets are good indicator of both class ADHD and NORMAL.

In our next step, we demonstrate the comparative results of power ratio features for each children. As shown in Fig. 5, we have evaluated 4-types of power ratios such as $Theta/alpha$, $Theta/Beta$ and $Relative - Theta$ finally $Relative - Beta$. Many studies have reported that power ratio features are the best indicators of ADHD by finding variability of a selected power ratios. In our study, we don't analyze the variability but to explore the most robust features for both groups. First, it has been found that $Relative - Theta$ provide maximal mutual information in average among all children, compared to other ratio features (see, Fig. 5a). Second, $Relative - Beta$ features as shown in (b) plane of Fig. 5 were slightly better than $Theta/Beta$ but in both cases some children has shown the less features in these ratios. The information measure in last feature set, namely $Theta/Alpha$ is less relevant but common in all children. Here, feature index represents the number of children while in Fig. 4 it represents the number of features in rows (8 channels × 10 children).

Significant findings were obtained when evaluating three different approach in feature selection. For example, Fig. 6 shows the results of selected features when using mutual information of power features, the power ratios when with maximal discrepancy features and semi-supervisedly selected features of the same power

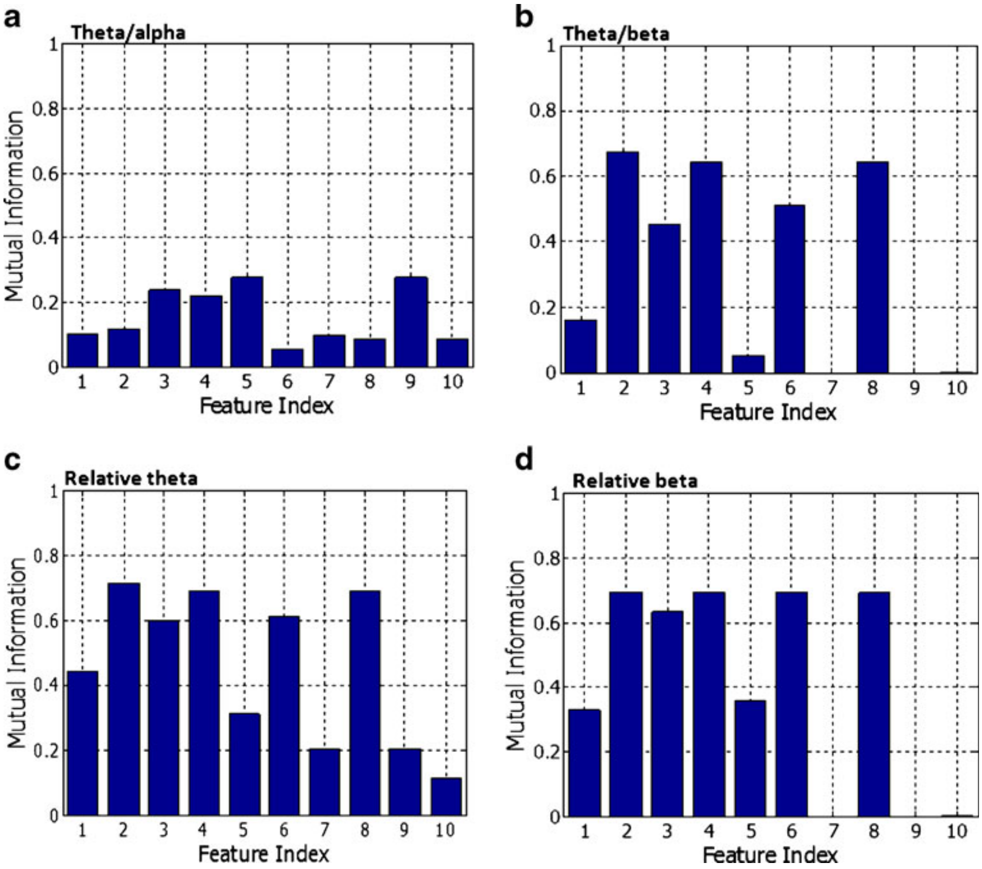

Fig. 5 Mutual information measures of power ratios **a** theta/alpha, **b** theta/beta, **c** relative theta and **d** relative beta. Most of the children showed similar distinguishing features relative theta and relative beta

features. The power features consists of all combination of the features and the results show us the best candidate feature among the selected features. Particularly, as shown in (b) plot of Fig. 6, the features consists of following sets $[\alpha, \beta, \theta, \theta/\alpha, \theta/\beta, relative\theta, relative\beta]$. The results are actually different from the ones which we analyzed in Figs. 4, and 5. Here, for example in (a) plane one can notice the minimum value of *delta* and *theta* in contrast to the results when they were found to be good features in in Fig. 4. It is due to the fact that mutual information changes as the number information about the feature increases, because the mutual information measures interaction information between random variable. Then from Fig. 6a we can conclude when all information are considered the most relevant feature sets to both group are ratio coefficients such as $\{\theta/beta, Rel.\theta, Rel.\beta\}$. Subsequently the same feature sets are selected by with maximal discrepancy criterion between each feature subset. The output is shown in Fig. 6b, in this case it can be noticed that appearance of $\{\beta, \delta, \theta\}$ feature sets. However the most predicting ones are found as the power ratios including the β feature set when using this criterion. The final case demonstrates more promising approach by semi-supervised feature selection. In this step, we analyze the same data with maximal discrepancy criterion and randomly provide various subsets of the features. Then, as explained earlier, the semi-supervised algorithm updates the feature set based on the available information. In addition to the dataset of 10 children we provided other data collection which were obtained from the same children but from different EEG recording sessions. It is noticable that in this feature selection method the relative power ratios remain stable in other words there's no update in the subset. However, the other specific bands also increase in mutual information measure. One thing to notice here is that this result have some information relation with the results obtained in Fig. 4 earlier where we assumed that good predictive features are *theta*, *beta* followed by *delta*. Compared to (b) plane of Fig. 6 we exactly notice the update of feature set in these frequency bands. However, the question is why *alpha* set increases (or updates)?; the cognitive state associated with alertness which is usually less in ADHD children. One of the answer would be the reference [8] where authors reported the difference in *alpha* state between ADHD and normal group. In addition our goal is not to analyze the difference, increase or decrease of powers instead we are trying to find the most distinguishing features of ADHD from normal children (or viceversa) for the classification purposes.

Subsequently, the performance of the proposed classification method is validated for selected data feature set. The dataset contains a variety of feature sets

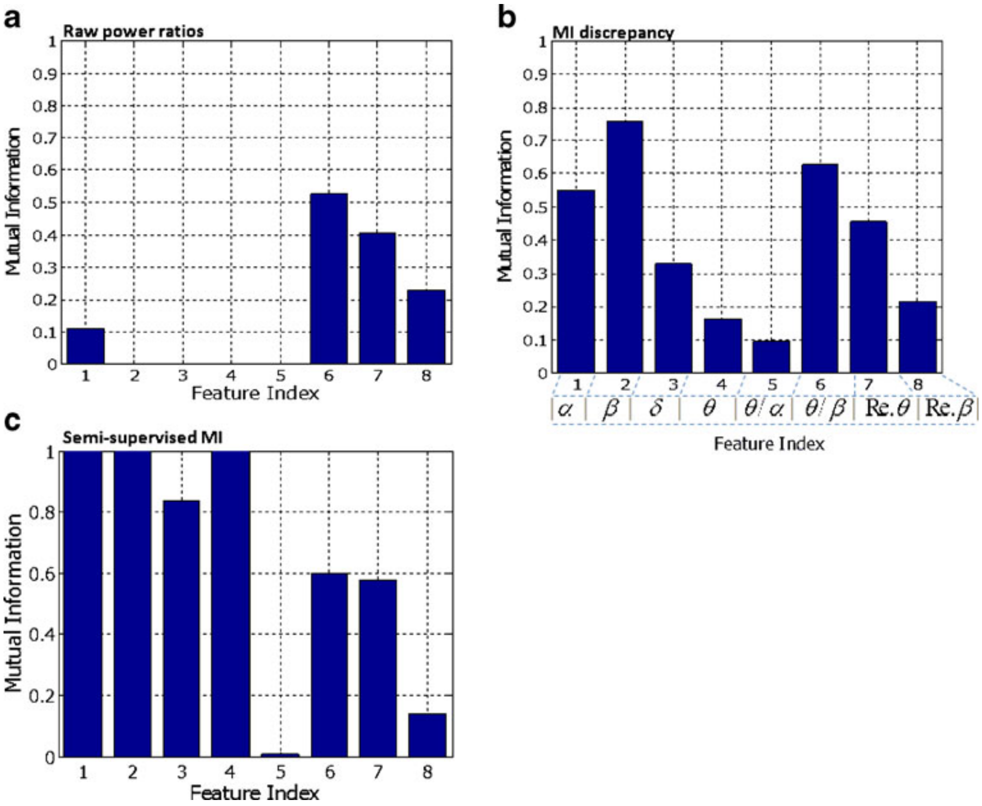

Fig. 6 Overall findings and comparison of the most predictive power among 8-analyzed features (see Table 1) and relative powers. Three types of feature selection methods are used and compared where **a** raw power features, **b** maximal discrepancy criterion based, and **c** semi-supervised selected features

Table 2 Types of input features provided to the SVM classifier

Input features	No. of features	
	ADHD	NORMAL
Raw power spectrum	223,560	3,160
Maximal discrepancy	120,246	4,620
Semi-supervised	902,462	11,354
Theta/alpha	2,530	365
Theta/beta	3,105	201
Relative powers	5,462	232

(considered as predictive powers). We tested the SVM on various combinations of input features (training set) which are organized as follows (also shown in Table 2); (1) Raw power spectrum feature sets of each class (ADHD or not); (2) Mutual information based feature maximum discrepancy feature sets; (3) Semi-supervised selected feature sets; (4) Power ratios and; (5) Relative powers. The performance of SVM were obtained considering all frequency features and their combinations.

Our first step is the exhaustive search of optimal parameter of the classifier, particularly C and σ. For each types of the feature set we performed a grid-search on C and σ using cross-validation. Various pairs of $(C;\sigma)$ values are tried and the one with the best cross-validation accuracy is selected. We noticed that trying growing sequences of C and sigma is a practical method to identify good parameters. Experimental results show that the minimum test error rate for the classification is acquired with the following parameters given in Table 3.

One can analyze that the training errors rates are minimized in some parameters but they vary. For instance the difference in the test error among selected four features can be seen. Especially, if you notice when using semi-supervised features that performance of the classifier achieves minimum error rate for the test data. Conversely raw power features provide less accuracy or higher error rate while trying to classify the two groups. It was also found that *Theta/alpha* ratios demonstrate more stable features however, the accuracy is degraded. Though our approach we achieved best minimum error rate of (5.2%) when using a semi-supervised algorithm.

Figure 7 demonstrates the overall classification results in terms of cross validation error, minimum cross validation error and most importantly minimum test error. However, using maximal discrepancy criterion the minimum test error rate (5.2%) was relatively good compared to using raw power features. Specific features sets such as *theta/alpha* and *theta/beta* ratios provided in average 8% of test error while the minimum training error was less 8%. Many studies report that *relative theta* ratios are best marker of the ADHD children. In this study we obtained best cross validation error using the *raw relative theta* features with 4.6%, however the minimum training error was comparatively less than the one obtained through other features. Since our goal is to find such a features that minimizes the cross validation error along with equally test error. In our next step we demonstrate the classification performance under the ROC curve for various types of input features to SVM.

Figure 8 shows the ROC curves of the SVM classifier together with the area under the ROC curve (AUC) measure. For the optimal classification we should obtain maximal *TPrate* and minimum *FPrate* that occurs at various instances of the threshold. Moreover classification power of a classifier is measured by the area under the ROC curve; AUC = 1 indicates perfect classification, while AUC = 0.5 means that the classifier does not perform better then random classifier. Figure 8a shows the SVM peformance for the test set with two types of feature selection method as well as raw sets consisting of power spectral features sets. We note that maximal classification performance obtained when using semi-supervisedly selected features AUC = 0.954. Since it is easy to see that the curve here stretches almost into the top left corner that represents the

Table 3 Cross validation results with the optimal parameters of SVM for 5—types of feature data set

Input features	SVM parameter	Kernel function	Min train error	Test error
	C,	σ		
Raw power spectrum	$C = 8.20$	$\sigma = 0.2727$	0.1096	0.1163
	$C = 10.0$	$\sigma = 0.1194$	0.1194	0.1163
Maximal discrepancy	$C = 0.10$	$\sigma = 0.4010$	0.4233	0.5104
	$C = 0.63$	$\sigma = 0.6589$	0.1656	0.5768
Semi-supervised	$C = 4.98$	$\sigma = 2.8070$	0.0532	0.05104
	$C = 10.9$	$\sigma = 0.8369$	0.0459	0.5268
Theta/alpha	$C = 6.72$	$\sigma = 0.3340$	0.0813	0.0856
	$C = 100$	$\sigma = 0.3340$	0.4452	0.8063
Relative ratio	$C = 55.0$	$\sigma = 0.4503$	0.05785	0.0747
	$C = 70.0$	$\sigma = 0.4503$	0.03254	0.0456

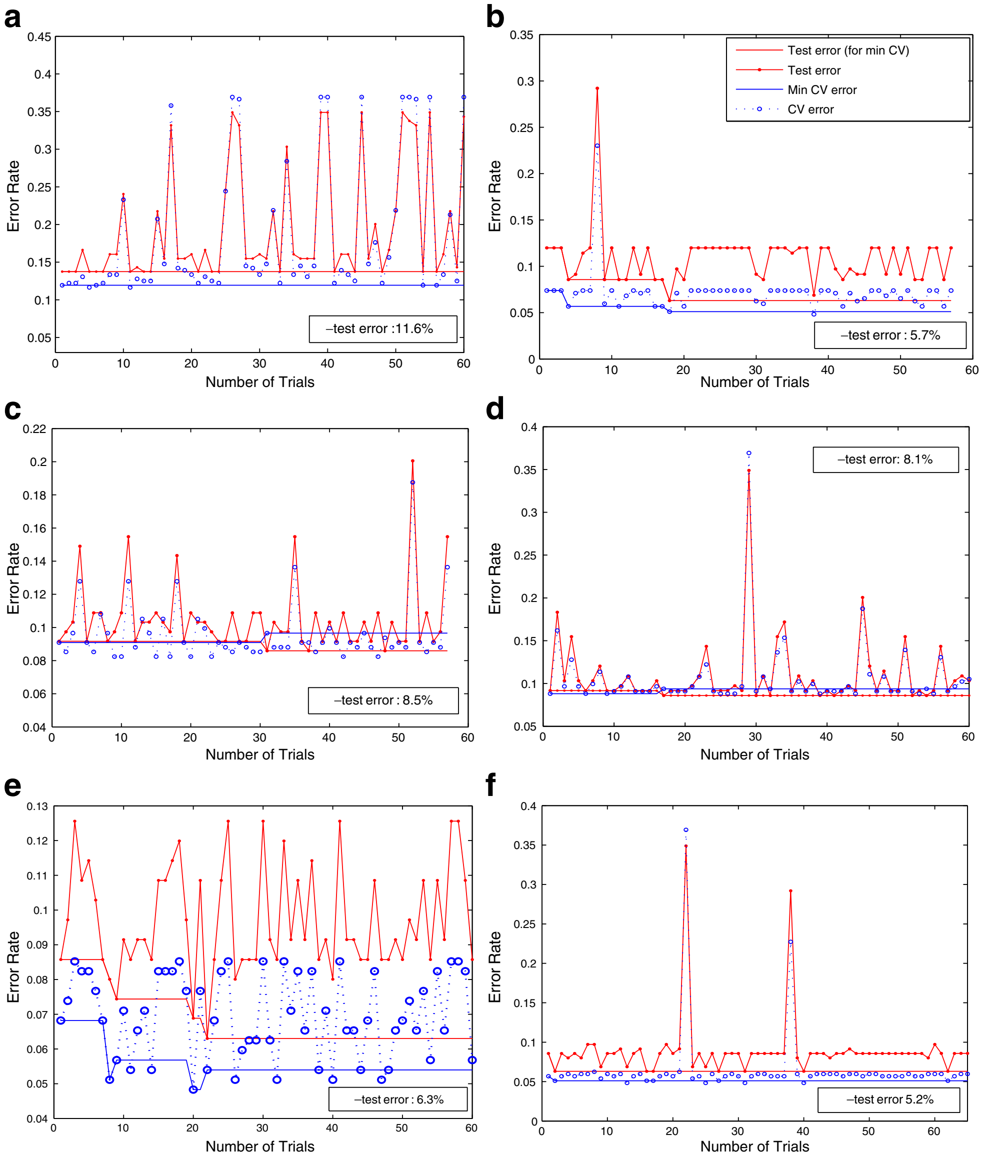

Fig. 7 SVM cross validation results using four types of EEG input features. **a** Raw power spectral features, **b** maximal discrepancy, **c** theta/alpha ratios and **d** theta/beta ratios, **e** relative theta, and **f** semi-supervisedly selected features consisting of all power spectral and ratio features

performance of the superior model (this model commits virtually no false positives). One can compare the TPrate in (x=0.13,y=1) achieves 100% classification but with the tradeoff 13% of the times FPrates. The ideal case would be to obtain (x=0,y=1) for perfect classification. When using raw features without any feature selection method we noted the classification performance is equal to AUC = 0.7 which is better than any random classifier. Lastly, maximal discrepancy feature selection approach results superior in AUC = 0.90 compared to selecting raw features.The remaining parts of the figure analyzes the power ratio input features and the effects of the classifier performance. For instance, Fig. 8b when using only power ratio features to the SVM one can find better features by analyzing the ROC curve and AUC. In this case, *Relative beta* found to be more robust in discrimination between ADHD and Normal than other features plotted in the figure. Simply, if a classifier has greater area (AUC value) it has a better average performance. SVM with *Theta/alpha* features becomes inferior to the remaining input features. Going further to the Fig. 8c results in more liberal performance by using maximally discrepant features. In this case we noted that all of the curves are virtually identical illustrating no big discrepancy in classifier performance. However, the maximum AUC value is obtained using again using *Relative beta* input features as in the previous Fig. 8b. Finally in Fig. 8d we can observe the classifier performance when using semi-supervisedly selected features. It's interesting to observe that the performance better given power ratio features where the maximal AUC = 0.97 achieved

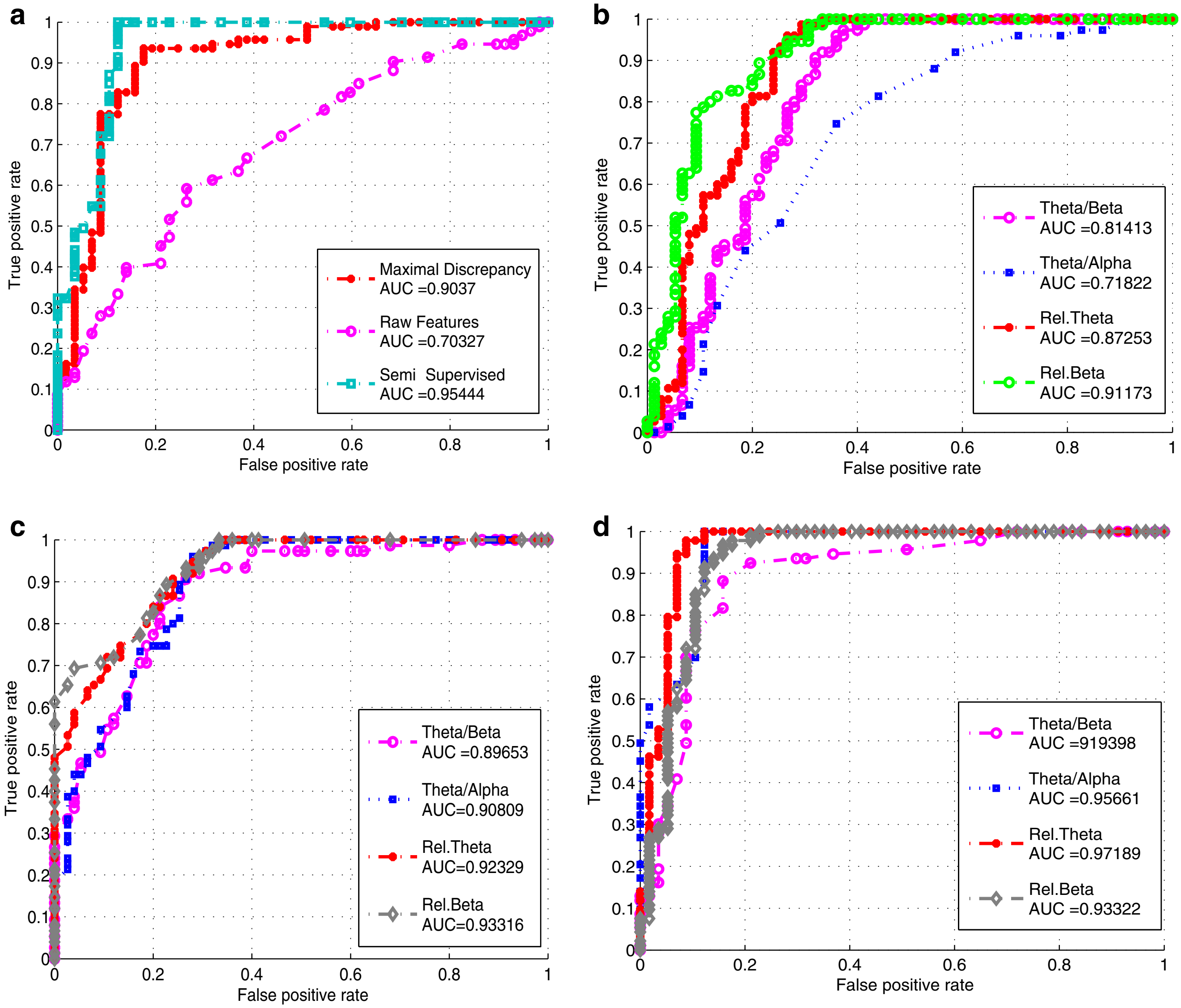

Fig. 8 ROC curve of the SVM performance when the input instances consisted of different combinations of EEG features such as: **a** power spectral features using three methods, **b** raw power ratio features, **c** power ratio features obtained through maximal discrepancy criterion, **d** power ratio features using semi-supervised method

using relative theta. This means that the classification rate is almost 97% times true. Besides, we see that the SVM only begins to commit no false positive errors after it has almost reached a true positive rate of 30–50%. The curve is in a near perfect classification model however this will not happen until the curve has almost reached the ‘perfect performancepoint. Compared to other methods (Fig. 8a, b, c), the model in Fig. 8d is not only superior because its curve is the closest to the ‘perfect performance curve but we can observe that for large ranges of the ranking values the model commits more true positive rates than it provides false positive classifications. We tried to demonstrated a classifier perfomance assessment depending on types of the input featues, that is, given two or more input features, we need to pick one in order to be deployed. The following criterion to validate one classifier with random features sets over the other(s) is considered: (a) our classifier should be general enough to describe its performance over a broad range of possible feature set and (b) it should be able to discern whether the performance difference between input features are statistically significant. It turns out that ROC curve analysis provides the validation for both of these criteria in a highly visual manner.

Conclusion

We demonstrated a decision support system in diagnosis of ADHD disorder. The contributions of the current study include developing an adaptive feature selection method in particular, a semi-supervised feature selection algorithm which is based on mutual information theory. We have shown the potential use of the algorithm is when the input available data is limited and one can define new features, based on which the algorithm updates its training set features. Compared to other studies that use a learning algorithm, this study is the first step in formulating the semi-supervised approach for ADHD feature selection. Through this approach it was possible to define new features of EEG signal which provide important information in discriminating between ADHD and normal group. The maximal accuracy of the SVM classifier was above 97% when when using a semi-supervisedly selected features. Particularly, we found that power ratio features are dominant features of EEG signal for ADHD, though other features provided relatively good accuracies. We believe the current method would assist the physicians in diagnostic process of ADHD disorder as well as lessen their workload. In our further work, we plan to increase the computational efficacy of the algorithm and test it on large numbers of datasets.

Acknowledgements We thank our colleagues who provided assistance in data acquisition sessions. Especially, we appreciate the efforts of Dr. Yunhee Shin and her research group during the research collaboration (Daegu Univ., ROK) and their fruitful discussions about the ADHD children. We are also immensely grateful to reviewers for their useful comments on an earlier version of the manuscript. This work was supported by the Daegu Gyeongbuk Institute of Science and Technology and funded by the Ministry of Education, Science and Technology of the Republic of Korea.

References

1. Biederman, J., and Faraone, S. V., Attention-deficit hyperactivity disorder. *Lancet* 366:237–248, 2005.
2. Castellanos, F. X., Toward a pathophysiology of attention-deficit/hyperactivity disorder. *Clin. Pediatr.* 36(7):381–393, 1997.
3. Barry, R. J., Clarke, A. R., and Johnstone, S. J., A review of electrophysiology inattention-deficit/hyperactivity disorder: I. Qualitative and quantitative electroencephalography. *Clin. Neurophysiol.* 114:171–183, 2003.
4. Loo, S. K., and Barkley, R. A., Clinical utility of EEG in attention deficit hyperactivity disorder. *Appl. Neuropsychol.* 12(2):64–76, 2005.
5. Barry, R. J., Johnstone, S. J., and Clarke, A. R., A review of electrophysiology in attentiondeficit/hyperactivity disorder: II. Event-related potentials. *Clin. Neurophysiol.* 114(2):184–198, 2003.
6. Barkley, R. A., Attention deficit hyperactivity disorder: A handbook for diagnosis and treatment, 2nd edn. Guilford Press, 1999.
7. Elder, T. E., The importance of relative standards in ADHD diagnoses: Evidence based on exact birth dates *J. Health Econ.* 29:642–656, 2010.
8. Barry, R., Kirkaikul, S., and Hodder, D., EEG alpha activity and the ERP to target stimuli in an auditory oddball paradigm. *Int. J. Psychophysiol.* 39:39–50, 2000.
9. Chabot, R. J., and Serfontein, G., Quantitative electroencephalographic profiles of children with attention deficit disorder. *Biol. Psychiatry* 40:951–963, 1996.
10. Clarke, A. R., Barry, R. J., McCarthy, R., and Selikowitz, M., EEG analysis in attention-deficit/hyperactivity disorder: A comparative study of two subtypes. *Psychiatry Res.*, 81:19–29, 1998.
11. Clarke, A. R., Barry, R. J. McCarthy, R., and Selikowitz, M., EEG-defined subtypes of children with attention-deficit/hyperactivity disorder. *Clin. Neurophysiol.*, 112:2098–2105, 2001.
12. Lazzaro, I., Gordon, E., Whitmont, S., Plahn,M., Li, W., Clarke, S., et al., Quantified EEG activity in adolescent attention deficit hyperactivity disorder. *Clin. Electroencephalogr.* 29:37–42, 1998.
13. Mann, C. A., Lubar, J. F., Zimmerman, A. W., Miller, C. A., and Muenchen, R. A., Quantitative analysis of EEG in boys with attention-deficit-hyperactivity disorder: Controlled study with clinical implications. *Pediatr. Neurol.*, 8(1):30–36, 1992.
14. Mueller, A., Candrian, G., Kroptov, J. D., Ponomarev, V., and Baschera, G. M., Classification of ADHD patients using a machine learning system. *Nonlinear Biomed. Phys.*, 4(1): 1–12, 2010.
15. Anduradha, J., Tisha, Ramachandran, V., Arulalan, K. V., and Tripathy, B. K., Diagnosis of ADHD using SVM algorithm. In: *COMPUTE2010, ACM Proceedings*. Bangalore, India, January 22–23, 2010.
16. Ahmadlou, M., and Adeli, H., Wavelet-synchronization methodology: A new approach for EEG-based diagnosis of ADHD. *Clin. EEG Nurosci.* 41(1):1–10, 2010.
17. American Psychiatric Association, American Psychiatric Association: Diagnostic and statistical manual of mental disorders DSM-IV., 4th edn., 1994.
18. Oh, K., Hong, K. M., Lee, H., and Ha, E., *K-CBCL*. Seoul, Korea: Chung Ang Aptitude Publishing Co., 1997.
19. Anastopoulos, A. D., Spisto, M., and Maher, M. C., The WISC-III freedom from distractibility factor: Its utiliity in identifying children with attention deficit hyperactivity disorder. *Psychol. Assess.* 6:368–371, 1994.
20. Cover, T. M., and Thomas, J. A., *Elements of information theory*. Wiley: Chichester, 2006.
21. Cristianini, N., and Shawe-Taylor, J., *Support vector machines and other kernel based learning methods*. Cambridge: Cambridge University Press, 2000.
22. Cortes, C., and Vapnik, V. N., Support-vector networks. *Mach. Learn.* 20:273–297, 1995.