33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Using Pre-treatment Electroencephalography Data to Predict Response to Transcranial Magnetic Stimulation Therapy for Major Depression

Ahmad Khodayari-Rostamabad, James P. Reilly, Gary M. Hasey, Hubert deBruin and Duncan MacCrimmon

Abstract— **We investigate the use of machine learning methods based on the pre-treatment electroencephalograph (EEG) to predict response to repetitive transcranial magnetic stimulation (rTMS), which is a non-pharmacological form of therapy for treating major depressive disorder (MDD). The learning procedure involves the extraction of a large number of candidate features from EEG data, from which a very small subset of most statistically relevant features is selected for further processing. A statistical prediction model based on *mixture of factor analysis* (MFA) model is constructed from a training set that classifies the respective subject into responder and non-responder classes. A *leave-2-out* (L2O) cross-validation procedure is used to evaluate the prediction performance. This pilot study involves 27 subjects who received either left high-frequency (HF) active rTMS therapy or simultaneous left HF and right low-frequency active rTMS therapy. Our results indicate that it is possible to predict rTMS treatment efficacy of either treatment modality with a specificity of 83% and a sensitivity of 78%, for a combined accuracy of 80%.**

I. Introduction

In this paper, specific EEG-based biomarkers for predicting efficacy of repetitive transcranial magnetic stimulation (rTMS) therapy for major depressive disorder (MDD) are identified and investigated. MDD is a common mental disorder and a major cause of workplace disability with costs very similar to those of diabetes and heart disease [1]. rTMS therapy, approved in Canada and the USA for use in patients with MDD, employs strong localized pulsed magnetic fields administered through a magnetic coil placed on the head of the subject, to induce electrical currents in the brain to change the activity of neuron populations. rTMS therapy has been proven to be at least as effective as pharmacological treatment. rTMS is commonly reserved for use when antidepressant medications prove ineffective. See, e.g. [2]. Typically, only 40% to 50% of MDD cases will respond to rTMS treatment. Since the duration of an rTMS trial is on the order of 4 weeks, our proposed prediction method could be of great value in the mental health care setting, in that considerable time and resources can be saved by avoiding rTMS treatment on the significant proportion of subjects who are likely to be non–responsive.

The few studies to date using EEG data and employing traditional clinical data analysis have shown limited ability to predict response to rTMS therapy. Price et al. [3] studied correlations between clinical response (after four weeks) and EEG features including individual alpha power (8–13 Hz), alpha frequency as well as asymmetry indexes in 39 depressed subjects. Their analysis found that there is weak evidence of predicted correlation between these features and clinical rating change. This and similar related literature demonstrate the need for more efficient models for prediction of response to rTMS therapy.

II. Methods

The data analysis methodology used in this paper is discussed at length in [4], [5]. Therefore, in this paper, we only present a summary overview of the proposed machine learning techniques.

A training set consisting of the pre-treatment resting EEG and the corresponding response of each subject to rTMS treatment is first collected. A machine learning procedure is developed from this training set in the following way. First, a large set of candidate features is extracted from the EEG signals. The dimensionality of these features is significantly reduced using a feature selection procedure. These reduced-dimensionality features are then fed into a classifier (trained using the known responses) that outputs the predicted response. The procedure is evaluated using a leave-n-out cross validation procedure.

A. Participants

Twenty-seven subjects diagnosed with unipolar MDD (using the internationally recognized Diagnostic and Statistical Manual - IV diagnostic criteria) and recruited into Research Ethics Board approved rTMS studies, were used in our analysis. All subjects had previously failed to respond to at least two courses of antidepressant medication therapy. Clinical information of participants are as follows: Age at start of treatment [years]: avg.=46.3, std=9.85, min=23.9, max=65.8, Gender: 20 female subjects (74%) and 7 male subjects (26%). Pre-treatment Hamilton depression rating scale (HamD) scores were: avg.=21.1, std=3.58, min=15, max=27.

B. rTMS Treatment

There are two modes of treatment used for this study: active and sham. Active treatment, when applied to the

A. Khodayari-R., J. P. Reilly and H. de Bruin are with Electrical and Computer Engineering Department, McMaster University, Hamilton, ON, L8S 4K1, Canada. emails: khodaya@mcmaster.ca, reillyj@mcmaster.ca and debruin@mcmaster.ca

A. Khodayari-R., G. Hasey and D. MacCrimmon are with Department of Psychiatry and Behavioural Neurosciences, McMaster University, and also with Mood Disorders Program, St. Joseph Hospital, Hamilton, ON, emails: ghasey@sympatico.ca and maccrim@mcmaster.ca

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), and by the Etherden Fellowship at St Joseph's Healthcare Foundation.

left dorsolateral prefrontal cortex (DLPFC), consisted of 10 Hz rTMS delivered as 20 trains of eight second duration using a figure of eight coil. Active treatment to the right DLPFC consisted of 1 Hz rTMS delivered as two trains of 60 second duration using a round coil. Intensity for active rTMS was set at 110% of motor threshold, defined as the lowest energy capable of inducing activation of the abductor pollicis brevis muscle of the contralateral thumb. Sham treatment was administered with one of the rTMS coils held at 90 degrees to a tangent at the scalp site and intensity set at a level sufficient to create a click audible with earplugs in place. Treatments were administered using a Dantec Magpro daily for 10 sessions over two weeks to a site five cm anterior, parasagittally, to the activation site for the abductor pollicis brevis muscle. True left, sham right (TLSR) therapy was administered to 18 subjects, while true left and true right (TLTR) therapy was administered to 9 subjects. The selection was done randomly. None of the subjects received sham rTMS over both the left and the right hemispheres. All subjects also received concurrent 'selective serotonin reuptake inhibitor' (SSRI) antidepressant medication during rTMS therapy, and for an additional four weeks thereafter.

C. Definition of Response

Subjects were classified as "responders" if the Hamilton Depression Rating (HamD) score at six weeks showed at least a 50% improvement over the pre-treatment HamD score. The HamD is a well-accepted means of quantifying the severity of depression. For our purposes, the HamD percentage change value is discretized into two values (or classes), corresponding to responder (R) when it is larger than or equal to 50%, and non-responder (NR) otherwise.

D. EEG Recordings and Quantitative Features

The international 10-20 EEG electrode placement system was used, referenced to linked ears. Data were recorded with a sampling frequency of 205 Hz. These data were collected after approximately 10 days off medication and just before beginning rTMS treatment. The data from 16 electrodes Fp1, Fp2, F3, F4, F7, F8, T3, T4, C3, C4, T5, T6, P3, P4, O1 and O2 were used in this study. A QSI-9500 EEG system is used, which filters the signals between [0.5Hz-80Hz] and applies a notch filter at 60 Hz. The patient was in a semi-recumbent position in a sound attenuated, electrically shielded room and an experienced EEG technician prompted patients on signs of drowsiness. Sessions were arranged in the mornings and patients were requested to avoid coffee, drugs, alcohol and smoking immediately prior to the recording. For each patient, a maximum of 3 spontaneous or resting EEG data files each of 3.5 minutes duration were collected while the subject's eyes were open.

For de-artifacting, the data were partitioned into segments of 1 sec. duration. If the input signal on any electrode saturated the acquisition hardware, the corresponding segment was rejected. The signals were then digitally bandpass filtered between 2.5 Hz and 39 Hz to partially mitigate the effects of eye movement, eye blinks and high-frequency muscle artifacts.

For each EEG file, the first 90 seconds of de-artifacted data are used. The selected data are divided into 5 epochs of 30 sec. duration with 50% overlap. Power spectral ratios (which become candidate features to be described later) are calculated using a Welch modified periodogram method over each epoch. The individual windows required for this process are obtained by dividing each epoch into windows of 2 sec. duration with 72% overlap[1]. These settings result in a nominal five epochs per file, times three files per subject to give 15 epochs per subject.

The set of N_c candidate features extracted from each data epoch consists of the anterior/posterior power ratios at various frequencies, and between various electrode pairs, in addition to some ratios involving more than two electrode pairs. The frequency resolution is 1 Hz and the processing bandwidth of 4 Hz up to 36 Hz is used, resulting in a value of $N_c = 1452$ candidate features. PSD ratios were expressed as base-10 logarithms. All candidate feature values were then normalized using their corresponding *z-score* value. The means and standard errors required for this process were calculated from the corresponding features measured from 91 normal i.e. healthy subjects.

E. Feature Selection

After normalization, the most relevant features are selected using the supervised, greedy method of [6]. This procedure is used to reduce the feature set from $N_c = 1452$ candidate features down to a set of only $N_r = 4$ most relevant features. See [4] for further details.

F. Classification and Performance Evaluation

Let the set of reduced features for the ith epoch for a particular subject be assembled into a vector $\mathbf{x}_i \in \mathbb{R}^{N_r}$ and the corresponding discrete-valued response or class be denoted by $y_i \in$ [R, NR], where R = responder and NR = non–responder. The resulting set $\{\mathbf{x}_i, y_i\},\ i = 1, \ldots, M_t$, constitutes a training set, where M_t is the total number of available epochs over all subjects. In this study one of the subject's EEG was heavily artifacted, so only 10 instead of the nominal 15 epochs were available. This results in a value of $M_t = 400$. Because the variance of features extracted across intra-subject epochs is large, all epochs were considered to be statistically independent.

In our study, we used the *mixture of factor analysis* (MFA) technique [7] to build the response predictor model (classifier), based on the *maximum likelihood* classification rule. This method is used previously in [5] to build statistical diagnosis models to discriminate psychiatric disorders.

Since we have multiple epochs for each subject, the final prediction result for each subject is obtained by averaging the MFA likelihood values over all available epochs for that

[1]The 50% and 72% overlap figures were chosen somewhat arbitrarily and can be altered with reasonable limits without significant impact on performance.

TABLE I

PREDICTION PERFORMANCE.

	predicted NR	predicted R	% correct
actual NR	15	3	83.3%
actual R	2	7	77.8%
			avg. = 80%

subject. The resulting averaged value is then quantized to generate a binary (i.e. R or NR) response value.

Kernelized principal component analysis (KPCA) [8] is used to visualize the clustering behaviour in a two-dimensional subset of the feature space. This gives us insight into the clustering and discriminating performance of the feature set, and aids in the identification of outliers.

The performance of the overall structure was evaluated using a leave-n-out (LnO) cross–validation process. In this study, $n = 2$ resulting in 14 folds. This is close to a 10–fold scenario, as suggested by [9]. In order to ensure unbiased results, classifier parameters were optimized, and features were selected, independently in every fold. This procedure ensures that the training process is completely independent of the data used for testing [9].

III. RESULTS

Table I shows the prediction performance in the form of a classification table, when $N_r = 4$ discriminating features are used. There are 18 subjects in NR group and 9 in the R group, for which 6 were treated with the TLSR, and 3 with the TLTR modality. The specificity is 83.3% and the sensitivity is 77.8%. When averaged, these figures result in an overall prediction accuracy of 80%.

A. List of Discriminating Features

A list of discriminating features is shown in Table II. In this table, anterior/posterior PSD ratio features involving more than two electrodes are calculated as $\log_{10}$ of the product of the numerator powers over the product of the denominator powers. Columns 3 and 4 reflect the means and standard deviations (std) of NR and R groups. These values however depend on the pre-processing, feature extraction and normalization procedures. To calculate standard deviation, we first determined the intra–subject average of each discriminating feature over all subject epochs, and then calculated the inter-subject standard deviation of the averaged feature values. The features are shown sorted based on their Fisher discriminant ratio (FDR) [10]. The FDR is defined as the squared difference of the means of that feature between the R and NR groups, normalized to the sum of the variances of the two groups. It is noted that FDR ranking gives a rough indication of the relevance of the feature, and in this study is used only for the purpose of ordering the features in this table. A feature appears in Table II if it is chosen at least once over all folds of the LnO cross-validation procedure, and if its FDR value is greater than approximately 0.8.

One may be tempted to consider some form of combination of the FDR of individual features as an indication of the resulting performance of the proposed predictor. However, this exercise involves a sequence of *one-dimensional* performance measures and as such is not an adequate performance indicator. The joint discriminating power of the features is only evident when observing the separation of the clusters in the *(multiple)* N_r*–dimensional* feature space. Thus care must be exercised when observing the data in columns 3 and 4 of Table II.

The clustering behaviour of the feature space is shown in Fig. 1. This shows a scatter plot of $M_t = 400$ available pre-treatment training samples projected onto only the two non-linear principal components (PC), selected on the criterion of maximum mutual information with the response variable. This figure was generated using the KPCA method with a Gaussian kernel. This figure shows one point per epoch, or a nominal 15 points per subject. Averaging the locations of the projected data samples belonging to each subject results in Fig. 2, where each subject is shown with only a single point. Each subject is arbitrarily assigned an exclusive index within the range $[1, \ldots, 27]$. For clarity of presentation, in Fig. 1 we label only the points corresponding to the two subjects 16 and 21 (R and NR respectively), to show how the projected data vary between epochs for a given subject. These two subjects were arbitrarily selected with one subject in each class. Each point in Fig. 2 is labelled with its corresponding subject index. The separation of the classes is clearly evident from the figure.

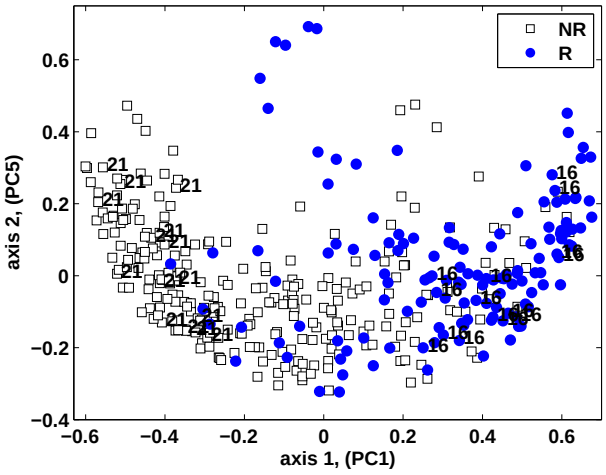

Fig. 1. Epoch-wise scatter plot of the N_r-dimensional feature space derived from the pre-treatment EEG data, projected onto two major nonlinear principal components. There are a nominal 15 epochs per subject. $N_r = 4$.

IV. DISCUSSION AND CONCLUSIONS

We have demonstrated a machine learning capability, based on the pre–treatment EEG, that can predict the response of a subject to rTMS therapy for MDD, with accuracies on the order of 80%. This could have clinical utility in administering rTMS therapy only to a targeted group of subjects who have been determined to have responsive characteristics, and therefore increase the treatment efficacy

TABLE II

A LIST OF MOST DISCRIMINATING FEATURES, SHOWING THE MEAN AND STANDARD DEVIATION (STD) OF EACH FEATURE (AFTER Z-SCORE NORMALIZATION) OVER THE NON-RESPONDER (NR) AND RESPONDER (R) GROUPS.

#	*Selected EEG-driven Numerical Feature*	average ($\pm$ std) for NR group	average ($\pm$ std) for R group	FDR
1	Front-to-Back PSD-ratio at f=6Hz, T3C3/P3O1	0.832 ($\pm$ 0.662)	-0.408 ($\pm$ 0.615)	1.883
2	Front-to-Back PSD-ratio at f=6Hz, F7F3/P3O1	0.104 ($\pm$ 0.437)	-0.912 ($\pm$ 0.634)	1.743
3	Front-to-Back PSD-ratio at f=6Hz, Fp1F7/P3O1	0.273 ($\pm$ 0.478)	-0.666 ($\pm$ 0.597)	1.506
4	Front-to-Back PSD-ratio at f=6Hz, T3C3/T5P3	0.592 ($\pm$ 0.691)	-0.361 ($\pm$ 0.541)	1.181
5	Front-to-Back PSD-ratio at f=6Hz, T3/T5	0.746 ($\pm$ 0.656)	-0.073 ($\pm$ 0.413)	1.115
6	Front-to-Back PSD-ratio at f=6Hz, F3/O1	0.319 ($\pm$ 0.592)	-0.726 ($\pm$ 0.794)	1.112
7	Front-to-Back PSD-ratio at f=24Hz, C3/O1	0.191 ($\pm$ 0.553)	-0.552 ($\pm$ 0.463)	1.062
8	Front-to-Back PSD-ratio at f=23Hz, C3/O1	0.195 ($\pm$ 0.546)	-0.501 ($\pm$ 0.4)	1.058
9	Front-to-Back PSD-ratio at f=7Hz, T3C3/P3O1	0.548 ($\pm$ 0.59)	-0.17 ($\pm$ 0.403)	1.011
10	Front-to-Back PSD-ratio at f=6Hz, Fp1/O1	0.581 ($\pm$ 0.621)	-0.37 ($\pm$ 0.759)	0.938
11	Front-to-Back PSD-ratio at f=6Hz, C3/O1	0.542 ($\pm$ 1.032)	-0.629 ($\pm$ 0.7)	0.882
12	Front-to-Back PSD-ratio at f=8Hz, T3C3/P3O1	0.617 ($\pm$ 0.504)	0.102 ($\pm$ 0.265)	0.82
13	Front-to-Back PSD-ratio at f=29Hz, C3T5/P3O1	-0.029 ($\pm$ 0.639)	-0.68 ($\pm$ 0.351)	0.799
14	Front-to-Back PSD-ratio at f=28Hz, C3/O1	0.222 ($\pm$ 0.458)	-0.553 ($\pm$ 0.737)	0.797

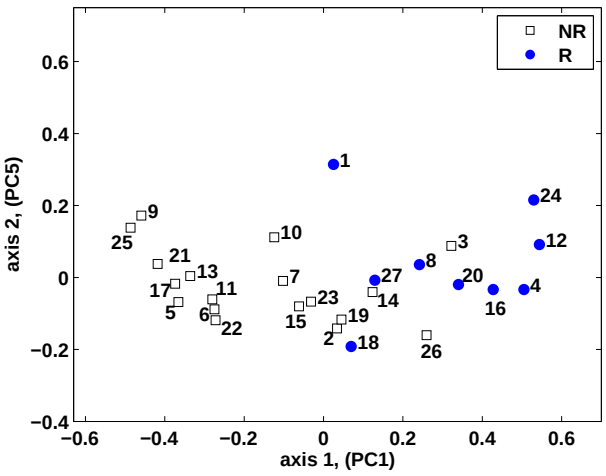

Fig. 2. Subject-wise scatter plot which is obtained from Fig. 1 by averaging all points (epochs) corresponding to each subject.

for this form of therapy. Since the number of subjects in this study is small, our findings must be replicated over a much larger sample before any definitive conclusions can be drawn.

We noticed that the feature selection algorithm of [6] is sensitive to many parameters, including the number N_c of candidate features, the feature normalization procedure, as well as the ratio of the number of subjects in the R versus NR groups. However, it was noted that any overall performance degradation due to such changes in selected features was not severe. The fact that selected features are sensitive in this way suggests that the feature selection procedure of [6] is suboptimal; indeed, this may be a consequence of the greedy nature of the algorithm. An avenue for future work is the investigation of alternative methods of feature selection for this application. In addition, since the effectiveness of rTMS therapy may be related to technical factors such as the frequency, intensity and site of stimulation, a further suggestion for future work is the determination of optimal settings for these parameters.

We examined using $N_r = 5, 6, 8$ discriminating features instead of $N_r = 4$. The prediction performance were unchanged indicating that the prediction methodology is robust and is not very sensitive to the choice of N_r.

Over–fitting is always a consideration in any machine learning application. An indication that over–fitting has not been a dominant factor in this application is provided by Fig. 2. Here, we see that a straight–line boundary (which is specified in terms of only two parameters) is sufficient to separate the two classes. Since the number of training points (in this case 27) is large in comparison to the number of parameters necessary to describe the boundary, it is unlikely that the boundary has over–fit the data.

REFERENCES

[1] B. G. Druss, R. A. Rosenheck and W. H. Sledge, "Health and disability costs of depressive illness in a major U.S. corporation," *American Journal of Psychiatry*, vol. 157, pp. 1274–1278, 2000.

[2] J. P. Lefaucheur, "Methods of therapeutic cortical stimulation," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 39, pp. 1–14, 2009.

[3] G. W. Price, J. W. Lee, C. Garvey and N. Gibson, "Appraisal of sessional EEG features as a correlate of clinical changes in an rTMS treatment of depression," *Clinical EEG and Neuroscience*, vol. 39, pp. 131–138, 2008.

[4] A. Khodayari-Rostamabad, *et al.*, "A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy," *Clinical Neurophysiology*, vol. 121, pp. 1998–2006, 2010.

[5] A. Khodayari-Rostamabad, J. P. Reilly, G. M. Hasey, H. DeBruin and D. J. MacCrimmon, "Diagnosis of psychiatric disorders using EEG data and employing a statistical decision model," in *Proc. Int. Conf. IEEE Eng. Medicine & Biology Society*, pp. 4006–4009, 2010.

[6] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, 2005.

[7] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," Department of Computer Science Technical Report, CRG-TR-96-1, University of Toronto, Toronto, Canada, 1996.

[8] K. R. Müller, *et al.*, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Networks*, vol. 12, pp. 181–201, 2001.

[9] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[10] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Academic Press, 2008.