



Classification of ADHD and non-ADHD subjects using a universal background model

Juan Lopez Marcano^a, Martha Ann Bell^b, A.A. (Louis) Beex^{a,*}

^a DSPRL-Wireless@VT-Electrical & Computer Engineering, Virginia Tech, Blacksburg, VA, 24060, United States

^b Psychology, Virginia Tech, Blacksburg VA 24060, United States

ARTICLE INFO

Article history:

Received 22 September 2016

Received in revised form 27 May 2017

Accepted 20 July 2017

Available online 30 August 2017

Keywords:

EEG

ADHD

Gaussian mixture models

Universal background model

AR models

ABSTRACT

ADHD affects a major portion of our children, predominantly boys. Upon diagnosis treatment can be offered that is usually quite effective. Diagnosis is generally based on subjective observation and interview. As a result, an objective test for the detection or presence of ADHD is considered very desirable.

Based on EEG, across multiple channels, using autoregressive model parameters as features, ADHD detection is approached here in analogy with the imposter problem known from speaker verification. Gaussian mixture models are used to define ADHD and universal background models so that a likelihood ratio detector can be designed. The efficacy of this approach is reflected in the traditional detector performance measures of the area-under-the-curve and equal-error-probability. The results – based on a limited database of males, approximately 6 years of age – indicate that high probability of detection and low equal error rate can be achieved simultaneously with the proposed approach, when using EEG collected during an attention network task. The effect of using contaminated data is investigated as well.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

In the US, ADHD is a condition that affects approximately 9.5% of children ages 4–17 [1]. Diagnosis of ADHD is done by using the Diagnostic and Statistical Manual of Mental Disorders (DSM), published by the American Psychiatric Association (APA) [2], which provides a list of symptoms that behavioral scientists use to determine whether or not a subject has a mental disorder. While DSM-V (2013) recognizes three different subtypes or presentations of ADHD, the data available for this effort provided the binary labels of Non-ADHD (NA) and ADHD (A) only.

Since diagnosis is done through subjective observations made by teachers, parents, and behavioral scientists, finding quantitative techniques to aid the diagnosis of ADHD has gained attention. In fact, classification of ADHD (A) and Non-ADHD (NA) has been done relatively successfully [3–8], which implies that A and NA subjects are separable to some extent in several feature domains.

This study concerns the use of a Gaussian-Mixture-Model-based universal background model (UBM) for the classification of A and NA subjects, consisting of 6-year old males. UBMs have been used in the past for speaker verification and identification, and have

achieved high levels of accuracy under different noise conditions [9,10]. Moreover, GMMs and UBMs have recently been studied for the detection and classification of EEG patterns [11,12].

The hypothesis addressed here is that a UBM can potentially address the shortcomings of other classification schemes. Over the last 30 years, the A/NA classification problem has been tackled by extracting features from EEG data when the subjects are resting with their eyes closed or performing some activity. However, when test subjects do not perform the activity they are instructed to perform, classification accuracy is more likely to be poor (perhaps even resembling guessing). Therefore, a UBM built using a large number of feature vectors, extracted from several activities, may make classification more robust.

To the best of the authors' knowledge, this is the first time a GMM-UBM is used for the classification of ADHD (A) and Non-ADHD (NA) subjects. In this study, the features evaluated are AR parameters, which were extracted from time intervals where A subjects and NA subjects were resting or performing an attention network task (ANT). UBMs were trained using a training dataset associated with 2 A subjects and 2 NA subjects, and then tested with a dataset associated with 1 NA subject and 2 A subjects that were not part of the training dataset. Performance was analyzed in terms of Receiver Operating Characteristics (ROC) and shown to vary depending on how much of the training and testing dataset came from ANT. When all the training and testing feature vectors

* Corresponding author.

E-mail address: beex@vt.edu (A.A. Beex).

originate from ANT activity (100% ANT, 0% resting EEG), a mean AUC (area under curve, for ROC) of 0.97 was obtained, with an EER (equal error rate, $P\{A/NA\} = P\{NA/A\}$) of 0.082. As resting data is added to the UBM and ADHD models, performance decreases, resulting in a mean AUC of 0.73 and a mean EER of 0.32 when 50% of the training and testing feature vectors come from ANT activity and the other 50% come from resting EEG.

The structure of this paper is as follows: Section 2 provides an overview of how EEG has been used for the discrimination between ADHD and Non-ADHD subjects. In Section 3, the methods used are described. Section 4 covers the experiments done as well as the corresponding results. Lastly, Section 5 provides the conclusions.

2. Related work

Since 1999, advances have been made towards quantitatively finding differences between ADHD subjects and Non-ADHD subjects during baseline eyes closed activity. In 1999, a study reported that the θ/β power ratio of ADHD subjects tends to be higher than that of Non-ADHD subjects [7]. In that study, the power ratio was obtained by computing the PSD estimates from the FFT. For classification, the θ/β power ratio of Non-ADHD subjects was averaged, and power ratios that were more than 1.5 standard deviations above the average θ/β power ratio of control subjects (the threshold) were classified as associated with ADHD subjects, whereas those that fell below the threshold were classified as Non-ADHD. This simple decision rule was reported to yield 98% of sensitivity. However, another study replicating the methodology of the latter study, reported 84% accuracy (sensitivity + specificity divided by 2).

Although a method that achieves 84% accuracy may not be accurate enough for diagnosis, it could be used for pre-screening. A study found that parents and teachers can detect ADHD with an accuracy ranging from 54% to 63%, which is equivalent to guessing, whereas the θ/β power ratio has been claimed to achieve 84% to 97% accuracy of classification [13]. The latter claim has been seriously called into question recently [14].

Another study [3] used power in frequency bands along with semi-supervised learning during eyes closed activity in order to diagnose ADHD subjects. In this study, the power and power ratios in the α , β , θ , and γ frequency bands were computed and the mutual information criterion was used to choose the least redundant features for training of a Gaussian support vector machine (SVM). The accuracy of classification was 97%.

In our earlier publication [4], AR parameters, extracted from attention activity, and supervised learning were used for the classification of ADHD and Non-ADHD subjects. AR(7) models were computed from windows of 2 s, and a KNN classification accuracy between 85% and 95% was obtained. In addition, a confidence metric was derived from the vote count of the KNN classifier, which ranged from 91% to 99%.

The effectiveness of event-related potentials (ERPs) has also been studied [5]; 74 control and 74 ADHD subjects performed a visual two-stimulus GO/NOGO task while their EEG data was recorded. Independent component analysis (ICA) was performed on the ERPs, and these features were used to train a SVM classifier, which achieved 92% accuracy of classification (90% sensitivity and 94% specificity).

UBMs have been studied for the purpose of classification and person verification. In a recent study, UBMs based on a multi-sphere support vector data description (MSSVDD) and based on GMMs were used to classify control subjects and alcoholic subjects [11]. The features extracted in this study were 12 power components in the 8–30 Hz frequency band and AR(21) coefficients. The EER of the GMM-UBM was found to be 0.221 and that of the MSSVDD was found to be approximately 0.1.

For EEG task classification, popular algorithms and frameworks involve Hidden Markov Models (HMM), since a task can be modeled as a sequence of mental states [15]. In fact, a study used HMM for mental task classification and modeled EEG as a chaotic signal. The models were tested using multiple datasets, and accuracy reported of approximately 72% for the worst case.

With the rise in popularity of deep learning, deep neural networks have been developed to detect patterns in EEG. With a training dataset of 50,900 feature vectors and a testing dataset of 500,000 feature vectors, a deep belief network (DBN) was developed for EEG anomaly detection [16]. The DBN was compared to SVM for the same task, and according to the F1 scores, DBNs slightly outperformed SVMs (0.475 vs 0.439).

Although some deep learning networks, such as recurrent neural networks (RNN), DBN, or long-short-term memories (LSTM) hold promise for the classification of A and NA, the dataset used for this paper is not large enough to be used for deep learning. To the best of our knowledge, there are large EEG datasets available online, such as [17], but not for ADHD.

3. Methods

An overview is provided in Section 3A of how the data were collected, in Section 3B of channel reduction, in Section 3C of details for AR modeling, and in Section 3D of GMM-UBMs.

A Data collection

Children between the ages of 6 and 8 years visited the research lab as part of an ongoing longitudinal study focused on frontal lobe development from infancy through childhood. Information regarding diagnosis of ADHD was obtained via maternal report. EEG was recorded using a stretch cap (Electro-Cap, Inc Eaton, OH: E1-series cap) in the extended 10/20 system pattern. Recordings were made from 26 electrodes located equidistant across the scalp.

Electrode impedances were kept under 20k ohms. The electrical activity from each lead was amplified using separate bioamps (James Long Company, Caroga Lake, NY). During data collection, the high-pass filter was a single pole RC filter with a 0.1 Hz cut-off (3 dB or half-power point) and 6 dB/octave roll-off. The low-pass filter was a two-pole Butterworth type with a 100-Hz cut-off (3 dB or half-power point) and 12 dB/octave roll-off. The EEG signal was digitized at 512 samples per second for each channel so that data were not affected by aliasing. The acquisition software was Snapshot-Snapstream (HEM Data Corp, Southfield MI). Prior to the recording of each subject, a 10 Hz, 50 μ V peak-to-peak sine wave was input through each amplifier and digitized for 30 s. This signal was analyzed and the resulting power values used to calibrate the EEGs.

After the EEG electrodes were applied, children participated in eyes open, eyes closed, and quiet VIDEO baseline events to collect resting EEG data. Then the children completed a battery of cognitive tasks designed to assess various aspects of attention [18] using the child version [19] of the Attention Network Task (ANT) and various aspects of cognition associated with executive functions (e.g., number Stroop, Dimensional Change Card Sort Task, Digit Span Task). Data from the ANT were used in the analyses that are the focus of this report.

The ANT was designed to assess Posner's brain-based attention networks [18] and yields measures of conflict, alerting, and orienting. The test requires the child to respond to a central target (a yellow fish on a light blue background) displayed on a computer screen and indicate whether the fish is facing left or right. The child is instructed to look at the fixation point, above or below which the target will appear. The target may appear with or without flankers (other fish), which may or may not be congruent with respect to

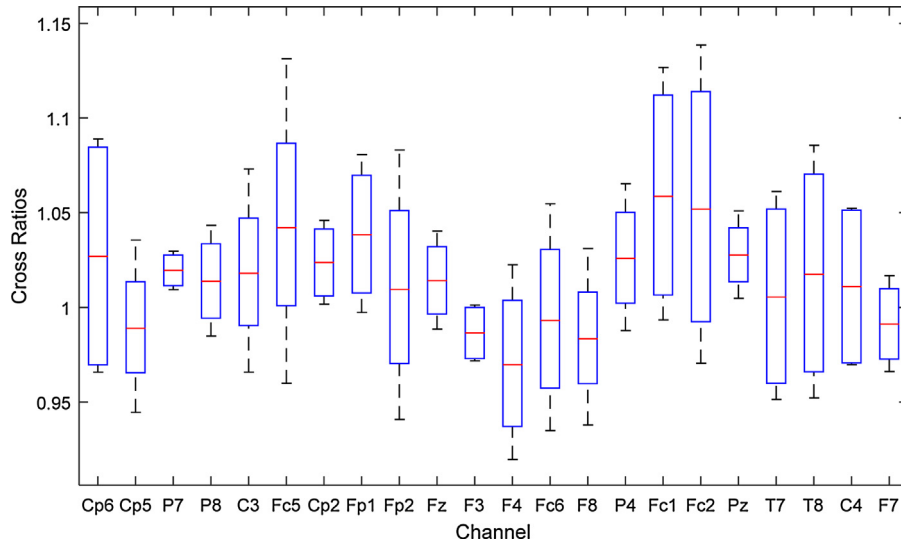


Fig. 1. Cross ratios for all channels.

direction they are facing. Reaction time responses to the alert cues, spatial cues, and flankers are manipulated to provide an assessment of the efficiency of each of the attention networks. The ANT is divided into 3 blocks of ~5 min each, with a brief rest period between blocks. The EEG during the first block and second block were used in these analyses.

After the research visit, EEG data were analyzed using EEG Analysis software developed by the James Long Company. Data were re-referenced via software to an average reference configuration and then analyzed with a discrete Fourier transform (DFT) using a Hanning window of 1 s width and 50% overlap. Power values were computed at each electrode site for theta (4–7 Hz) and beta (13–30 Hz) frequency bands. Power was expressed as mean square microvolts.

The EEG data available for this study was for 7 subjects, 4 with ADHD diagnosis and on medication, and 3 not diagnosed as ADHD. All 7 subjects were about 6 years of age and all were males. No ADHD subtype information was available for these subjects. However, statistically speaking the inattentive/hyperactive or combined subtype is most prevalent in boys, which could mean that our ADHD subjects form a relatively homogeneous group.

B Channel reduction

To reduce the number of channels to be used for the analyses described herein, the aim was to determine five channels that probably will provide good discrimination. Previous research indicates that resting state eyes-opened and eyes-closed theta/beta power ratios (TBPR) tend to be higher for ADHD subjects than for Non-ADHD subjects [7,8]. The preliminary step of channel reduction is therefore executed based on TBPR; however TBPR were evaluated here during ANT activity for all recorded channels, for all subjects, i.e. not during resting state. The next step was computation of all cross ratios, defined as the ratio of TBPR-ADHD over TBPR-Non-ADHD. Fig. 1 reflects the distribution of cross-ratios for each of the EEG channels, in the form of boxplots.

The lines in the middle of the boxes represent the means, while the top and bottom sides of the boxes represent the 75th and 25th percentiles respectively, with the upper and lower horizontal lines representing the maximum and minimum values respectively. The cross ratios fall between 0.9 and 1.15. Based on this preliminary analysis step, the channels chosen to proceed with are Fc2, Fc1, Fc5, Cp6, and C3. This choice does not necessarily mean that these

five channels produce the very best possible discrimination; after all, the performance of various methods is yet to be analyzed, and the results of such analysis may indicate that optimization of the channel choice needs refinement when targeting a specific application.

C Feature extraction

The features used are the autoregressive (AR) parameters a_k , extracted from finite length observation records for the selected channels, according to the following model.

$$x_n = -\sum_{k=1}^p a_k x_{n-k} + \varepsilon_n \quad (1)$$

where p is the order of the AR model and ε_n is the prediction error process. The Burg method [20] was used for AR parameter estimation, after finding a reasonable order for the model by using Akaike's Information Criterion [21]:

$$AIC(p) = N \ln(\sigma^2) + 2p \quad (2)$$

where N is the number of observed samples, and σ^2 is the estimated prediction error variance. The “best” AR model order is the one that minimizes the value of $AIC(p)$.

D GMM-UBM

A Gaussian Mixture Model (GMM) is a model for a probability density function (pdf) expressed as a weighted sum of Gaussian probability density functions. The main reason for using GMMs for classification problems is that Gaussian distributions can approximate any arbitrary pdf [10]. The pdf of a GMM λ is thus expressed as

$$p(\mathbf{v}|\lambda) = \sum_{m=1}^M w_m g_m(\mathbf{v}|\mu_m, \Sigma_m) \quad (3)$$

where \mathbf{v} is an N -dimensional feature vector, w_m are the weights, and g_m the individual N -variate Gaussian pdf, which have the following form:

$$g_m(\mathbf{v}|\mu_m, \Sigma_m) = \frac{1}{(2\pi)^{N/2} |\Sigma_m|} \exp\left(-\frac{1}{2}(\mathbf{v} - \mu_m)^T \Sigma_m^{-1} (\mathbf{v} - \mu_m)\right) \quad (4)$$

where μ_m is an N -dimensional mean column vector and Σ_m is a diagonal covariance matrix of the feature element variances.

To train the GMMs, i.e. finding the model parameters, the expectation maximization (EM) algorithm was used. In the expectation-step (E-step) of the EM algorithm, the a posteriori probabilities $\psi_{t,c,m}$ of a feature vector \mathbf{v}_t belonging to the Gaussian mixture model λ_c , also known as class membership weights, are computed in this fashion:

$$\psi_{t,c,k} = P(w_{c,k}|\mathbf{v}_t, \lambda_{c,i}) = \frac{w_{c,k}g_k(\mathbf{v}_t|\mu_{c,k,i}, \Sigma_{c,k,i})}{\sum_{m=1}^M w_{c,m}g_m(\mathbf{v}_t|\mu_{c,m,i}, \Sigma_{c,m,i})} \quad (5)$$

where $i = 1, 2, \dots, I$, $c = 1, 2, \dots, C$, $m = 1, 2, \dots, M$, and $t = 1, 2, \dots, T$, where I is the total number of iterations, C is the total number of classes, M is the total number of mixture components, and T is the total number of feature vectors.

In the maximization step (M-step), the weights, means, and covariance matrices that parameterize the Gaussian mixture models λ_c are computed as follows:

$$w_{c,k,i+1} = \frac{1}{T} \sum_{t=1}^T \psi_{t,c,k} \quad (6)$$

$$\mu_{c,k,i+1} = \frac{\sum_{t=1}^T \psi_{t,c,k} \mathbf{v}_t}{T w_{c,k,i+1}} \quad (7)$$

$$\Sigma_{c,k,i+1} = \frac{\sum_{t=1}^T \psi_{t,c,k} (\mathbf{v}_t - \mu_{c,k,i+1})^2}{T w_{c,k,i+1}} \quad (8)$$

Note that the square of a vector denotes a diagonal matrix with diagonal elements equal to the squared elements of the vector. With every iteration i the likelihood for which the parameters are computed increases, so that a maximum in the likelihood occurs at the last iteration; however, that maximum may have reached a plateau at an earlier iteration. In other words,

$$l(\mathbf{v}_t|\lambda_{c,i+1}) \geq l(\mathbf{v}_t|\lambda_{c,i}) \quad (9)$$

In this study, EM was executed for a maximum of 15 iterations. This number represents a safe choice. For speech data, when the parameters are initialized randomly, the number of iterations needed tends to be over 10 [9], whereas the number of iterations needed is approximately 7 when using k-means clustering for initialization [9], as is used here.

Once the GMMs have been formed, a UBM λ_{UBM} can be created based on the GMMs. In order to do so, each GMM λ_c that will be part of the UBM is adapted by performing the maximum a posteriori (MAP) adaptation. This is done by, first, computing the posterior probabilities of each feature vector belonging to the UBM.

$$\tilde{\psi}_{t,c,k} = P(w_{c,k}|\mathbf{v}_t, \lambda_{UBM}) = \frac{w_{c,k}g_k(\mathbf{v}_t|\mu_{c,k}, \Sigma_{c,k})}{\sum_{m=1}^M w_{c,m}g_m(\mathbf{v}_t|\mu_{c,m}, \Sigma_{c,m})} \quad (10)$$

In our classification problem, $C=2$. Therefore, there will be 2 GMMs, λ_{ADHD} and $\lambda_{non-ADHD}$, and $\lambda_{non-ADHD}$ will be adapted to form λ_{UBM} .

Sufficient statistics are then computed to obtain the weights, means, and variances of λ_{UBM} . These parameters are the count, first, and second moment of the posterior probabilities found in (10).

$$n_{c,k} = \sum_{t=1}^T \tilde{\psi}_{t,c,k} \quad (11)$$

$$E_{c,k}(\mathbf{v}_t) = \frac{\sum_{t=1}^T \tilde{\psi}_{t,c,k} \mathbf{v}_t}{\sum_{t=1}^T \tilde{\psi}_{t,c,k}} \quad (12)$$

$$E_{c,k}(\mathbf{v}_t^2) = \frac{\sum_{t=1}^T \tilde{\psi}_{t,c,k} \mathbf{v}_t^2}{\sum_{t=1}^T \tilde{\psi}_{t,c,k}} \quad (13)$$

Once the sufficient statistics have been computed, the weights, means, and variances are adapted by using (14) through (16)

$$\tilde{w}_{c,k} = \left[\frac{a_{c,k,w} n_{c,k}}{T} + (1 - a_{c,k,w}) w_{c,k} \right] \tilde{\psi}_{t,c,k} \quad (14)$$

$$\tilde{\mu}_{c,k} = a_{c,k,\mu} E(\mathbf{v}_t) + (1 - a_{c,k,\mu}) \mu_{c,k} \quad (15)$$

$$\tilde{\sigma}_{c,k}^2 = a_{c,k,\sigma} E_{c,k}(\mathbf{v}_t^2) + (1 - a_{c,k,\sigma}) (\sigma_{c,k}^2 + \mu_{c,k}^2) - \tilde{\mu}_{c,k}^2 \quad (16)$$

where $a_{m,w}$, $a_{m,\mu}$, $a_{m,\sigma}$ are the adaptation coefficients for the weights, means, and variances respectively. These coefficients can be computed by using the following formulae:

$$a_{c,k,w} = \frac{n_{c,k}}{n_{c,k} + \alpha_w} \quad (17)$$

$$a_{c,k,\mu} = \frac{n_{c,k}}{n_{c,k} + \alpha_\mu} \quad (18)$$

$$a_{c,k,\sigma} = \frac{n_{c,k}}{n_{c,k} + \alpha_\sigma} \quad (19)$$

where α_w , α_μ , α_σ are the relevance factors of the weights, means, and variances. In this study, relevance factors were set to 10. Relevance factors can potentially affect classification performance, an occurrence not observed here.

In this study, UBMs were found using various combinations of features extracted from Non-ADHD subjects (impostors). Models were also found to fit the class of ADHD subjects (targets). Fig. 2 summarizes how classification is done in this study.

For classification, the log-likelihood ratio (LLR) is used, i.e. the ratio of the likelihood of a test vector \mathbf{v}_t belonging to the ADHD model over the likelihood of \mathbf{v}_t belonging to the universal background model. If the LLR is greater than or equal to zero, the subject is classified as ADHD, otherwise the subject is classified as Non-ADHD.

$$LLR = \log \left(\frac{p(\mathbf{v}_t|\lambda_{ADHD})}{p(\mathbf{v}_t|\lambda_{UBM})} \right) \quad (20)$$

To speed up the process of training the GMM-UBMs, the MSR Identity toolbox [22] was used.

4. Experiments

The experiments were conducted based on 7 subjects, all 6 year old males: 4 A and 3 NA subjects.

As indicated in Section 3B, these 5 channels were used: Fc1, Fc2, Fc5, Cp6, and C3. AR(7) parameters were extracted from these

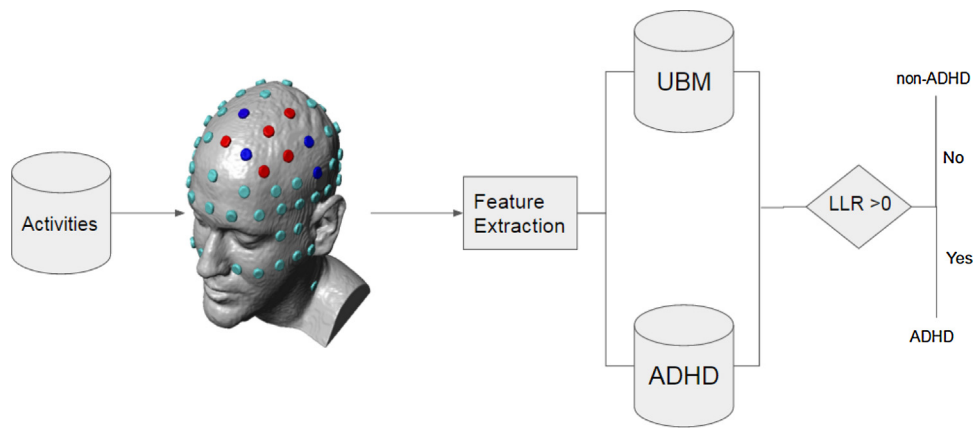


Fig. 2. GMM-UBM for the classification of A/NA subjects.

channels and these parameters concatenated, forming 35-D feature vectors. Feature vectors were extracted during ANT, VIDEO, and eyes closed (EC) activities.

By using 4 subjects for training (2 NA and 2 A) and the other 3 for testing (1 NA and 2 A), GMM-UBMs were trained. The training dataset consisted of the 35-D AR parameter vectors extracted from 2-s windows with 50% overlap during the ANT, VIDEO, and/or eyes closed (EC) activities. Given the available recorded data, when overlapping 2-s windows were used, for every subject 423 feature vectors were extracted during ANT; 112 during VIDEO; and 56 during EC.

Throughout, the performance of classification in the form of the receiver operating characteristic (ROC) was analyzed. When characterizing the performance of systems that employ biometrics, EER is often used. EER corresponds to the point on the ROC curve where the miss rate or false negative/Non-ADHD rate (FNR) is equal to the false positive/alarm/ADHD rate (FPR). In addition, the area under the curve (AUC) of the ROC plot can be used to describe the detection (or true positive/ADHD) rate vs the false positive rate (FPR). In this study, AUCs and EERs are used as performance indicators.

A Effect of number of mixture components

For speaker verification and/or identification problems, it is common practice to use one mixture model per speaker [9]. Because this paper concerns a binary classification problem, two mixture models are formed (one for A and one for NA subjects). To decide how many mixture components should be computed, the effect of the number of mixture components is explored.

As stated earlier in this section, 4 subjects were used for training. The results in Fig. 3 were obtained using 2 A subjects and 0 NA subjects for testing during ANT. Peak values are found when the number of mixture components equals the number of subjects used for training. When 4 mixture components are used to fit the data, the AUC is 0.981 and the EER is 0.047. When 8 or more mixture components are computed, the AUC fluctuates between 0.972 and 0.981 and the EER between 0.059 and 0.066. To minimize the computational burden and to maximize detection, the number of mixture components used in the rest of the paper is set equal to the number of training subjects.

B Effect of activities

Our hypothesis is that the accuracy of ADHD detection depends on the activity performed by the subjects, i.e. some activities elicit stronger discrimination statistics than others. Furthermore, if the activity a test subject performs differs from that presumed to be

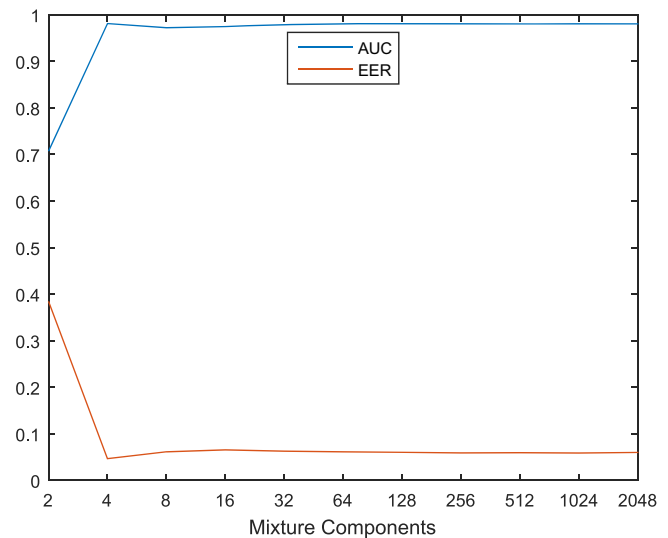


Fig. 3. Effect of the number of mixture components.

performed by the training subjects, the miss rate is expected to increase.

For the results reflected in Fig. 4, GMM-UBMs are trained with 2 subjects (1 NA and 1 A) and tested with the other 5. The training subjects were paired in $(4 \times 3 =)$ 12 different ways in order to train and test with all possible combinations. Four scenarios were considered for training and testing: using eyes closed (EC) data for training and testing (dark blue histogram); using ANT for training and testing (light blue histogram); using ANT for training and EC for testing (yellow histogram), and using EC for training and ANT for testing (brown histogram).

Fig. 4 shows that detection performance is poor when ANT data is used for training and EC is used for testing (yellow) and vice versa (brown); apparently there is not much commonality in the functioning of the brain between when the eyes are closed and when actively paying attention to some task. For both scenarios, there is a large concentration of miss rates between 0.35 and 0.65, which is the equivalent of guessing. The average EER for these cases is 0.45. Slightly better results are found when EC data is used for training and testing (dark blue), but with the EERs spread out more. For the fourth case, when ANT data is used for training and for testing, the AUC tends to be higher than 0.6 and the EERs tend to be below 50%, with a mean EER of 39%.

Hypothesizing that doing so would increase the detection performance, training was done next with 4 subjects and testing with

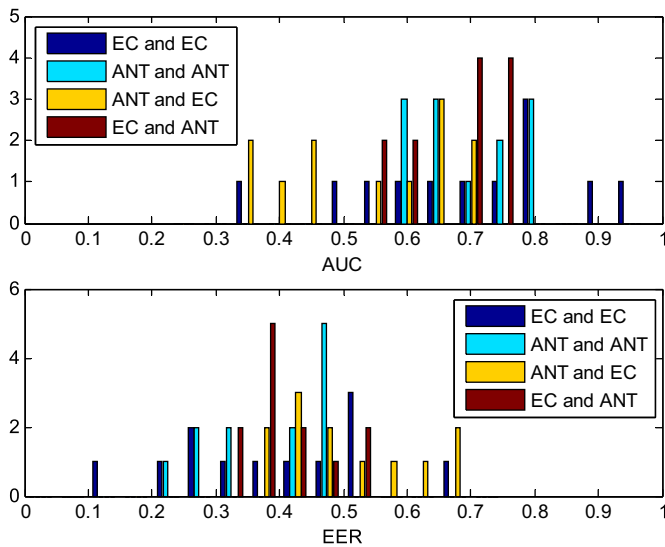


Fig. 4. Distribution of AUCs (top) and EERs (bottom) when training/testing = EC/EC (dark blue), ANT/ANT (light blue), ANT/EC (yellow), and EC/ANT (brown); all combinations of 2 subjects (1A and 1 NA) used for training and all other non-overlapping subjects for testing. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

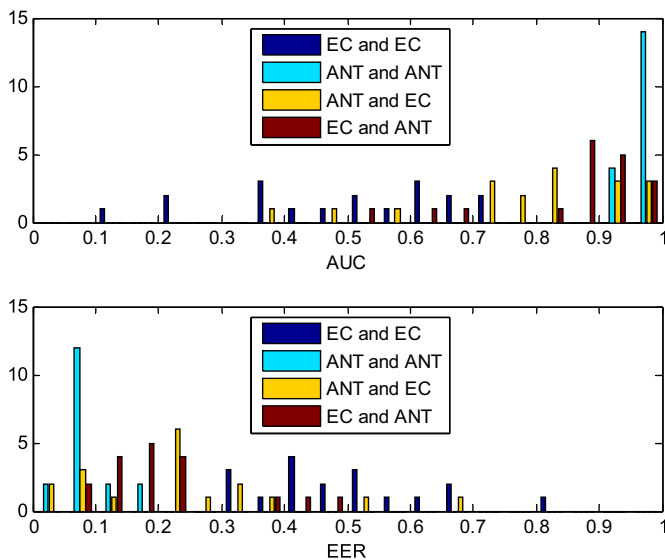


Fig. 5. Distribution of AUCs (top) and EERs (bottom) for training/testing cases EC/EC (dark blue), ANT/ANT (light blue), ANT/EC (yellow), and EC/ANT (brown); all combinations of 4 subjects (2 A and 2 NA) used for training and all other non-overlapping subjects for testing. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the other 3. The training/testing cases EC/EC, ANT/ANT, ANT/EC, and EC/ANT were explored, and the EERs and AUCs were computed for all 18 combinations of 4 (picking 2 A and 2 NA, from the available 4 A and 3 NA). Fig. 5 summarizes the result of these experiments.

When using EC for training and testing, the EERs range from 0.30 to 0.81 with a mean of 0.50, indicative of a classifier that is (as with less training), still equivalent to guessing. Likewise, the AUCs for this EC/EC scenario are spread between 0.11 and 0.74 with a mean of 0.49, representative of guessing as well. When using ANT for training and EC for testing, the results improve: the EERs are now spread between 0.025 and 0.65 with a mean of 0.24. The AUCs for this case follow a similar pattern: they range between 0.37 and 0.99, with a mean of 0.79, which means that the classifier makes correct decisions most of the time. When EC is used for training and ANT

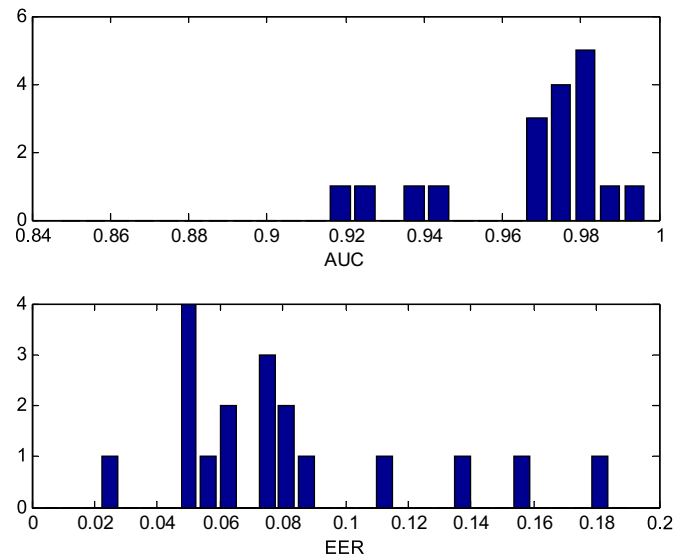


Fig. 6. Distribution of AUCs (top) and EERs (bottom) when ANT feature vectors from 4 subjects are used for training (2A and 2 NA) and from the other 3 subjects for testing.

for testing (yellow), the AUCs and EERs are distributed very much as in the ANT/EC case, and the mean AUC and EER are 0.86 and 0.21 respectively. The fourth case (ANT for training and ANT for testing) has improved substantially, with all of the AUCs over 0.92 and most over 0.95. The mean AUC is 0.97. The EERs for this ANT/ANT case are now spread between 0.025 and 0.18. The mean EER is now 0.082 with most of the EERs below 0.11. As expected, these results make a strong case for using more subjects for training.

Fig. 6 explores in finer detail the distribution of the case where 4 subjects during ANT activity are used for training. The distribution of AUCs (top graph) has a long-tail-like shape, with most of the values over 0.97. As pointed out for Fig. 5, there is a handful of cases with AUC below 0.95. Corresponding to the latter cases, there is the same handful of cases with EER (bottom graph) above 0.11. Except for these – what seem to be outlier – cases, the EERs are distributed between 0.02 and 0.085 and centered at 0.065, which happens to be almost 2% points below the mean EER of 0.082. Note that even the worst of the 18 possible subject combinations for training and testing produces AUC above 0.90 and EER below 0.20.

These experiments suggest that the more ANT data is used during training, the better the performance; when 2 subjects were used for training, EERs were larger not only because the number of feature vectors was smaller, the number of mixture components was smaller also. When only EC is used for training, classification performance becomes closer to guessing. When ANT was used for training and EC for testing and 2 subjects are used for training, the mean EER was found to be 0.45. On the other hand, when 4 subjects are used, doubling the number of feature vectors in the training dataset, the mean EER drops to 0.24. Finally, when only ANT data is used, the mean EER is 0.39 when two subjects are used for training and drops to less than 0.09 when four subjects are used for training.

C GMM-UBM with EC and ANT data

In this section, the hypothesis will be explored that classification performance degrades as resting data is mixed in – as a contamination – with the modeled training data. GMM-UBMs were trained using 4 subjects and tested with the remaining 3, those that were not used for training (recall that 4 A and 3 NA subjects are used for these experiments).

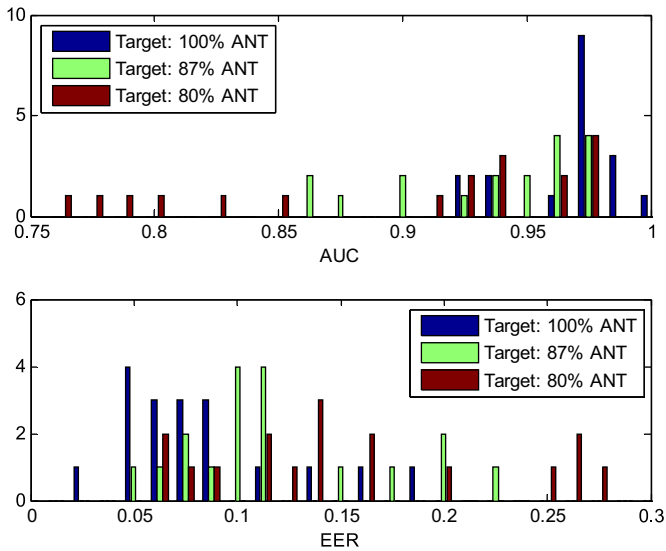


Fig. 7. AUCs (top) and EERs (bottom) of GMM-UBMs with ANT + EC (mixed) composition training datasets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 7 shows how performance degrades as ANT data is combined with EC data for training of the UBM. As was shown in Fig. 6, when only features extracted during ANT activity were used for training, AUCs vary between 0.92 and 0.99 with a mean of 0.97; EERs vary between 0.025 and 0.18 with a mean of 0.082. When 87% of the training dataset comes from ANT activity (384 feature vectors during ANT and 56 during EC), performance degrades slightly: AUCs now vary between 0.86 and 0.98 with a mean of 0.94; EERs vary between 0.053 and 0.22 with a mean of 0.12. When 80% of the data comes from ANT activity (220 feature vectors from ANT and 56 from EC), the AUCs are spread between 0.76 and 0.98 with a mean of 0.90. Lastly, for this case, the EERs are distributed between 0.057 and 0.30 with a mean of 0.16.

Fig. 8 contains sample AUCs and EERs from the distributions shown in Fig. 7. The top figures and left plots show how performance degrades as some EC (resting) data is mixed in with ANT data to train the UBM. The average ROC plot for the top left graph has an AUC of 0.97, whereas the worst ROC has an AUC of 0.92. The average AUC decreases for all the other ROCs: 0.94 (top right ROC), 0.90 (bottom left), and 0.55 (bottom right), which is equivalent to guessing. Similarly, the least favorable AUCs of the ROCs decrease as more resting data is added to the training mixture.

Fig. 9 shows a different representation of the ROC plots shown in Fig. 8 in log scale. Through the use of EERs, Fig. 9 confirms that performance decreases as resting data is added to the dataset. When 100% of the training dataset is ANT data, the average EER is 0.082, and the worst EER is 0.18. The latter implies that the worst case probability of detection is above 81%. When 87% of the dataset is ANT data, the average EER becomes 0.12 and the worst EER becomes 0.22. When 80% of the data is ANT data, the average EER becomes 0.16 and the worst becomes 0.30. Lastly, when 0% of the dataset is ANT data, the average EER is 0.46. In other words, most GMM-UBM models made using EC data will tend to guess whether a test vector belongs to the A class or to the NA class. For models using only EC data, the distribution of EERs goes from 0.30 to 0.81, and is centered around 0.5 (see Fig. 5).

Fig. 10 shows how performance decreases when the percentage of resting data in the training dataset increases even further. Since only 56 feature vectors could be extracted during resting EC activity, VIDEO activity was used. The latter is another baseline (resting activity) during which EEG recordings were taken from the subjects. In Fig. 10, the scenario where 60% of the data was during ANT consisted of 166 feature vectors extracted during ANT, 56 during EC, and 54 during VIDEO. The scenario where 50% of the dataset was during ANT consisted of 166 vectors extracted during ANT, 56 during EC, and 110 during VIDEO.

Fig. 10 shows an exaggeration of the behavior patterns observed in Fig. 7. When 60% of the training and testing data comes from ANT, the AUCs are distributed between 0.62 and 0.86 with a mean of 0.76, which is where most of the AUCs are concentrated. The EERs, for this 60% mixed case, are scattered between 0.21 and 0.42 with a

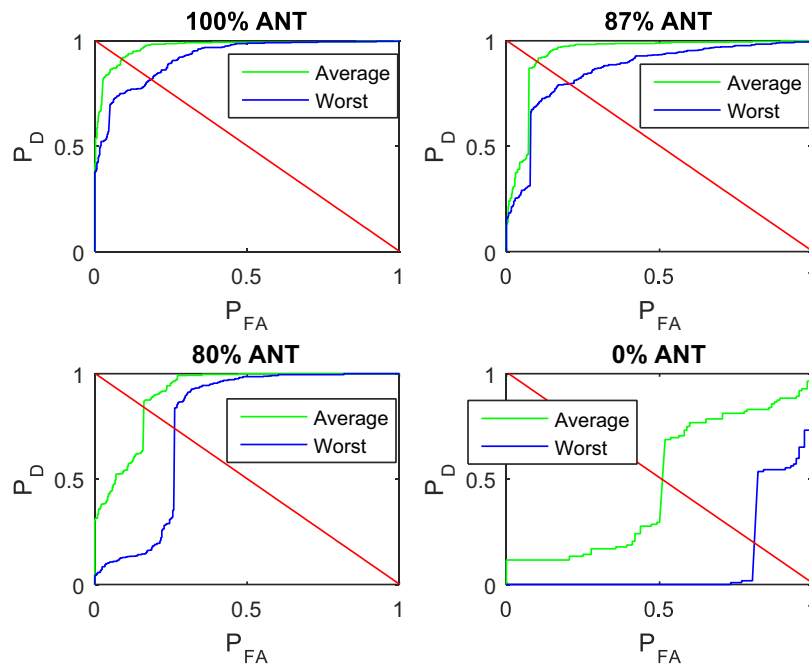


Fig. 8. Sample ROC plots with different training datasets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

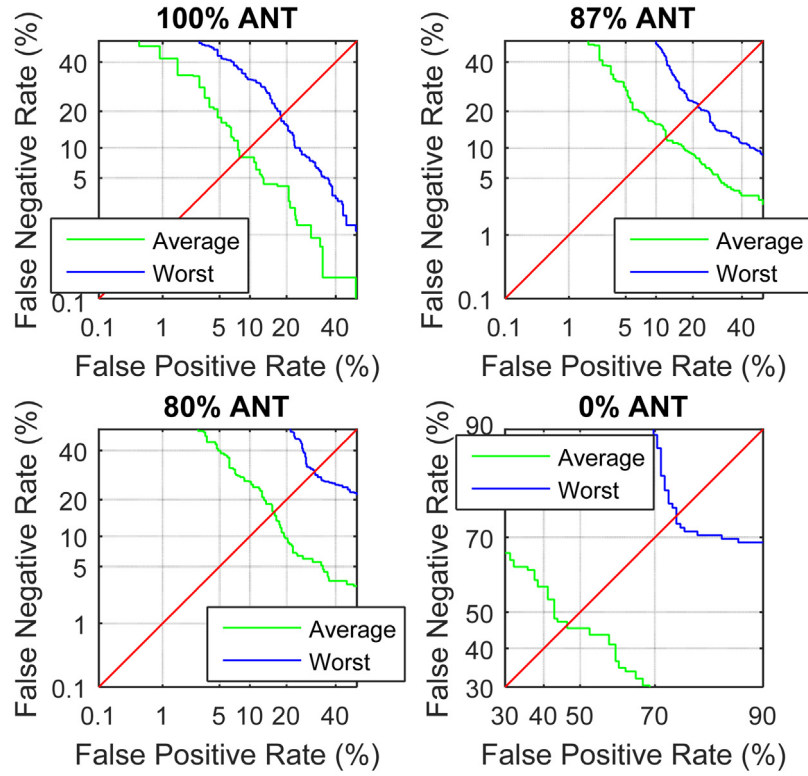


Fig. 9. Sample DET curves with different datasets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

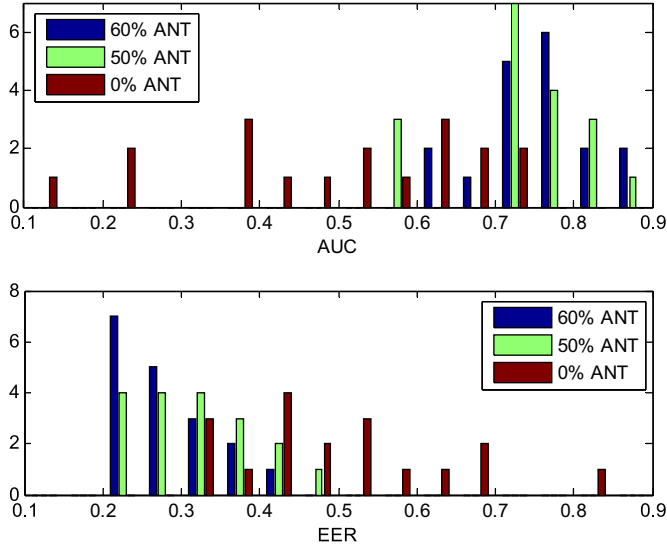


Fig. 10. AUCs (top) and EERs (bottom) of GMM-UBMs with ANT+EC+VIDEO (mixed) composition training datasets and same composition testing sets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

mean of 0.29. When 50% of the training and testing data comes from ANT, performance drops slightly. The distribution of AUCs for this case is similar to the latter, but there are some cases where the AUCs are slightly below 0.6 and the EERs are between 0.2 and 0.5. The mean AUC and EER for this 50% mixed case are 0.73 and 0.32 respectively. For these 60% and 50% mixed cases, the performance is slightly, but consistently, better than guessing.

The final case shown in Fig. 10 is when all the training and testing feature vectors are extracted from EC. The AUCs and EERs are spread between 0.11 and 0.73, and 0.30 and 0.81 respectively, meaning

Table 1

Summary of AUC under different training and testing scenarios (percentages in mix).

Testing Training	ANT (100)	ANT + EC (87, 13)	ANT + EC + VIDEO (50, 33, 17)	EC (100)
ANT (100)	0.92 to 1.00 Mean 0.97	0.90 to 0.99 Mean 0.96	0.74 to 0.98 Mean 0.92	0.37 to 0.98 Mean 0.79
ANT + EC (87, 13)	0.87 to 1.00 Mean 0.96	0.86 to 0.98 Mean 0.94	0.75 to 0.98 Mean 0.90	0.34 to 0.98 Mean 0.77
ANT + VIDEO + EC (50, 33, 17)	0.92 to 1.00 Mean 0.96	0.89 to 0.98 Mean 0.94	0.76 to 0.98 Mean 0.81	0.27 to 0.92 Mean 0.68
EC (100)	0.55 to 0.97 Mean 0.86	0.55 to 0.96 Mean 0.85	0.44 to 0.97 Mean 0.80	0.11 to 0.77 Mean 0.49

that classification is akin to guessing. The AUC for 0% ANT, i.e. 100% EC, falls roughly in the 15–75% range. This is just a little better than a pure guess. This performance is very similar to the result found based on phase synchrony processing of eyes closed EEG [23], an entirely different approach than GMM-UBM. As the performance of a KNN approach [4] was also good when using ANT data, improved detection performance appears to be highly correlated with using an attention task instead of the eyes closed condition.

The mixed training data cases show that even 13% of resting data in the training phase has a perceptible effect on classification performance. While the results make clear that it is inadvisable to use resting EEG as part of the training data, note that inadvertent, temporary inattention during an ANT task could well look like resting data. Mitigation is provided by collecting more data across more subjects, and/or actively detecting (and removing from consideration) resting data segments.

Table 1 summarizes the results of the experiments done in order to study the effect of different activities on the classification of A/NA. The left column represents the activities used for training and the top row represents the activities used for testing. A tabular presentation of the data (histogram range and mean) is given to avoid using too many histograms.

Table 1 shows the pattern observed in Figs. 7 and 10: The higher the proportion of ANT data that is used for training and testing, the higher the AUC. When only features extracted during ANT activity are used for training (first row) the AUCs are higher than those in any of the other rows. For instance, the AUC of entry (1,1) (train/test = ANT/ANT) is statistically (on average) higher than for cells (2,1), (3,1), and (4,1). This pattern is observed to hold for the other rows. For instance, the AUC of entry (1,2) is higher than that of (2,2), which is equal to or higher than that of (3,2) and (4,2). These cases show that detection degrades gracefully when the proportion of EC data increases in the training dataset. However, when EC is used for training (last row), the worst-case and mean AUCs drop substantially, which suggests that training models with EC data only should be avoided. When ANT, EC, and VIDEO are used for training, classification performance based on testing with EC (last column) produces very poor worst-case AUC results. Training and testing with EC only yields performance akin to guessing. Thus – for the channels and features used in this study – training with ANT data only, or as much ANT data as possible, yields the highest classification performance.

5. Discussion

While the number of subjects in this study was small, the results show that good discrimination of 6-year old male subjects with and without ADHD appears very possible when EEG collected during an attention network task is used. The significance of this result is that for males diagnosis of ADHD at about 6 years of age is quite common. This may be related to the inattentive-hyperactive subtype being prevalent in males. For females the diagnosis of ADHD generally happens a bit later, which may be related to the inattentive subtype being prevalent in females. Encouraged by the results to date, our future efforts are directed towards more fine-grained methods, subtype detection, and generalization to a larger database including males and females.

6. Conclusions

On the basis of data from 5 channels of EEG, selected based on theta to beta power ratios during an attention task, a Gaussian mixture model for ADHD was developed for likelihood ratio comparison with a (impostor) universal background model for Non-ADHD. When training and testing is done with data extracted during the attention network task (ANT), the worst case detection performance was at least 81% at an equal error rate of at most 18%, while the respective average values were 92% and 8.2%. Therefore, good discrimination performance between 6-year old male ADHD and Non-ADHD subjects can be achieved based on attention task data. The fact that poor discrimination performance results when training and testing is done with data extracted during rest, with eyes closed, supports the hypothesis that ADHD detection performance is task dependent. When training data is contaminated across different tasks discrimination performance deteriorates, even when testing uses only ANT data. If the testing data is also contaminated, discrimination performance generally worsens on average and becomes very poor in terms of worst case performance.

References

- [1] Data and Statistics | ADHD | NCBDDD | CDC, Cdc.gov, 2016. [Online]. Available: <http://www.cdc.gov/ncbddd/adhd/data.html>.
- [2] Diagnostic and Statistical Manual of Mental Disorders, American Psychiatric Association, Washington, DC, 2000.
- [3] B. Abibullayev, J. An, Decision support algorithm for diagnosis of ADHD using electroencephalograms, *J. Med. Syst.* 36 (4) (2011) 2675–2688.
- [4] J. Lopez Marcano, M.A. Bell, A.A. (Louis) Beex, Classification of ADHD and non-ADHD using AR models, Orlando, FL, in: The Proceedings of Engineering in Medicine and Biology Conference (EMBC2016), 363–366, 2017, 16–20 August.
- [5] A. Mueller, G. Candrian, J. Kropotov, V. Ponomarev, G. Baschera, Classification of ADHD patients on the basis of independent ERP components using a machine learning system, *Nonlinear Biomed. Phys.* 4 (1) (2010) S1.
- [6] R. Chabot, H. Merkin, L. Wood, T. Davenport, G. Serfontein, Sensitivity and specificity of QEEG in children with attention deficit or specific developmental learning disorders, *Clin. EEG Neurosci.* 27 (1) (1996) 26–34.
- [7] V. Monastera, J. Lubar, M. Linden, P. VanDeusen, G. Green, W. Wing, A. Phillips, T. Fenger, Assessing attention deficit hyperactivity disorder via quantitative electroencephalography: an initial validation study, *Neuropsychology* 13 (3) (1999) 424–433.
- [8] M.M. Lansbergen, M. Arna, M. van Dongen-Boomsma, D. Sronk, J.K. Buitelaar, The increase in theta/beta ratio on resting-state EEG in boys with attention-deficit/hyperactivity disorder is mediated by slow alpha peak frequency, *Prog. Neuropsychopharm. Biol. Psychiatry* 35 (1) (2011) 47–52.
- [9] P. Raman, Speaker Identification and Verification Using Line Spectral Frequencies, MS Thesis, Virginia Polytechnic Institute and State University, 2015, 2017.
- [10] D. Reynolds, Gaussian mixture models, in: *Encyclopedia of Biometrics*, Springer, 2009, pp. 659–663.
- [11] P. Nguyen, D. Tran, T. Le, X. Huang, W. Ma, EEG-based person verification using multi-sphere SVDD and UBM, in: 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD, Gold Coast Australia, 2013, pp. 289–300.
- [12] E. Thomas, A. Temko, G. Lightbody, W. Marnane, G. Boylan, A Gaussian mixture model based statistical classification system for neonatal seizure detection, in: 2009 IEEE International Workshop on Machine Learning for Signal Processing, Grenoble, France, 2009, pp. 1–6.
- [13] S. Snyder, H. Quintana, S. Sexson, P. Knott, A. Haque, D. Reynolds, Blinded, multi-center validation of EEG and rating scales in identifying ADHD within a clinical sample, *Psychiatry Res.* 159 (3) (2008) 346–358.
- [14] S.K. Loo, M. Arns, Should the EEG-based theta to beta ratio Be used to diagnose ADHD? *ADHD Rep.* 23 (8) (2015) 8–13, <http://dx.doi.org/10.1521/adhd.2015.23.8.8>.
- [15] S. Solhjo, A.M. Nasrabadi, M.R.H. Golpayegani, Classification of chaotic signals using HMM classifiers: EEG-based mental task classification, 13th European Signal Processing Conference (2005) 1–4.
- [16] D. Wulsin, J. Blanco, R. Mani, B. Litt, Semi-supervised anomaly detection for EEG waveforms using deep belief nets, Ninth International Conference on Machine Learning and Applications (ICMLA) (2010) 436–441.
- [17] A. Harati, S. Lopez, I. Obeid, J. Picone, M.P. Jacobsen, S. Tobochnik, The TUH EEG CORPUS: a big data resource for automated EEG interpretation, *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (2014) (December).
- [18] M.I. Posner, S.E. Peterson, The attention system of the human brain, *Annu. Rev. Neurosci.* 13 (1990) 25–42.
- [19] M.R. Rueda, J. Fan, B.D. McCandliss, J. Halparin, D. Gruber, L.P. Lercari, M.I. Posner, Development of attentional networks during childhood, *Neuropsychologia* 42 (2004) 1029–1040.
- [20] S.J. Orfanidis, *Optimum Signal Processing*, Collier Macmillan, 1988.
- [21] H. Akaike, A new look at the statistical model identification, *IEEE Trans. Autom. Control* 19 (6) (1974) 716–723.
- [22] M.S.R. Identity Toolbox, 2017, Available <https://www.microsoft.com/en-us/download/details.aspx?id=52279>.
- [23] G.V. Tcheslavski, A.A. (Louis) Beex, Phase synchrony and coherence analysis of EEG as tools to discriminate between children with and without attention deficit disorder, *Biomed. Signal Process. Control* 1 (2006) 151–161, <http://dx.doi.org/10.1016/j.bspc.2006.08.001>.