35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Classification of Schizophrenia using Genetic Algorithm-Support Vector Machine (GA-SVM)

Ming-Hsien Hiesh, Yan-Yu Lam Andy, Chia-Ping Shen, *Student Member*, *IEEE*, Wei Chen, Feng-Shen Lin, Hsiao-Ya Sung, Jeng-Wei Lin, Ming-Jang Chiu, and Feipei Lai, *Senior Member, IEEE*

Abstract— **Recently, Event-Related Potential (ERP) has being the most popular method in evaluating brain waves of schizophrenia patients. ERP is one of the electroencephalography (EEG), which is measured the change of brain waves after giving patients certain stimulations instead of resting state. However, with traditional statistical analysis method, both P50 and MMN showed significant difference between controls and patients but not in Gamma band. Gamma band is a 30-50 Hz auditory stimulation which had been suggested may be abnormal in schizophrenia patients. Our data are recruited from 5 schizophrenia patients and 5 controls in National Taiwan University Hospital have been tested with this platform. The results showed that detection rate is 88.24% and we also analyzed the importance of features, including Standard Deviation (SD) and Total Variation (TotalVar) in different stage of wavelet transform. Therefore, this proposed methodology could serve as a valuable clinical decision support for physiologists in evaluating schizophrenia.**

I. Introduction

Schizophrenia is a severe mental illness which afflicts approximately 1% of the general population [1]. Schizophrenia is a chronic and disabling mental disorder, affecting memory, attention and executive function in human [2]. Moreover, the precise illness mechanisms underlying is still poorly understood. Therefore, electroencephalogram (EEG) signals analysis may be helpful for the scientists to acknowledge the mechanisms of such disorder. Analysis of time series may provide an insight into some new EEG patterns in schizophrenia. Schizophrenia EEG can indicate underlying synchronous neural activity of the brain. Several methods have been devised for detecting schizophrenic patient base on EEG analysis. However, the quantitative EEG analysis still remains a challenging task because of the complexity and irregular nature of brain activity. In the past, because of the limitations of statistical software and computer power, only 1% of schizophrenia EEG data had been used. It means only 1% or less information had been extracted or learned from these valuable data. Recently, with the growth of computing power, there are many cloud computing techniques have been applied to this area [3-5].

Various approaches and models providing classification of schizophrenia have been proposed. Aviyente used dynamic Bayesian network to measure effective brain connectivity [6]. A. Khodayari-R. proposed using EEG data and employing a statistical decision model for automated diagnosis procedure [7]. Y. J. Li had proposed using Artificial Neural Network (ANN) for classification of schizophrenia and depression patients [8]. M. Liu employed Principal Components Analysis (PCA) with nonlinear SVM [9] and Alba-Sanchez, F. utilized Self-Organizing Maps (SOM) in pattern recognition of mental disorder disease [10] . In this paper, we proposed a system built with wavelet transform (WT), Genetic Algorithm (GA) and Support Vector Machine (SVM) to analyze Gamma band synchronization test data. The system architecture is shown as Fig. 1, including four parts: data pre-processing, feature extraction, feature selection and classification.

P50, Mismatch Negativity (MMN) and Synchronous Gamma Activity (Gamma band) are 3 commonly adopted ERP in schizophrenia research. Data used in this research is the Gamma band synchronization test, which is giving subjects 3 types of auditory stimulations (20Hz, 30Hz and 40Hz) continuously for 30 minutes, 10 minutes for each frequency [11]. Gamma band synchronization test is one of Event-Related Potential (ERP), which measures the differences of brain waves before and after giving stimulations, instead of measuring EEG under resting state. Unlike P50 and Mismatch Negativity (MMN), another 2 widely adopted methods in schizophrenia research will take over half an hour in recording EEG, Gamma band synchronization test could be accomplished in 10 minutes. In previous study, 40Hz has been suggested may be abnormal in schizophrenia patients (Fig. 2) [12]. With our platform, we are looking into further findings of the Gamma band synchronization test, including whether the time of the test could be shorter, quantify the level of schizophrenia patients.

In this study, we demonstrated a system built with Genetic Algorithm (GA) and Support Vector Machine (SVM), which analyzed Gamma band synchronization EEG data from schizophrenia patients. With further development, this system could help automated quantifying diagnosis of schizophrenia,

H. S. Hsieh is with the Department of Psychiatry, National Taiwan University Hospital, Taipei, Taiwan.

Y. Y. Lam, C. P. Shen, W. Chen and F. P. Lai are with the Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan.

F. S. Lin and H. Y. Sung are with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

J. W. Lin is with the Department of Information Management, Tunghai University, Taichung, Taiwan.

M. J. Chiu is with the Department of Neurology, National Taiwan University Hospital, Taipei, Taiwan.

T. M. Liu is with the Institute of Biomedical Engineering, National Taiwan University, Taipei, Taiwan.

Correspondence to Mr. Feng-Shen Lin, No. 1, Sec. 4, Roosevelt Rd., Daan District, Taipei City, Taiwan; TEL: 886-958775761; E-mail: r98922096@ntu.edu.tw;

by linking EEG data with genetic data, demographic data and family/trio study.

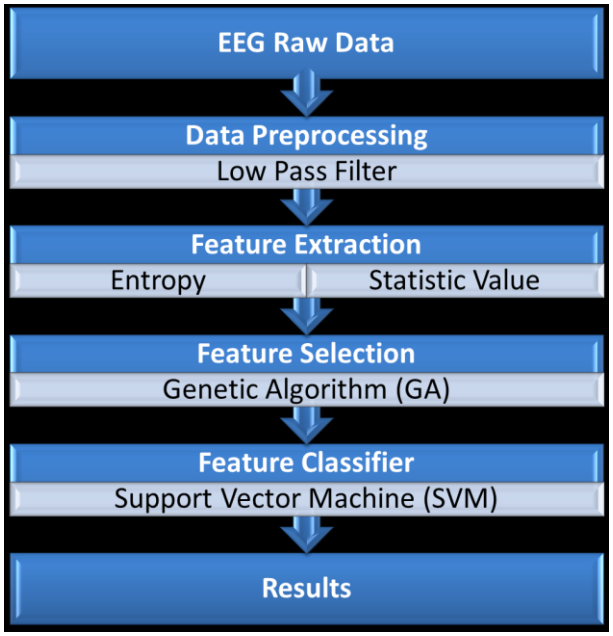

Figure 1. System architecture

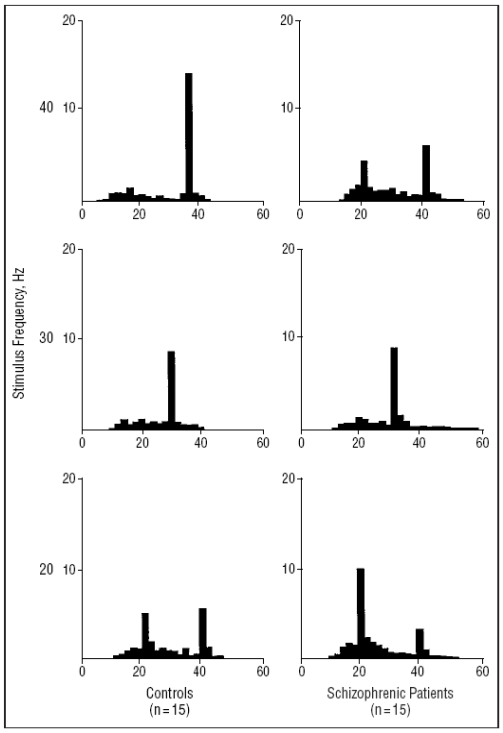

Figure 2. 40Hz has been suggested may be abnormal in schizophrenia patients

II. Materials and Methods

A. Data Acquisition and Preprocessing

We collected 5 controls and 5 schizophrenia patients from National Taiwan University Hospital who underwent Gamma band synchronization by giving them 20Hz, 30Hz and 40Hz auditory stimulations for 10 minutes (Fig. 3). The collected data were segmented into 2-second fragment, which is following by a 4th-order wavelet transform (Fig. 4).

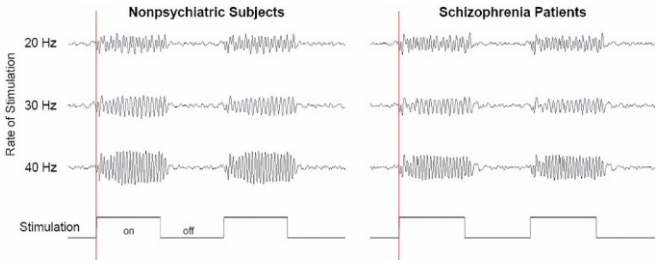

Figure 3. Auditory stimulations in Gamma band synchronization test

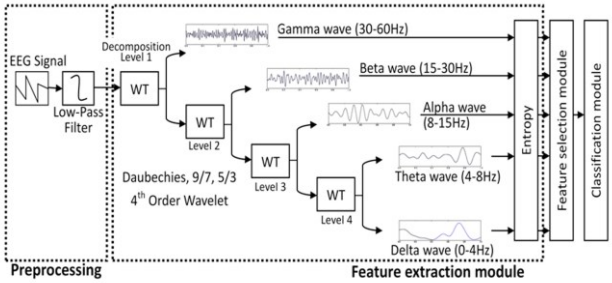

Figure 4. From preprocessing to feature extraction

B. Wavelet Transform

The four steps of wavelet transform (WT), which was applied with Daubechies, tap 5/3, and tap 9/7 filter pairs, decomposing the signal into high and low frequency components [13]. The *g* indicated the high-pass filter, and *h* indicated the low-pass filter. The results stored in low-pass channels {A1, A2, A3, and A4} and in high-pass channels {D1, D2, D3, and D4}. These results then extracted by the values of their mean, standard deviation, maximum, minimum and entropy.

C. Feature Extraction

All 16 channels EEG data were calculated to obtain features. Sample Entropy (SampEn) is usually used in quantifying the regularity of time-series data, which represented as a simple index for the overall complexity and predictability of each time series. More regular data shows lower value of SampEn index.

We also calculate the maximum, minimum, mean, and standard deviation for all of the nodes in the EEG fragments. These parameters were further used to extract total variation, standard deviation, sample entropy, skewness [14], and energy as statistical features.

D. Feature Selection

Genetic Algorithm (GA) is used as feature selection in our study. GA could generate optimizations and solutions to search problems [15-16]. The main concept of GA is to

simulate the selection processes happened in natural, including mating, mutations, inheritance, crossover and selection. GA starts with creating a group of randomly generated individuals represented as *0*s and *1*s, followed by mutation, recombination, and mating. The fitness of each generation are evaluated and only the fittest individual will be selected and survive in the next generation. Implementing this concept on solving search problems, we can have the fitness of parameters and determines optimal parameters. These selected features will be used in the following Support Vector Machine (SVM) classification.

The GA can be summarized in four steps: (1) it randomly generates the initial chromosomes population (M_1); (2) the fitness u(p, q, g) is computed for each chromosome g in current population M_k, where u(p, q, g) is the accuracy of the SVM(p, q, g) classification, and used the corresponding features indicated by g; (3) selecting good chromosomes from M_k to generate offspring M_{k+1} using genetic operators, where the proportion of u(p,q,g) and selection probability for a chromosome g in M_k is designed, and (4) returns to Step 2 until obtaining the satisfactory condition.

When the GA terminates, the most accurate SVM classification is chosen, as shown in the Equation (1).

$$\Omega \boldsymbol{p,q} = \boldsymbol{\Omega p,q,g^*}, \text{ where } \boldsymbol{g^*} = \arg \boldsymbol{g} \max(\{\boldsymbol{u(p, q, g)}\}) \quad (1)$$

If the GA evolves R generations, the number of standard SVM invocations is $|M_1|*|M_2|*\ldots*|M_R|$, where $|M_k|$ is the number of chromosomes in M_k.

E. Classification

The SVM realize a linear classification system by mapping the inputted feature vectors into a high-dimensional space [17]. A training set of instance-label pairs ($\boldsymbol{x}$, y) and weighting vector w can be written as Equation (2), which $\boldsymbol{x}$ and y denote the input and output domains respectively [18].

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & y_i \left(w^T x_i + b \right) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \quad (2)$$

The penalty term C in Equation (2), is chosen by user to assign a penalty to errors, and ξ is a slack variable. In order to construct models, SVM maps the instances into high-dimensional domains to separate training data, and the number of variables becomes large, or even infinite. Equation (3) is a typical method to addressing this problem.

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) - \sum_{i=1}^{l} \alpha_i \\ \text{subject to} \quad & 0 \leq \alpha_i \leq c, \qquad i = 1, \ldots, l \\ & \sum_{i=1}^{l} y_i \alpha_i = 0 \end{aligned} \quad (3)$$

III. Results and Discussion

A. Features Extracted

Total 800 features have been extracted from all 16 channels EEG data and is shown in Table I. Each of total variation, standard deviation, sample entropy, skewness and energy contains 160 features. We also analyzed the importance of features by linear SVM, and the result is shown as Table II. These features are selected from wavelet transform Stage 1 (Gamma band; 30-60Hz), Stage 2 (Beta band; 15-30Hz), Stage 3 (Alpha band; 8-15Hz) and Stage 5 (Delta band; 0-4Hz). All of these features weight over 50% in our analysis.

TABLE I. Feature List

Total variation	Standard deviation	Sample entropy	Skewness	Energy
160	160	160	160	160

TABLE II. Important Features

Feature name	Stage	Weight
Standard deviation	WT (Stage 3)	24.02%
Standard deviation	WT (Stage 5)	7.89%
Total variation	WT (Stage 1)	6.59%
Skewness	WT (Stage 2)	6.34%
Standard deviation	WT (Stage 2)	5.64%

B. Performances of Classification

The performances of classifications are evaluated for three parameters, namely, sensitivity, specificity, and accuracy. A sample is one of EEG signal. The definitions of sensitivity, specificity, and accuracy are described as follows:

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)$$

where TP represents the number of true positives, FP represents the number of false positives, TN represents the number of true negatives, and FN represents the number of false negatives.

We trained and predicted 5000 segments of 2-second fragment after 40Hz auditory stimulation. Our proposed system performs 89.48% sensitivity and 87% specificity. Furthermore, we obtained 87% and 89.48% accuracy in control group and patient group respectively, and overall 88.24% accuracy of classification. The comparison of our performance with previous studies is shown in Table 3.

TABLE III. Comparsion of Classification Methods

Methods	Sensitivity	Specificity	Accuracy
GA-SVM	89.48%	87%	88.24%

Methods	Sensitivity	Specificity	Accuracy
Artificial Neural Network [8]	N/A	N/A	60~80%
Self-Organizing Maps [10]	N/A	N/A	70%

IV. Conclusion

Schizophrenia is a severe mental illness which afflicts approximately 1% of the general population. Schizophrenia is a chronic and disabling mental disorder, affecting memory, attention and executive function in human. Currently, P50, MMN and Gamma band synchronization test are commonly used Event-Related Potential (ERP) in schizophrenia research. ERP measures the differences of brain waves before and after giving subjects stimulations. The 40Hz auditory stimulation in Gamma band synchronization test had been suggested may be abnormal in schizophrenia patients in previous study. However, in our experience, it is difficult to distinguish controls and patients with traditional statistical methods.

In this paper, we demonstrated a system built with Wavelet Transform (WT), Genetic Algorithm (GA) and Support Vector Machine (SVM) to classify the results of Gamma band synchronization test. This system includes four steps: (1) data preprocessing, (2) feature extraction, (3) feature selection and (4) classification.

The input data are segmented into 2-second fragments of brain waves after giving subjects 40Hz auditory stimulation, which is following by a 4^{th}-order WT and we calculate the maximum, minimum, mean, and standard deviation for all of the nodes in the segments in this step. These values will then further used to extract total variation, standard deviation, sample entropy, skewness, and energy as statistical features. GA and SVM are feature extraction and classification respectively.

Our proposed system performs 89.48% sensitivity and 87% specificity. Furthermore, we obtained 87% and 89.48% accuracy in control group and patient group respectively, and overall 88.24% accuracy of classification. Furthermore, we also analyzed and sorted the importance of features with linear SVM. The most important features including standard deviation in WT stage 2/3/5, skewness in WT stage 2 and Total Variation (TotalVar) in WT stage 1. In addition, all of these features weight over 50% in our analysis.

With further development of this system and joined the recruited EEG data with genome data, demographic data and family/trio study, this proposed methodology could serve as a valuable clinical decision support tool and references for physiologists in schizophrenia evaluation.

Acknowledgment

The authors would also like to thank Prof. Chih-Jen Lin and his research team members for providing the LIBSVM tool.

References

[1] B.S. Raghavendra, and D.N. Dutt. "A Study of Long-range Correlations in Schizophrenia EEG using Detrended Fluctuation Analysis", *Signal Processing and Communications*, pp. 1-5, 2010.

[2] P.K. McGuire and C.D. Frith, "Disordered functional connectivity in schizophrenia" *Psychol. Med.*, vol. 26, pp. 663-667, 1996.

[3] C.P. Shen, "A Multiclass Classification Tool Using Cloud Computing Architecture," *International Symposium on Network Enabled Health Informatics, Biomedicine and Bioinformatics (HI-BI-BI 2012),* Istanbul , June. 2012.

[4] C.P. Shen, "Detection of cardiac arrhythmia in electrocardiograms using adaptive feature extraction and modified support vector machines," *Expert Systems With Applications*, vol. 39, no. 9, pp. 556 - 561, July 2012.

[5] C.P. Shen, "Bio-signal Analysis System Design with Support Vector Machines based on Cloud Computing Service Architecture" in *Proc. IEEE Engineering in Medicine and Biology Society (EMBS)*, Argentina, Aug. 2010, pp. 1421-1424.

[6] A.Y. Mutlu, "Inferring Effective Connectivity in the Brain from EEG Time Series Using Dynamic Bayesian Networks", *31st Annual International Conference of the IEEE EMBS*, Minnesota, September 2009.

[7] A. Khodayari-R., "Diagnosis of Psychiatric Disorders Using EEG Data and Employing a Statistical Decision Model", *32nd Annual International Conference of the IEEE EMBS*, Argentina, September 2010.

[8] Y.J. Li, "Classification of Schizophrenia and Depression by EEG with ANNs", 27^{th} *IEEE Engineering in Medicine and Biology Annual Conference*, Shanghai, September 2005.

[9] M. Liu, "A Study of Schizophrenia Inheritance through Pattern Classification", 2^{nd} *International Conference on Intelligent Control and Information Processing*, Changsa, 2011.

[10] Alba-Sanchez F., "Assisted Diagnosis of Attention-Deficit Hyperactivity Disorder through EEG Bandpower Clustering with Self-Organizing Maps", *32nd Annual International Conference of the IEEE EMBS*, Argentina, September 2010.

[11] J.S. Kwon, "Gamma Frequency–Range Abnormalities to Auditory Stimulation in Schizophrenia", *Arch Gen Psychiatry*., 56(11):1001-1005, 1999.

[12] G.A. Light, "Gamma Band Oscillations Reveal Neural Network Cortical Coherence Dysfunction in Schizophrenia Patients", *Biol. Psychiatry*, vol. 60, pp. 1231–1240, 2006.

[13] A. Cohen, "Bi-othogonal bases of compactly supported wavelets," *IEEE Trans. Communication on Pure Applied mathematics*, vol. 45, pp. 485-560, 1992.

[14] http://en.wikipedia.org/wiki/Skew_distribution

[15] J.H. Holland, "Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence", *MIT Press*, 1992.

[16] J. Yang, "Feature subset selection using a genetic algorithm", *Intelligent Systems and Their Applications*, vol. 13, pp. 44-49, 1998.

[17] C. Cortes, *Support-vector network*, 1995.

[18] C.C. Chang, "LIBSVM: a library for support vector machines", 2001, software is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm