

Diagnosis of Psychiatric Disorders Using EEG Data and Employing a Statistical Decision Model

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

Abstract—An automated diagnosis procedure based on a statistical machine learning methodology using electroencephalograph (EEG) data is proposed for diagnosis of psychiatric illness. First, a large collection of candidate features, mostly consisting of various statistical quantities, are calculated from the subject’s EEG. This large set of candidate features is then reduced into a much smaller set of most relevant features using a feature selection procedure. The selected features are then used to evaluate the class likelihoods, through the use of a *mixture of factor analysis* (MFA) statistical model [7].

In a training set of 207 subjects, including 64 subjects with major depressive disorder (MDD), 40 subjects with chronic schizophrenia, 12 subjects with bipolar depression and 91 normal or healthy subjects, the average correct diagnosis rate attained using the proposed method is over 85%, as determined by various cross-validation experiments. The promise is that, with further development, the proposed methodology could serve as a valuable adjunctive tool for the medical practitioner.

I. INTRODUCTION

An essential criterion for successful treatment of a mental illness is a correct diagnosis. However, the diagnostic process can be a more difficult task than it may first appear. Even though clinical diagnostic guidelines are well established, i.e., through the *Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association* (DSM) and designed to differentiate various psychiatric conditions, often specific symptoms can appear in more than one diagnostic category, and diagnostic criteria can overlap to the point where confident differentiation is often impossible. Even the psychiatric expert can have difficulty distinguishing certain psychiatric conditions, e.g. psychotic depression from schizophrenia or, most notably, differentiating major depressive disorder (MDD) from bipolar depression (BD). This distinction is highly relevant as the antidepressant medications that would be quite appropriate for MDD may, in the patient with BD, induce mania, or rapid cycling between depression and mania thus making the condition considerably worse [1].

In summary, at best the consequence of a diagnostic error is that an ineffective treatment would be prescribed.

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).

However, even this best scenario is highly undesirable, since it adversely affects recovery of the patient. At worst, a diagnostic error results in exacerbation rather than the expected mitigation of symptoms. It is very important therefore that a correct psychiatric diagnosis be obtained before treatment is initiated. To this end, in this paper we propose an automated machine learning procedure based on the patient’s resting electroencephalogram (EEG) that can diagnose specific forms of psychiatric illness. The classes of psychiatric illness considered in this paper include schizophrenia (SCZ), MDD and BD (also referred to as bipolar affective disorder–depressed phase). A sample of healthy or ‘normal’ (N) subjects is also included. The proposed methodology could in principal be extended to other forms of illness; however, in this study we consider only these classes, due to the availability of training data.

There are previous studies based on the EEG for diagnosis. For example, the study [2] (based on a sample of 11 subjects with a history of depression and 11 normal control subjects) reported that frontal brain asymmetry (FBA) is a potential marker for depression. In [3], EEG data is analyzed to compare normal subjects versus subjects suffering various mental disorders. This group found that a decrease in delta or theta band EEG power can be regarded as a specific sign of brain dysfunction.

The machine leaning paradigm has also been used in this context. For example, see [4]–[6]. Li and Fan [4] used an artificial neural network (ANN) fed with EEG data to differentiate three classes of subjects: those with schizophrenia, those with depression, and healthy subjects. In a sample with 10 subjects in each of these three classes, the multi-layer perceptron neural network (trained with the back-propagation technique) obtained a correct diagnosis rate of 60%, 60%, and 80% for each of the three respective classes.

In this study, we demonstrate the use of statistical machine learning methodologies for psychiatric diagnosis. Specifically, we propose a maximum likelihood decision approach based on the use of *mixture of factor analysis* (MFA) models [7] for distinguishing between the four classes of diagnostic illness: SCZ, MDD, BD and N. On the basis of the available data, we show that the proposed approach to automated diagnosis offers a rather satisfactory level of performance.

II. METHODS

Resting or spontaneous EEG signals, denoted by $E_i, i = 1, \dots, M_t$ where M_t is the number of available training epochs, are collected from M subjects who suffer from one

of the $J = 4$ classes of psychiatric illness considered in this study. The true diagnostic class (as determined below) of the patient is denoted as y_i . The possible values for the y_i are [SCZ, MDD, BD and N]. The set of EEG recordings and the corresponding diagnostic class is referred to as a *training set*, denoted by $\mathcal{D} = \{(\mathbf{E}_i, y_i), i = 1, \dots, M_t\}$. These EEG signals from each subject are pre-processed to extract a large number N_c of candidate features $\tilde{\mathbf{x}}_i \in \mathbb{R}^{N_c}$ that might be relevant for diagnosis. These candidate features consist of various statistical quantities such as spectral coherence at all frequencies between 4Hz and 36Hz with 1Hz resolution and between all electrode pairs, as well as mutual information between all electrode pairs. In addition, the candidate features also include absolute and relative power spectral density (PSD) levels, the log ratio of left-to-right hemisphere powers, and anterior/posterior power ratios, between relevant electrodes and at all relevant frequencies. For the number of EEG electrodes and the frequency resolution values considered in this study, the number of candidate features is $N_c = 6988$. These candidate features are then reduced to a set of most relevant features $\mathbf{x}_i \in \mathbb{R}^{N_r}$, where $N_r \ll N_c$, to select only those features which are most statistically indicative of the diagnostic class. For this study, we choose N_r to be of the order of 10. We use the ‘‘greedy’’ feature selection method proposed by [9]. These reduced-dimensionality features are then fed into a classifier (in the form of a maximum likelihood decision model) that outputs the estimated diagnostic class.

A. Participants

Our study sample consists of a total of 207 adults, including 64 MDD, 40 SCZ, 12 BD and 91 N subjects. All 116 subjects with psychiatric conditions were recruited from the case load of the St Josephs Hospital, Center for Mountain Health Services, Hamilton, Ontario. They were carefully diagnosed using the appropriate DSM criteria by experienced psychiatrists specializing in the management of either mood disorders or schizophrenia. In most subjects with MDD, the diagnosis was confirmed using the Structured Clinical Interview for DSM (SCID). All subjects with schizophrenia met both DSM-IV criteria for schizophrenia and the Kane et al [10] criteria for treatment resistance.

B. EEG Recordings

In subjects with MDD or BD, resting EEG is measured after 10 days medication withdrawal. In subjects with schizophrenia, the subjects maintain their prescribed medication. Signals from sixteen EEG electrodes (configured according to the standard 10-20 system referenced to linked ears), consisting of Fp1, Fp2, F3, F4, F7, F8, T3, T4, C3, C4, T5, T6, P3, P4, O1 and O2, are recorded using a QSI-9500 EEG system, with a sampling frequency of 205Hz. For each subject, 6 EEG files each of 3.5 minutes duration are collected, consisting of 3 with eyes open and 3 with eyes closed; however, some subjects had only 5 files available. 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 38 Hz to partially mitigate the effects of eye movement and muscle artifacts. In each EEG file, in an effort to conserve computational demands, only the first 45 seconds of the de-artifacted portion of the EEG data are considered. These segments are divided into 2 overlapping epochs of 30 sec. duration, to give a nominal 12 epochs of EEG data per subject.

C. Classification using MFA Modeling

The MFA approach to classification [7] is based on maximum likelihood principles. Let the probabilistic event that a reduced feature vector \mathbf{x} belongs to the j -th class be denoted by Ω_j , where $j \in \{1, \dots, J\}$ and J is the number of classes. Assume that we have constructed the probability models for all classes, $\{p(\mathbf{x}|\Omega_j, \hat{\Theta}^{(j)}), j = 1, \dots, J\}$, where $\hat{\Theta}^{(j)}$ denotes the estimated parameters for the probabilistic generative model corresponding to the probabilistic event Ω_j . Based on the maximum likelihood classification rule, the class of a test feature vector \mathbf{x} is given by

$$\hat{y} = \arg \max_j p(\mathbf{x}|\Omega_j, \hat{\Theta}^{(j)}). \quad (1)$$

In this study, we separately estimate the statistical generative model of each class with MFA modeling using the corresponding training subset. MFA is based on ordinary factor analysis, which is a method for modeling correlated high-dimensional data with a low-dimensional oriented subspace. For the training data set of each class indexed by superscript j , the model assumes that each N_r -dimensional data vector $\mathbf{x}_i^{(j)}$ was generated by linearly transforming an $m < N_r$ dimensional vector of hidden independent zero-mean unit-variance Gaussian sources (factors) $\mathbf{z}_i = [z_{i1}, \dots, z_{im}]$, as follows

$$\mathbf{x}_i^{(j)} = \mathbf{A}^{(j)} \mathbf{z}_i + \boldsymbol{\mu}^{(j)} + \mathbf{n}_i, \quad i = 1, \dots, M_{t_j} \quad (2)$$

where M_{t_j} is the number of training samples in class j , $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$, and $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}; \boldsymbol{\Psi}^{(j)})$, where $\mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\Sigma})$ denotes a multi-variate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The matrix $\mathbf{A}^{(j)} \in \mathbb{R}^{N_r \times m}$, is a linear transformation known as the ‘factor loading matrix’ for the j th class, and $\boldsymbol{\mu}^{(j)}$ is the mean vector of the analyzer. The factor analysis methodology discussed above invokes a linear model, but in practice the data manifold generated by the set of reduced feature vectors is generally nonlinear. To address this issue, a mixture of factor analysis (MFA) model is used. The MFA procedure models the density for a data point \mathbf{x}_i as a weighted average of factor analyzer densities. Let us assume that for class j , the number of mixture components is K , the factor loading matrix for component k is denoted by $\mathbf{A}^{(j,k)}$, and $\boldsymbol{\mu}^{(j,k)}$ is the corresponding mean vector. The distribution for $\mathbf{x}_i^{(j)}$ is therefore modelled as

$$p(\mathbf{x}_i^{(j)}|\Omega_j, \Theta^{(j)}) = \sum_{k=1}^K \alpha_k^{(j)} p(\mathbf{x}_i|s_k^{(j)}, \mathbf{A}^{(j,k)}, \boldsymbol{\mu}^{(j,k)}, \Psi^{(j)})$$

where $\alpha_k^{(j)} = p(s_k^{(j)}|\alpha^{(j)})$ is the mixing proportion, $s_k^{(j)} \in \{1, 2, \dots, K\}$ represents a discrete random variable indicating the component from which $\mathbf{x}_i^{(j)}$ has been generated, $\alpha^{(j)} = [\alpha_1^{(j)}, \dots, \alpha_K^{(j)}]^T$ such that $\sum_{k=1}^K \alpha_k^{(j)} = 1$ and $\alpha_k^{(j)} > 0$. Based on the MFA model, the parameter set corresponding to the j -th class of data is $\Theta^{(j)} = \left\{ \{\mathbf{A}^{(j,k)}, \boldsymbol{\mu}^{(j,k)}\}_{k=1}^K, \boldsymbol{\alpha}^{(j)}, \Psi^{(j)} \right\}$. These parameters are estimated, using the information in the training data, by an *expectation-maximization* (EM) algorithm [7]. The number of mixture components (K) and number of factors (m) are assumed equal over all classes. The best values for these quantities are found using a nested cross-validation procedure.

After the model is trained, the likelihood that a test data vector \mathbf{x} belongs to j -th diagnostic class can be shown to be

$$p(\mathbf{x}|\Omega_j, \hat{\Theta}^{(j)}) = \sum_{k=1}^K \hat{\alpha}_k^{(j)} \mathcal{N}(\mathbf{x}|\hat{\boldsymbol{\mu}}^{(j,k)}; \mathbf{A}^{(j,k)} \mathbf{A}^{(j,k)T} + \hat{\Psi}^{(j)}) \quad (3)$$

The estimated class is then determined according to (1).

Finally, considering the fact that we have multiple epochs (values of \mathbf{x}_i) for each subject, the diagnosis result for each subject is determined by averaging the MFA probabilities from (3) over all corresponding epochs before quantization.

To evaluate the diagnosis performance, we used an iterative ‘leave- n -out’ (L n O) cross-validation (CV) procedure, as follows. First, subjects are divided into contiguous subsets, each consisting of n subjects. Then, in each fold (iteration) of the cross-validation loop, all epochs associated with a specific subset are omitted from the training set. The remaining data are used to train the diagnosis model. The model is tested using the omitted subset. The process repeats, where in each iteration each respective subset is omitted, until all subjects have been tested once. In this study, the value of n varied with the number of training samples in the respective experiment, so that the number of folds remains constant at a value of approximately 10 iterations.

III. RESULTS

Three diagnostic experiments are studied in this paper and the results are shown in Tables I–III. We show results from selected two-way diagnostic classifications in addition to a multi-class diagnosis result. Experiment 1 is a two-way diagnosis test between the classes SCZ and MDD. The results of the experiment are contained in the contingency table shown in Table I. The first row of this table represents the 64 patients who have been assumed correctly diagnosed by the psychiatrist to have MDD. The diagonal entry of this row (first column) indicates the number of correct diagnoses made by the proposed method, whereas the second entry gives the number of cases incorrectly diagnosed as SCZ. Row 2 represents the SCZ performance in a corresponding manner. As may be seen, the overall correct diagnosis rate is in excess of 88%.

Experiment 2 is an additional binary classification example, this time between MDD and BD subjects. These

TABLE I
EXPERIMENT 1: BINARY DIAGNOSIS PERFORMANCE RESULTS FOR
MDD VS. SCHIZOPHRENIA WITH $N_r = 14$.

true	Estimated as MDD	Estimated as SCZ	Total No.
MDD	57 (89%)	7	64
SCZ	5	35 (87.5%)	40
avg. = 88.3%			104

TABLE II
EXPERIMENT 2: BINARY DIAGNOSIS PERFORMANCE RESULTS FOR
MDD VS. BD WITH $N_r = 8$.

true	Estimated as MDD	Estimated as BD	Total No.
MDD	60 (93.8%)	4	64
BD	4	44 (91.7%)	12 × 4
avg. = 92.7%			76

two conditions are very difficult to distinguish, and even in the absence of a past history of an episode of mania or hypomania, impossible to differentiate in a clinical setting. Due to the imbalance in the number of training samples in the BD and MDD groups (12 and 64 samples respectively), the classification procedure becomes biased toward the majority population, reducing the level of performance. To avoid this difficulty we used the following procedure. We divided the 64 MDD subjects into 4 subsets each of size 16 subjects. Then 4 separate diagnosis experiments are performed where in each experiment each of the MDD subsets are sequentially tested against the 12 subjects with BD. The overall contingency table is then constructed by adding the respective entries from the individual tables obtained from each experiment, as shown in Table II. Again, the combined performance level of approximately 92% offers promising potential for the proposed method. It is to be noted that this experiment was repeated using the linear discriminant analyzer (LDA), the nearest-neighbor (NN) and support vector machine (SVM) classifiers [8]. The performance of the MFA classifier exceeded that of the others by a margin of at least 2.5%.

In experiments 1 and 2, the values for K and m of the MFA model were determined in each fold of the cross-validation procedure from the candidate sets $[1, \dots, 5]$ and $[1, \dots, 4]$, respectively.

We now investigate the clustering behaviour of the feature space for the MDD vs. BD classification example (Experiment 2). In this vein, we construct a two-dimensional representation of the feature space, compressed from N_r dimensions to 2 dimensions using the Kernel PCA (KPCA) method (with a Gaussian kernel) [11]. To balance the number of subjects in each diagnostic group, as previously discussed, we arbitrarily chose only the first subset of 16 MDD subjects for this clustering analysis. Fig. 1 shows a scatter plot of 336 feature vectors corresponding to the 336 available EEG epochs projected onto the first two major nonlinear principal

IV. DISCUSSION AND CONCLUSIONS

In conclusion, the proposed EEG-based methodology, consisting of the feature selection method of Peng [9] and the MFA classification procedure [7] is found to be very efficient for diagnosis of psychiatric disorders. The superior performance of the MFA method for this application, in comparison to other forms of classifier, is very likely due to its ability to model a low-dimensional nonlinear manifold using a combination of linear components. Furthermore, the proposed method outputs a soft decision in the form of a likelihood statistic for each of the classes, as opposed to a hard decision as in the case of other common forms of classifier. This provides the clinician with the likelihood of occurrence of each of the illnesses, for a given patient. This can be of value e.g., in prescribing treatment in the case of a co-morbid illness.

Findings such as these suggest that machine learning may find an important place in the tool chest of the medical practitioner, particularly when experienced psychiatric personnel are not readily available. Confirmation of diagnosis may permit the clinician to initiate appropriate treatment while awaiting expert psychiatric assessment, which, even in urban areas, may not be available for weeks or months.

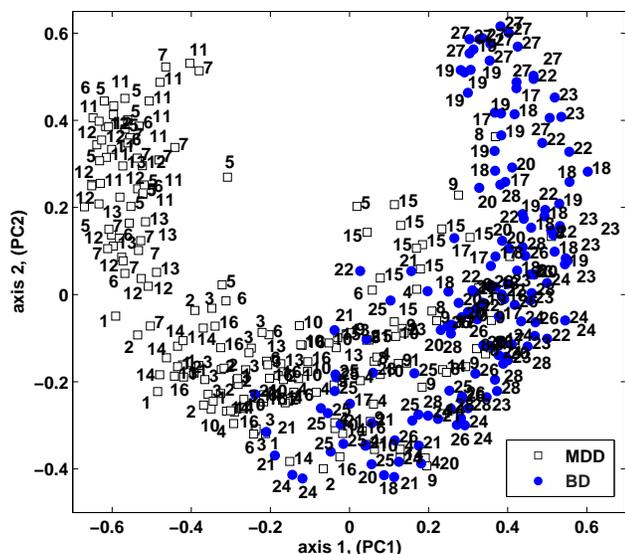


Fig. 1. From experiment 2: Scatter plot of the projection of the selected feature vectors onto the first 2 nonlinear principal components, obtained using the KPCA method. Clustering between the BD versus MDD features is evident, which suggests discrimination between these diagnostic classes is possible. $N_r = 8$.

TABLE III

EXPERIMENT 3: THREE-CLASS DIAGNOSIS PERFORMANCE RESULTS: MDD VS. SCZ VS. N.

(true)	Est. as MDD	Est. as SCZ	Est. as N	Total No.
MDD	55 (85.9%)	6	3	64
SCZ	3	35 (87.5%)	2	40
N	4	7	80 (87.9%)	91
			avg. = 87.1%	195

components. The patient index is written beside each data sample. It may be observed that this figure shows a noticeable clustering of the subjects into the MDD and BD groups.

Finally, in experiment 3, a three-class diagnosis problem between the classes MDD, SCZ and N is studied. Table III shows the results for $N_r = 42$. Because the training set for the BD case is imbalanced, and the diagnostic performance for the BD class is consequently degraded, we have omitted this class from this example. A diagnostic capability for the BD case could nevertheless be obtained by combining e.g., the MDD and BD classes into one and then performing a 3-way classification with the remaining classes as shown in this experiment. A subsequent binary classification as in experiment 2 could then distinguish between the MDD and BD classes. Note that the features are determined in this multi-class case by finding separate lists of discriminating features for all binary combinations of diagnosis class. These features are then concatenated into a single larger list, to be used in the multi-class classifier.

REFERENCES

- [1] V. Kusumakar, "Antidepressants and antipsychotics in the long-term treatment of bipolar disorder," *Journal of Clinical Psychiatry*, vol. 63, Suppl 10, pp. 23–28, 2002.
- [2] A. J. Niemiec and B. J. Lithgow, "Alpha-band characteristics in EEG spectrum indicate reliability of frontal brain asymmetry measures in diagnosis of depression," in *Proceedings Int. Conf. of the IEEE Eng. in Medicine and Biology Society*, pp. 7517–7520, Sept. 2005.
- [3] P. Coutin-Churchman, et al., "Quantitative spectral analysis of EEG in psychiatry revisited: drawing signs out of numbers in a clinical setting," *Clinical Neurophysiology*, vol. 114, no. 12, pp. 2294–2306, Dec. 2003.
- [4] Y.-J. Li and F.-Y. Fan, "Classification of Schizophrenia and depression by EEG with ANNs," in *Proceedings Int. Conf. of the IEEE Eng. in Medicine and Biology Society*, pp. 2679–2682, Sept. 2005.
- [5] S. Ghosh-Dastidar, H. Adeli and N. Dadmehr, "Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection," *IEEE Trans. Biomedical Engineering*, vol. 55, no. 2, pp. 512–518, Feb. 2008.
- [6] I. Güler and E. D. Übeyli, "Multiclass support vector machines for EEG-signals classification," *IEEE Trans. Information Technology in Biomedicine*, vol. 11, no. 2, pp. 117–126, Mar. 2007.
- [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] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Academic Press, 2008.
- [9] 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.
- [10] J. Kane, G. Honigfeld, J. Singer, H. Meltzer and the Clozaril Collaborative Study Group, "Clozapine for the treatment-resistant schizophrenic: A double-blind comparison with chlorpromazine," *Archives of General Psychiatry*, vol. 45, pp. 789–796, Sept. 1988.
- [11] K. R. Müller, S. Mika, G. Rätsch, K. Tsuda and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Networks*, vol. 12, no. 2, pp. 181–201, Mar. 2001.