Using Pre-treatment EEG Data to Predict Response to SSRI Treatment for MDD

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

***Abstract*— The problem of identifying in advance the most effective treatment agent for various psychiatric conditions remains an elusive goal. To address this challenge, we propose a machine learning (ML) methodology to predict the response to a selective serotonin reuptake inhibitor (SSRI) medication in subjects suffering from major depressive disorder (MDD), using pre-treatment electroencephalograph (EEG) measurements.**

The proposed feature selection technique is a modification of the method of Peng et al [10] that is based on a Kullback–Leibler (KL) distance measure. The classifier was realized as a kernelized partial least squares regression procedure, whose output is the predicted response. A low-dimensional kernelized principal component representation of the feature space was used for the purposes of visualization and clustering analysis. The overall method was evaluated using an 11-fold nested cross-validation procedure for which over 85% average prediction performance is obtained. The results indicate that ML methods hold considerable promise in predicting the efficacy of SSRI antidepressant therapy for major depression.

I. Introduction

Major depressive disorder (MDD) is a serious mental disorder and is now the third largest cause of workplace disability. By the year 2020, depression is expected to account for about 15% of total global disease burden, second only to ischemic heart disease. In industrialized countries mental illnesses may account for about 16% of total health care costs and for about 30% of disability claims [1].

Despite the significance of MDD, objective procedures for selecting optimal treatments are lacking. The choice of antidepressant therapy is currently based on personal preference, weighted by clinical factors such as family history, symptom clustering and previous medication history. An effective algorithm for selecting the optimal antidepressant treatment on the basis of symptomatic presentation and other clinical data has proven to be an elusive objective, probably because the same collection of depressive symptoms may be produced by several different neurobiological pathologies. Typically, 60 to 70% of subjects do not remit after the first antidepressant medication trial [2]. Although 67% of those treated for MDD will eventually reach remission, up to 4 different antidepressant treatment trials may be required, each taking 6 weeks or longer. The personal and economic cost of delayed or ineffective therapy is substantial. Clearly, choosing an effective treatment during the initial trial would be of immense clinical and economic value.

A methodology that can employ pre-treatment measures to predict the response to an SSRI treatment, such as the one proposed in this paper, would eliminate the inefficient trial-and-error process that often characterizes the management of MDD.

Several works analyzed resting electroencephalograph (EEG) data for predicting treatment outcome in depressed subjects, e.g. [4]–[9]. In this paper, we have extended the state of the art to develop high performance machine learning (ML) methods of analyzing pretreatment EEG to predict response to SSRI treatment for MDD.

II. Methods

Pre-treatment resting or spontaneous EEG signals, denoted by $\boldsymbol{E}_i, i = 1, \ldots, M_t$ where M_t is the number of training epochs, are collected from M subjects who participated in the study and suffer from MDD. They were then prescribed an SSRI medication. The corresponding response outcome y_i of the patient to the treatment, after completion of a full treatment plan, is recorded. The possible values for the y_i are either "R" (responder), or "NR" (non–responder). The set of EEG recordings and the corresponding outcomes is referred to as a *training set*, denoted by $\mathcal{D} = \Big\{(\boldsymbol{E}_i, y_i), \quad i = 1, \ldots, M_t\Big\}$. These EEG signals from each subject are pre-processed to extract a large number N_c of candidate features $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{N_c}$ that might be relevant for prediction. These are then reduced to a set of most relevant features $\boldsymbol{x}_i \in \mathbb{R}^{N_r}$, where $N_r \ll N_c$, to extract those features which are most indicative of the response outcome. These reduced-dimensionality features are then fed into a classifier which outputs the predicted response to the treatment.

A. Subjects and Clinical Details of the Study

Twenty two subjects (9 males, 13 females, age 20.6 to 62.6, mean 48.9 years) diagnosed with MDD using the internationally recognized Diagnostic and Statistical Manual–IV diagnostic criteria were treated with a 6 week course of an SSRI antidepressant (particularly, the drug Sertraline hydrochloride, with the trade name Zoloft).

The definition of a responder to the SSRI medication in this case was taken to be at least a 25% improvement between the pre– and post–treatment Hamilton depression

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: akhodayari@ieee.org, reillyj@mcmaster.ca and debruin@mcmaster.ca

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 by the Natural Sciences and Engineering Research Council of Canada (NSERC).

rating scales. This is a 17 item clinical procedure undertaken by psychiatric interview with the patient and yields a quantitative indication of the severity of depression.

The EEG recording procedure is described as follows. Sixteen channels of EEG (standard 10–20 system referenced to linked ears) were recorded at a sampling frequency of 205 Hz, after approximately 10 days of medication withdrawal and before a 6-week trial of antidepressant treatment was administered. The EEG electrodes used in this study are Fp1, Fp2, F3, F4, F7, F8, T3, T4, C3, C4, T5, T6, P3, P4, O1 and O2. For each patient, a maximum of 6 EEG files of 3.5 minutes duration were collected, 3 with eyes open (EO), and 3 with eyes closed (EC) conditions. For de-artifacting, the data were partitioned into segments of 1 second duration. If the input signal on any electrode saturated the acquisition hardware, the entire segment was rejected. The signals were then digitally bandpass filtered after recording between 3 and 32 Hz to partially mitigate the effects of eye movement and muscle artifacts. All available EO and EC EEG measurements were used in this study.

For each EEG file, only the first 45 seconds of de-artifacted data are used, (with no apparent degradation in performance) in an effort to reduce computational demands. The selected data are divided into 2 epochs of 30 sec. duration with 50% overlap. Each epoch is further divided into 1 sec. windows with 50% overlap to calculate the statistical quantities which become the candidate features as described below. This makes a total of 12 epochs per subject. However, for one particular subject, only 10 epochs are available. Therefore, the total number of available epochs is $M_t = (12 \text{ epochs/subject } \times 22 \text{ subjects } - 2) = 262$. All candidate features extracted from the training set are normalized before further processing to lie within the interval $[-1, 1]$. We note that similar response prediction performance is obtained when a *z-score normalization* method was used instead.

B. Feature Selection

The set of N_c candidate features extracted from each data epoch consist of the following statistical quantities: spectral coherence between all electrode/channel pairs [1], the mutual information between all electrode pairs, absolute and relative power spectral density (PSD) levels, the log ratio of left-to-right hemisphere powers, and anterior/posterior power ratios at all frequencies between 4Hz and 23Hz with 1Hz resolution and between all electrode pairs. Such quantities have been used in previous related work; e.g., [8] used coherence. Also [7], [8] have used inter and intra-hemispheric power ratios as numerical indicators of treatment response. With a 1 Hz frequency resolution and 16 electrodes, we have $N_c = 4336$ candidate features.

Before describing the method for feature selection, we first define some required variables. Let $\boldsymbol{X} \in \mathbb{R}^{M_t \times N_r}$ be the matrix of selected features. Its rows $\boldsymbol{x}_i^T \in \mathbb{R}^{N_r},\ i = 1, \ldots, M_t$ contain all selected features for the ith training sample (i.e., $\boldsymbol{x}_i$ are the feature vectors). The column vectors $\boldsymbol{\chi}_\ell \in \mathbb{R}^{M_t},\ \ell = 1, \ldots, N_c$ contain the values of the ℓth feature extracted from all training samples.

We propose a sub–optimal approach for feature selection which is a modification of the method of [10] (as well as [11]) based on the Kullback-Leibler (KL) distance [12]. We consider the KL distance between the *pdf* of the ℓ-th candidate feature given that the subject is a responder, and the *pdf* of the ℓ-th feature given that the subject is a non-responder. "Good" features are those for which this KL distance is large. To form these distributions, we divide the column vector $\boldsymbol{\chi}_\ell\ \ell = 1, \ldots, N_c$ into two subsets $U_\ell^{(\mathrm{R})}$ and $U_\ell^{(\mathrm{N})}$ (using the known response y_i corresponding to each element as a response indicator). These subsets contain the values of respective feature over the responder and non-responder groups, respectively. Approximations to the *pdf's* of these subsets can be evaluated using histograms or Parzen windows. The idea of the proposed method is to choose the feature at the jth step, $j = 1, \ldots, N_r$ so that the selected feature is a combination of maximum relevance and minimum redundancy.

More precisely, the first column of $\boldsymbol{X}$ is the vector whose corresponding feature has maximum KL distance between responders and non-responders. Then, at the jth step, we already have $\boldsymbol{X}(j-1) \in \mathbb{R}^{M_t \times (j-1)}$, the matrix corresponding to the previously selected most relevant features. Let us define sets $\mathcal{L} = \{\ell \mid \ell = 1, \ldots, N_c\}$, $\mathcal{J}_1 = \{n_q \mid q = 1, \ldots, j-1\}$ (the set of indexes already chosen in previous steps) and $\bar{\mathcal{J}} = \{\mathcal{L} - \mathcal{J}_1\}$, the set of remaining indexes. The task is to select the jth feature vector whose index is $n_j \in \bar{\mathcal{J}}$ (i.e., the jth column $\boldsymbol{\chi}(n_j)$ of $\boldsymbol{X}(j)$). In a manner similar to [10], this can be done by solving the following "regularized" optimization problem which implements a tradeoff between maximum relevance and minimum redundancy. The index n_{opt} corresponding to the optimal feature is then given by

$$
\begin{aligned}
n_{\text{opt}} \;=\; & \arg\max_{n \in \bar{\mathcal{J}}} \Bigg\{ D_{\text{KL}}\left(U_n^{(\mathrm{R})} \| U_n^{(\mathrm{N})}\right) \\
& +\; \xi\, \frac{0.5}{j-1} \sum_{n_q \in \mathcal{J}_1} \Big[D_{\text{KL}}\left(U_n^{(\mathrm{R})} \| U_{n_q}^{(\mathrm{R})}\right) \\
& \qquad + D_{\text{KL}}\left(U_n^{(\mathrm{N})} \| U_{n_q}^{(\mathrm{N})}\right) \Big] \Bigg\}
\end{aligned} \tag{1}
$$

In the above, $\xi > 0$ is a regularization parameter (with default value of $\xi = 1$) which controls the relative weighting between the relevance and the redundancy of each feature, the sets $U_n^{(\cdot)}$ (indexed by n) represent the R and NR subsets formed from the feature column $\boldsymbol{\chi}(n),\ n \in \bar{\mathcal{J}}$ under test, and subsets $U_{n_q}^{(\cdot)}$ (indexed by $n_q \in \mathcal{J}_1$) are the R and NR subsets corresponding to the features already chosen in previous iterations. $D_{\text{KL}}(\cdot\|\cdot)$ denotes the KL distance between distributions of its two arguments. The resulting column $\boldsymbol{\chi}(n_{\text{opt}})$ from (1) is appended to $\boldsymbol{X}(j-1)$, j is incremented, and the process repeats until all N_r features are found. The first D_{KL} term in (1) expresses relevance of

[1]The magnitude squared coherence estimate was calculated using the Welch averaged periodogram method.

the proposed feature; i.e., we desire discriminating features for which there is a large KL distance between their R and NR subsets. The second two terms express redundancy; here, the objective is to encourage the choice of features whose corresponding subsets have small statistical dependence with those already chosen. The criterion of (1) is a compromise between these two competing measures.

C. The Classification Method

The kernel partial least squares regression (KPLSR) method [13], with a Gaussian Kernel, was employed for the classification procedure. With the KPLSR method, a regression function $f(\boldsymbol{x})$ is trained so that $f(\boldsymbol{x}) = 1$ if $\boldsymbol{x}$ belongs to the non-responder group, and $f(\boldsymbol{x}) = 2$ if $\boldsymbol{x}$ corresponds to a responder. (The values 1 and 2 are chosen arbitrarily). Since there are M_p epochs per subject, a single decision is produced by evaluating $\hat{y}_p = f(\boldsymbol{x}_p), p = 1, \ldots, M_p$, where the $\boldsymbol{x}_p$ are all the feature vectors for the subject under consideration. The mean over all the $\hat{y}_p$ values is then quantized to the nearest integer $[1, 2]$ from which the NR or R prediction for that subject is made.

D. The Nested Cross-Validation Procedure

The performance of the proposed treatment-response prediction process was evaluated using a *nested cross-validation procedure* described as follows. The outer loop is a 11-fold cross-validation, in which the 22 subjects available in our study are divided into 11 contiguous subsets each of size of 2 subjects, so that each subject is tested once. In each fold, the epochs corresponding to 2 subjects are considered as test data and are removed from the training set. The remaining epochs (belonging to the remaining 20 subjects) are used for training, which includes feature selection and specifying the classifier/regressor. This is equivalent to a *leave-2–out* (L2O) testing procedure. The performance index used in this case is the average correct prediction rate, which is evaluated as the arithmetic mean of the sensitivity and specificity values obtained using the result of the outer 11-fold cross-validation procedure just described.

Optimal values for the two design parameters associated with the KPLSR method (i.e., the number of major latent vectors and the standard deviation value of the Gaussian kernel function) are determined inside each fold (according to [14]) using a simple two–dimensional grid search. This is accomplished by further dividing the training subset (including 20 subjects as described above) into contiguous validation subsets of size 2 and using an inner 10-fold cross-validation loop. This means 18 subjects are used for training and 2 subjects for validation in each fold of the inner loop. The average validation performance by this inner loop is used as the optimality measure for selecting the best design parameters. Note that with this cross-validation procedure, the test subject is not used in feature selection or classifier design.

III. Results

The performance of the proposed methodology for prediction of response to SSRI medication is indicated by the contingency table shown in Table I for $N_r = 8$ using an 11-fold nested cross-validation procedure. The average correct prediction performance is 86.6% (specificity=85.7%, sensitivity=87.5%).

TABLE I

Contingency table for the proposed prediction procedure.

	predicted NR	predicted R	% correct
actual NR	12	2	85.7%
actual R	1	7	87.5%

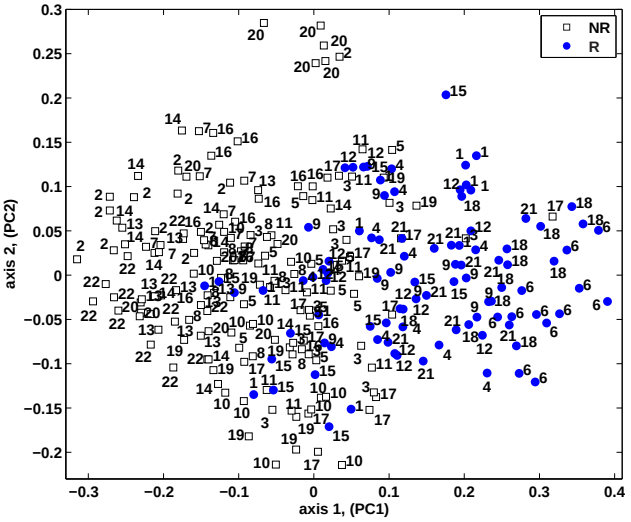

Fig. 1. Scatter plot of the projection of feature vectors from all available training epochs onto the first 2 major principal components, using KPCA. The numbers identify epochs belonging to each subject.

Fig. 1 shows a scatter plot of the $M_t = 262$ available training samples (epochs) in the N_r–dimensional feature space, projected onto two major components which are obtained using a kernelized principal component analysis (KPCA) method [15] with a Gaussian kernel. The patient index is written beside each data point. This figure shows a noticeable clustering of the data samples into two classes, although the clustering is not perfect.

A list of the most relevant discriminating features is shown in Table II. Columns 2 and 3 reflect the means and standard deviations of non-responder (NR) and responder (R) groups. A feature is listed in this table if it is used at least once throughout cross-validation procedure.

IV. Discussion and Conclusions

The performance of the proposed method suggests that suitably–selected features extracted from the EEG cluster according to how the patient responds to the treatment under consideration. Thus, the pretreatment EEG appears to contain information regarding brain functioning that is relevant to, and predictive of, the therapeutic effect of the SSRI antidepressant medication.

Our proposed feature selection process is novel in the respect that we have considered a large number of features

TABLE II

A LIST OF MOST DISCRIMINATING FEATURES, SHOWING THE MEAN AND STANDARD DEVIATION OF EACH FEATURE OVER THE NON-RESPONDER (μ_N, σ_N) AND RESPONDER GROUPS (μ_R, σ_R).

Selected Feature	μ_N, ($\pm$ σ_N)	μ_R, ($\pm$ σ_R)
Mutual Info. T3&T5	0.59 ($\pm$ 0.19)	0.37 ($\pm$ 0.18)
Mutual Info. T3&P3	0.56 ($\pm$ 0.19)	0.39 ($\pm$ 0.13)
Mutual Info. F4&T4	0.51 ($\pm$ 0.2)	0.33 ($\pm$ 0.14)
Coherence f=9Hz T3&T5	0.68 ($\pm$ 0.2)	0.42 ($\pm$ 0.2)
Coherence f=10Hz F7&C3	0.57 ($\pm$ 0.2)	0.38 ($\pm$ 0.23)
Coherence f=10Hz T3&T5	0.69 ($\pm$ 0.22)	0.40 ($\pm$ 0.22)
Coherence f=10Hz T3&P3	0.60 ($\pm$ 0.21)	0.37 ($\pm$ 0.2)
Coherence f=10Hz C3&T5	0.47 ($\pm$ 0.24)	0.29 ($\pm$ 0.17)
Coherence f=12Hz T3&T5	0.67 ($\pm$ 0.19)	0.40 ($\pm$ 0.21)
Coherence f=12Hz T3&O1	0.37 ($\pm$ 0.23)	0.20 ($\pm$ 0.14)
Coherence f=13Hz T3&T5	0.68 ($\pm$ 0.17)	0.42 ($\pm$ 0.2)
Coherence f=14Hz T3&T5	0.67 ($\pm$ 0.17)	0.41 ($\pm$ 0.2)
PSD-ratio f=14Hz Fp1/F3	-0.19 ($\pm$ 0.16)	0.01 ($\pm$ 0.23)
Coherence f=14Hz C3&T5	0.47 ($\pm$ 0.22)	0.28 ($\pm$ 0.18)
PSD-ratio f=14Hz Fp1/C3	-0.25 ($\pm$ 0.21)	-0.04 ($\pm$ 0.23)
Coherence f=15Hz T5&P3	0.68 ($\pm$ 0.15)	0.50 ($\pm$ 0.19)
Coherence f=16Hz T3&T5	0.67 ($\pm$ 0.19)	0.43 ($\pm$ 0.23)

including those, or similar ones, already cited in the literature, and reduced them, using the proposed feature selection procedure, into a small, maximally discriminative set. This is in contrast to the previous approaches, which hypothesize that a feature may be discriminative. An experiment is then required to verify or reject the hypothesis. Our method automatically identifies salient features without the need for experiment, thus saving considerable effort.

Our findings are consistent with the results of Cook et al [4], who found that absolute and relative power in all four EEG bands recorded pre-treatment over the regions implicated in mood disorders; i.e., prefrontal Fp1-Fp2-Fpz FC1-FC2-Cz, left temporal T3-T5; and right temporal T4-T6 are not significant predictors of response.

Several research groups have only used alpha power, considering an increase indicative of less cognitive neural activity, and that depressed patients had greater inter-hemispheric alpha asymmetries. They only considered inter-hemispheric asymmetries and found significantly greater overall pre-treatment alpha asymmetry with the right hemisphere more active in non-responders to fluoxetine (a different SSRI compared to this study) than responders and responders had significantly greater alpha in the occipital regions than non-responders and controls, and also greater right over left alpha asymmetry with non-responders having the opposite asymmetry [7].

It is interesting to note that our optimized feature selection process chose 14 of the 17 features from the alpha or low beta frequency bands. However, unlike other studies, e.g., [7], our method did not favour any inter-hemispheric power asymmetries. Instead, our results show that responders have more uniform alpha and low beta power anterior to posterior in the left hemisphere than non-responders who showed relatively greater posterior power.

As coherence and mutual information between several electrode pairs appear to be among the most highly predictive features (especially in the T3–T5 region), we might speculate that neural interaction or connectivity between the regions corresponding to the respective channels is highly relevant to SSRI response.

Because some of the features have strong statistical dependencies, the set of selected features in Table II is not unique. Some of the features may be replaced with others, with a small or no loss in performance. However, because of the inter-dependence of these features, a replaced feature set could be indicative of the same neurological information as the original and therefore likely correspond to closely related EEG electrodes and frequencies.

REFERENCES

[1] C. S. Dewa, A. Lesage, P. Goering, and M. Craveen, "Nature and prevalence of mental illness in the workplace," *Healthcare Papers*, vol. 5, no. 2, pp. 12–25, 2004.

[2] M. H. Trivedi, et al., "Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice," *American Journal of Psychiatry*, vol. 163, no. 1, pp. 28–40, Jan. 2006.

[3] D. C. Malone, "A budget-impact and cost-effectiveness model for second-line treatment of major depression," *Journal of Managed Care Pharmacy*, vol. 13, no. 6 (Suppl A), pp. S8–S18, 2007.

[4] I. A. Cook, et al., "Early changes in prefrontal activity characterize clinical responders to antidepressants," *Neuropsychopharmacology*, vol. 27, pp. 120–131, 2002.

[5] C. Mulert, G. Juckel, M. Brunnmeier, S. Karch, G. Leicht, R. Mergl, H.-J. Moller, U. Hegerl, and O. Pogarell, "Prediction of treatment response in major depression: integration of concepts," *Journal of Affective Disorders*, vol. 98, pp. 215–225, Mar. 2007.

[6] A. M. Hunter, I. A. Cook and A. F. Leuchter, "The promise of the quantitative electroencephalogram as a predictor of antidepressant treatment outcomes in major depressive disorder," *Psychiatric Clinics of North America*, vol. 30, no. 1, pp. 105–124, Mar. 2007.

[7] G. E. Bruder, J. P. Sedoruk, J. W. Stewart, P. J. McGrath, F. M. Quitkin and C. E. Tenke, "Electroencephalographic alpha measures predict therapeutic response to selective serotonin reuptake inhibitor antidepressant: pre- and post-treatment findings," *Biological Psychiatry*, vol. 63, pp. 1171–1177, 2008.

[8] H. Hinrikus, et al., "Electroencephalographic specral asymmetry index for detection of depression," *Medical and Biological Engineering and Computing*, vol. 47, pp. 1291–1299, 2009.

[9] V. Henkel, et al., "Does early improvement triggered by antidepressants predict response/remission? — Analysis of data from a naturalistic study on a large sample of inpatients with major depression," *Journal of Affective Disorders*, vol. 115, pp. 439–449, June 2009.

[10] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[11] R. Battiti, "Using mutual information for selecting features in supervised neural net learning", *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–550, July 1994.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. USA: John Wiley & Sons, 2006.

[13] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection Techniques*, C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, Eds. Lecture Notes in Computer Science, Springer, 2006, pp. 34–51.

[14] S. Varma and R. Simon, "Bias in error estmation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, no. 91, Feb. 2006.

[15] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, July 1998.