Relevance Vector Machine Applied to EEG Signals Classification

Sandro Chagas, Marcio Eisencraft, Clodoaldo Ap. M. Lima

Abstract—**The electroencephalogram (EEG) is a complex and aperiodic time series, which is a sum over a very large number of neuronal membrane potentials. Despite the rapid advances of neuroimaging techniques, EEG recording continues playing an important role in both the diagnosis of neurological diseases and understanding of the psychological process. In order to extract relevant information of brain electrical activity, a variety of computerized-analysis methods have been used. In this paper, we propose the use of a recently developed machine-leaning technique – relevance vector machine (RVM) – for EEG signals classification. RVM is based on Bayesian estimation theory, which has as distinctive feature the fact that it can yield a sparse decision function defined only by a very small number of so-called relevance vectors. From the experimental results, we can see that estimation and classification based on RVM perform well in EEG signals classification problem compared with traditional approach support vector machine (SVM), which indicates that this classification method is valid and has promising application.**

Keywords - **Electroencephalogram, Support Vector Machine, Relevance Vector Machine, Bayesian estimation theory, classification problem**

I. INTRODUCTION

The human brain is a complex system, and exhibits rich spatiotemporal dynamics. Among the techniques for investigating human brain dynamics, electroencephalography (EEG) provides a non-invasive and direct measure of cortical activity with temporal resolution in milliseconds. EEG is a record of the electrical potentials generated by the cerebral cortical neurons. Early on, EEG analysis was restricted to visual inspection of EEG records. Since there is no definite criterion evaluated by the experts, visual analysis of EEG signals is insufficient. Routine clinical diagnosis requires the analysis of EEG signals. Therefore, some automation and computer techniques have been used for this aim [5]. Since the early days of automatic EEG processing, representations based on a Fourier transform have been most commonly applied. This approach is based on earlier observations that the EEG spectrum contains some characteristic waveforms that fall primarily within four primary components. Such methods have proved beneficial for various EEG characterizations, but fast Fourier transform (FFT), suffers from large noise sensitivity. Parametric power spectrum estimation methods such as AR, reduces the spectral loss problems and gives better frequency resolution. Also AR method has an advantage over FFT that, it needs shorter duration data records than FFT [20].

In the late 1890s, a powerful method was proposed to perform time-scale analysis of signals: the wavelet transforms (WT). This method provides a unified framework for different techniques that have been developed for various applications. Since the WT is appropriate for analysis of non-stationary signals and this represents a major advantage over spectral analysis, it is well suited to locating transient events. Wavelet's feature extraction and representation properties can be used to analyze various transient events in biological signals. In [1] was presented an overview of the discrete wavelet transform (DWT) developed for recognizing and quantifying spikes, sharp waves and spike-waves. Through, wavelet decomposition of the EEG records, transient features are accurately captured and localized in both time and frequency context.

Various other techniques from the theory of signal analyses have been used to obtain representations and extract the features of interest for classification purposes. Neural networks and statistical pattern recognition methods have been applied to EEG analysis. In [11] was used the raw EEG data as an input to a neural network while in [19] was used the features proposed by Gotman with an adaptive structure neural network, but his results show a poor false detection rate. In [10] a recurrent neural network combined with wavelet pre-processing was proposed to predict the onset of epileptic seizures both on scalp and intracranial recordings only one-channel of electroencephalogram.

However, most of the techniques used to train the neural network classifiers are based on the idea of minimizing the training error, which is usually called empirical risk. As a result, limited amounts of training data and over high training accuracy often lead to over training instead of good classification performance. In addition, its classification accuracy is also sensitive to the dimension of the training set.

On the other hand, the Support Vector Machine (SVM) approach is based on the minimization of the structural risk [18], which asserts that the generalization error is delimited by the sum of the training error and a parcel that depends on the Vapnik-Chervonenkis dimension. By minimizing this summation, high generalization performance may be obtained. Besides, the number of free parameters in SVM does not explicitly depend upon the input dimensionality of the problem at hand. Another important feature of the support vector learning approach is that the underlying optimization problems are inherently convex and have no local minima, which comes as the re-

Sandro Chagas, Marcio Eisencraft, Clodoaldo Ap. M. Lima, Escola de Engenharia, Universidade Presbiteriana Mackenzie, São Paulo, Brasil, E-mails: schagas@gmail.com, marcioft@mackenzie.br, moraes@mackenzie.br.

sult of applying Mercer's conditions on the characterization of kernels [13].

Although the SVM classification provides successful results, a number of significant and practical disadvantages are identified as follows [15][16]:

- Although SVMs are relatively sparse, the number of support vector (SVs) typically grows linearly with the size of the training set and therefore, SVMs make unnecessarily liberal of basic functions.
- Predictions are not probabilistic, and therefore, SVM is not suitable for classification tasks in which posterior probabilities of class membership are necessary.
- In SVM, it is required to estimate the error/margin trade tradeoff parameter C, which generally entails a cross-validation procedure which can be a waste of data as well as computation.
- In SVM, the kernel function must satisfy Mercer' condition, hence, it must be a continuous symmetric kernel of a positive integral operator.

The Relevance Vector Machine (RVM) has been introduced by [15][16] as a Bayesian treatment alternative to the SVM that does not suffer from the aforementioned limitations. The RVM is a statistical learning method and it is a probabilistic sparse kernel model identical in functional form to the SVM. It represents a new approach to pattern classification that has recently attracted a great deal of interest in the machine learning community. RVM can be seen as a new way to train polynomial, neural network, or Radial Basic Functions classifiers. It operates on the induction principle of structural risk minimization, which minimizes an upper bound on the generalization error. Because in this approach no probability density is estimated, it becomes highly insensitive to the curse of dimensionality. In many problems RVM classifiers have been shown to perform much better than other non-linear classifiers such as artificial neural networks. However, their application to EEG signals classification problem has been very limited.

In this paper, we propose the use of RVM for EEG signals classification. Furthermore, a comparative study in terms of performance and complexity (number of relevance vectors versus number support vectors) is realized. As in traditional pattern-recognition systems, our approach consists of two main modules: a feature extractor based on DWT that generate a feature vector from the EEG signals and feature classifier that output the class based on the features vector.

In the next section we describe the data analyzed and the techniques used to preprocess it. In Section 3 we present the classifiers implement using the SVM and RVM; in Section 4 we show the results of the classifications and the accuracy rate; and then in Section 5 we conclude the study.

II. Data Analysis

A. *Data selection*

In this work, we have used the EEG data publicly available at (http://www.meb.uni-bonn.de/epileptologie/science/physik/eegdata.html [2]. The complete data set consists of five sets (denoted A-E) each containing 100 single-channel EEG segments. These segments were selected and cut out from continuous multi-channel EEG recordings after visual inspection for artifacts, e.g., due to muscle activity or eye movements. Sets A and B consisted of segments taken from surface EEG recordings that were carried out on five healthy volunteers using a standardized electrode placement scheme (Figure 1).

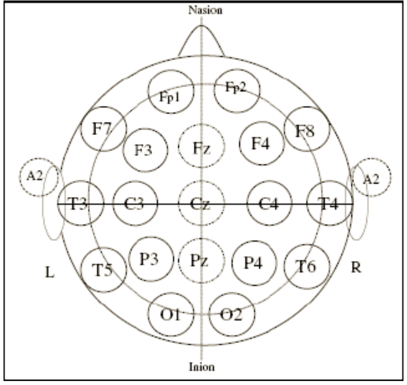

Figure 1: The 10-20 international system of electrode placement c images of normal and abnormal cases.

Volunteers were relaxed in an awake-state with eyes open (A) and eyes closed (B), respectively. Sets C, D, and E originated from EEG archive of pre-surgical diagnosis. EEGs from five patients were selected, all of whom had achieved complete seizure control after resection of one of the hippocampal formations, which was therefore correctly diagnosed to be the epileptogenic zone. Segments in set D were recorded from within the epileptogenic zone, and those in set C from the hippocampal formation of the opposite hemisphere of the brain. While sets C and D contained only activity measured during seizure free intervals, set E only contained seizure activity. Here segments were selected from all recording sites exhibiting ictal activity. All EEG signals were recorded with the same 128-channel amplifier system, using an average common reference. The data were digitized at 173.61 samples per second using 12 bit resolution. Bandpass filter settings were 0.53-40 Hz (12dB/oct). In this work, like in [14], we used two sets (A and E).

B. *Data preprocessing*

Wavelet transform is a spectral estimation technique in which any general function can be expressed as an infinite series of wavelets. The basic idea underlying wavelet analysis consists of expressing a signal as a linear combination of a particular set of functions (WT), obtained by shifting and dilating one single function called a mother wavelet. The decomposition of the signal leads to a set of coefficients called wavelet coefficients. Therefore the signal can be reconstructed as a linear combination of the wavelet functions weighted by the wavelet coefficients. In order to obtain an exact reconstruction of the signal, adequate number of coefficients must be computed. The key feature of wavelets is the time-frequency localization. It means that most of the energy of the wavelet is restricted to a finite time interval. Frequency localization means that the Fourier transform is band limited. When compared to FFT, the advantage of time-frequency localization is that wavelet analysis varies the time-frequency aspect ratio, producing good frequency localization at low frequencies (long time windows), and good time localization at high frequencies (short time windows). This produces a segmentation, or tiling of the time-frequency plane that is appropriate for most physical signals, especially those of a transient nature. The wavelet technique applied to the EEG signal will reveal features related to the transient nature of the signal which are not obvious by the Fourier transform [6].

Selection of suitable wavelet and the number of decomposition levels is very important in analysis of signals using the DWT. The number of decomposition levels is chosen based on the dominant frequency components of the signal. The levels are chosen such that those parts of the signal that correlate well with the frequencies necessary for classification of the signal are retained in the wavelet coefficients. In the present study, since the EEG signals do not have any useful frequency components above 30 Hz, the number of decomposition levels was chosen to be 5. Thus, the EEG signals were decomposed into details D1-D5 and one final approximation, A5 [14].

We have used a Daubechies order-4 wavelet (db4), its smoothing feature made it more appropriated to detect changes of EEG signal [14].

C. *Feature Extraction*

In order to reduce the dimensionality of the extracted features vector [4][14][17][12], statistics over the set of the wavelet coefficients were used to generate the input to the SVM:

Statistics over wavelet coefficients obtained:

- Average of wavelet coefficients in each sub-band (W_Avg).
- Standard deviation of wavelet coefficients in each sub-band (W_Std).
- Maximum of wavelet coefficients in each sub-band (W_Max).

III. Variants of Support Vector Machines

In this section we revise the use of SVM and RVM in classification problems.

A. *Standard SVM*

Let the training set be $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, with input $\mathbf{x}_i \in \Re^{m}$ and $y_i \in \{\pm 1\}$. The SVM first accomplishes a mapping ϕ: $\Re^{m} \rightarrow \Re^{n}$. Usually, n is much higher than m in such a way that the input vector is mapped into a high-dimensional space. When data are linearly separable, the SVM builds a hyperplane in $\Re^{n}$ $\mathbf{w}^T\phi(\mathbf{x})+b$ in which the boundary between positive and negative samples is maximized. It can be shown that $\mathbf{w}$, for this optimal hyperplane, may be defined as a linear combination of $\phi(\mathbf{x}_i)$, that is $\mathbf{w}=\sum_{i=1}^{N}\alpha_i y_i \phi(\mathbf{x}_i)$.

The generalized optimal separating hyperplane is determined by the vector **w**, that minimizes the functional:

$$\min_{w,b,e} J(\mathbf{w},b,e) = \tfrac{1}{2}(\mathbf{w}^T\mathbf{w}) + C\sum_{t=1}^{N}\xi_t \tag{1}$$

(where C is a given value) subject to the constraints

$$y_t[\mathbf{w}^T\varphi(\mathbf{x}_t)+b] \geq 1-\xi_t,\ t=1,\cdots,N \tag{2}$$

Then, the resulting quadratic programming (QP) problem may be written as:

$$\max_{\alpha} J(\boldsymbol{\alpha}) = \max_{\alpha}\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) \tag{3}$$

subject to $\sum_{i=1}^{N}\alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$, for $i=1,\cdots,N$, for $i=1,\cdots,N$. To obtain $\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ in QP problem we do not need to calculate $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$ explicitly. Instead, for some ϕ, we can design a kernel K(.,.) such that $K(\mathbf{x}_i,\mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$.

Then, the expression of QP problem becomes:

$$\max_{\alpha} J(\boldsymbol{\alpha}) = \max_{\alpha}\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) \tag{4}$$

For the training samples along the decision boundary, the corresponding $\alpha_i{}'s$ are greater than zero, as ascertained by the Kuhn-Tucker Theorem [13].These samples are known as support vectors. The number of support vectors is generally much smaller than *N*, being proportional to the generalization error of the classifier [18]. A test vector $\mathbf{x} \in \Re^{m}$ is then assigned to a given class with respect to the expression

$$f(\mathbf{x}) = sign(\mathbf{w}^T\phi(\mathbf{x})+b) = sign\left(\sum_{i=1}^{N}\alpha_i y_i K(\mathbf{x},\mathbf{x}_i)+b\right)$$

B. *Relevance Vector Machine*

The RVM introduces a priori over the model weights governed by a set of hyper-parameters, in a probabilistic framework. One hyperparameter is associated with each weight, and the most probable values are iteratively estimated from the training data. The most compelling feature of the RVM is that it typically utilizes significantly fewer kernel function compared to the SVM, while providing a similar performance.

For two-class classification, any target can be classified into two class such that $t_n\{0,1\}$. A Bernoulli distribution can be adopted for $p(t|\mathbf{x})$ in the probabilistic framework because only two classes (0 and 1) are possible. The logistic sigmoid link function $\sigma(y) = 1/(1+\exp(-y))$ is applied to $y(\mathbf{x})$ to link random and systematic components, and generalize the linear model. Following the definition of the Bernoulli distribution, the likelihood is written as

$$p(t|\mathbf{w}) = \prod_{n=1}^{N}\sigma\{y(\mathbf{x}_n;\mathbf{w})\}^{t_n}(1-\sigma\{y(\mathbf{x}_n;\mathbf{w})\})^{1-t_n} \tag{5}$$

for the targets $t_n \in \{0,1\}$.

The likelihood is complementary by a prior over the parameter (weights) in the form of

$$p(t|\mathbf{w}) = \prod_{n=1}^{N}\frac{\sqrt{\alpha_i}}{\sqrt{2\pi}}\exp\left(-\frac{\alpha_i w_i^2}{2}\right) \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_n)^T$ shows the hyperparameter introduced to control the strength of the priori over its associated weight. Hence, the prior is Gaussian, but conditioned on **α.**

For a certain **α** value, the posterior weight distribution conditioned on the data can be obtained using Bayes' rule, i. e,

$$p(\mathbf{w}|t,\boldsymbol{\alpha}) = \frac{p(t|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})}{p(t|\boldsymbol{\alpha})} \qquad (7)$$

where p(t|**w**) is likelihood of t, p(**w**| $\boldsymbol{\alpha}$) is the prior density of **w**, and p(t| $\boldsymbol{\alpha}$) is referred to as evidence.

The weight cannot be analytically obtained, and therefore, a Laplacian approximation procedure is used [8].

1) Since p(**w**|t,**α**) is linearly proportional to $p(t|\mathbf{w}) \times p(\mathbf{w},\boldsymbol{\alpha})$, it is possible to aim to find the maximum of

$$\log\{p(t|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})\} = \sum_{n=1}^{N}[t_n log\, y_n + (1-t_n)\log(1-y_n)]\frac{1}{2}\mathbf{w}^T A\mathbf{w}$$

for the most probable weight $\mathbf{w}_{MP}$, with $y_n = \sigma\{y(\mathbf{x}_n;\mathbf{w})\}$ and $\mathbf{A} = diag(\alpha_0,\alpha_1,\cdots,\alpha_N)$ being composed of the current values of **α**. This is a penalized logistic log-likelihood function and requires iterative maximization. The iteratively reweighed least-squares algorithm can be used to find $\mathbf{w}_{MP}$ [15][16][16].

2) The logistic log-likelihood function can be differentiated twice to obtain the Hessian in the form of

$$\nabla_w\nabla_w logp(\mathbf{w}|t,\boldsymbol{\alpha})|_{\mathbf{w}_{MP}} = -(\emptyset^T\mathbf{B}\emptyset + \mathbf{A}) \qquad (8)$$

where $\mathbf{B} = diag(\beta_1,\beta_2,\cdots,\beta_N)$ is a diagonal matrix with $\beta_n = \sigma\{y(\mathbf{x}_n;\mathbf{w}_{MP})\}[1-\sigma\{y(\mathbf{x}_n;\mathbf{w}_{MP})\}]$ and $\emptyset$ is the design matrix with $\emptyset_{nm} = K(\mathbf{x}_n,\mathbf{x}_{m-1})$ and $\emptyset_{n1} = 1$. This result is then negated and inverted to give the covariance Σ, as shown as follows, for a Gaussian approximation to the posterior over weights centered at $\mathbf{w}_{MP}$:

$$\sum = (\emptyset^T\ \mathbf{B}\emptyset + \mathbf{A})^{-1} \qquad (9)$$

In this way, the classification problem is locally linearized around $\mathbf{w}_{MP}$ in a effective way with

$$\mathbf{w}_{MP} = \boldsymbol{\Sigma}\emptyset^T\mathbf{B}\hat{t}$$
$$\hat{t} = \emptyset\mathbf{w}_{MP} + \mathbf{B}^{-1}(t-y)$$

These equations are basically equivalent to the solution of a generalized least-square problem. After obtaining $\mathbf{w}_{MP}$ the hyper-parameters α_i are updated using $\alpha_i^{new} = \lambda_i/w_i^2$, where w_i^2 is the ith posterior mean weight and λ_i is defined as $\lambda_i = 1 - \Sigma_{ii}$ where Σ_{ii} is the ith diagonal element of the covariance, and can be regarded as a measure of how well determined each parameter w_i is by the data. During the optimization process, many α_i will have large values, and thus, the corresponding model weights are pruned out, realizing sparsity. The optimization process typically continues until the maximum change in α_i values is below a certain threshold the maximum of iteration number of iterations is reached.

IV. Experimental Results

In what follows, we provide details on how the experiments have been conducted, and present the comparative analysis of the two types of SVMs with respect to the kernel parameter and number of support vector variation.

The most popular kernels used in SVM and RVM are the linear, polynomial, radial basis function (RBF) and exponential radial basis function (ERBF) kernels. The linear kernel typically shows a lower performance and is therefore not employed in the provided results. Note that σ^2 determines the variance in the case of the RBF and ERBF kernel.

- Linear Kernel

$$K(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i.\mathbf{x}_j$$

- Polynomial Kernel

$$K(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i.\mathbf{x}_j)^d$$

- RBF Kernel

$$K(\mathbf{x}_i,\mathbf{x}_j) = exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}\right)$$

- ERBF Kernel

$$K(\mathbf{x}_i,\mathbf{x}_j) = exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|}{2\sigma^2}\right)$$

A. *Configuration of the experiments*

In the experiments accomplished, as mentioned before, we have assessed the performance of the RVM and SVM models with regard to the variation of kernel parameter, keeping the value of the regularization parameter C constant in 100. This value was achieved after some preliminary experiments and agrees with the fact that SVM models with low values of C tend in general to achieve better performance than those with high values of this parameter. Although we know that there are several rules-of-thumb to select the values kernel parameter [3], for RBF and ERBF kernel we have opted to set the values of σ as $0.5, 1, 2, 4$ and for polynomial kernel we have opted to set the values of d as 1, 2, 3, 4. For each of the four values in this range, a 10-fold cross-validation was performed to better gauge the average performance of the models. To define whether a sample will be a support vector or relevance vector was used a threshold equal to 10^{-6}, i. e, $\|\boldsymbol{\alpha}\| > 10^{-6}$.

B. *Results*

In Table 1, we provide the value(s) of the kernel parameter and the correspondent number of support vector (SV) for the SVM and number of relevance vector (RV) for the RVM, in terms of cross-validation, for each quadruple <features vector, model type, kernel type, kernel parameter>. Besides the features vector considered, was conducted simulation with the EEG series without pre-processing.

C. *Discussion*

Considering the results presented in Table 1, one can observe that, in most of the cases, the performance indices (i.e. misclassification rate) showed by the two types of vector machines were quite similar to each other, with a slight prevalence of the standard SVM models. However, RVM requires a significantly less number of relevance vectors (RVs) as compared with the number of support vector (SVs) used in SVM; hence the classification time is considerably reduced.

Except for the feature vector extracted through standard deviation and maximum of the coefficients of wavelet, the SVM, with RBF and ERBF kernel and parameter σ = 3, 4; produced a smaller number of SVs when compared to number of RVs used in RVM. This can be justified due to the very low value of the threshold used to define SVs and RVs. For other kernel parameters value and features vector the number of the SVs was much bigger than the number of RVs.

From these results, it is possible to conclude that the choice of the kernel parameter value and kernel function was not so much an important factor to be considered to distinguish between the overall best error rates exhibited by the machines.

When comparing the results obtained using features vector and EEG data without pre-processing, one observed that the performance is quite similar. This indicates that the selection of parameters to the kernel is more important than the technique used for the extraction of features.

V. CONCLUDING REMARKS

In this paper, we have presented a preliminary analysis study contrasting as the performance exhibited by SVM and RVM classifiers as number of the SVs and RVs with respect to the calibration of the kernel parameter value and vector feature applied to EEG signals classification problem. Such study is interesting as it can provide hints on how these machines are affected by the hyper-parameter tuning process and feature vector extracted in the EEG signal classification problem.

Experimental results show that RVM is superior to SVM in terms of the number of kernel functions that needs to be used in the classification phase. Therefore, RVM is preferable SVM in applications that require low complexity and, possibility, real-time classification with a priori training.

As ongoing work, we are currently extending the scope of investigation by considering multiclass problems, other kernel functions (and parameters), other types of vector machines — such as the Proximal SVMs, the Lagrangian SVMs [7], as well as (and most importantly) the conjoint influence of the hyper-parameters. In the future, we plan to investigate how the combination of models coming from different types of vector machines, each configured with the same values of the control parameters, can improve the levels of performance, in terms of accuracy and generalization, from that achieved by each vector machine type alone.

VI. REFERENCES

[1] Adeli H, Zhou Z, Dadmehr N. "Analysis of EEG records in an epileptic patient using wavelet transform." Journal Neuroscience Methods, vol. 123, no. 1, pp. 69–87, 2003.

[2] Andrzejak, R. G.; Lehnertz, K.; Mormann, F.; Rieke, C.; David, P. Elger, C. E. "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state." Phys. Rev. E, vol. 64, 061907, 2001.

[3] Cherkassky, V ; Ma, Y., "Practical selection of SVM parameters and noise estimation for SVM regression." Neural Networks, vol. 17, pp. 113-126, 2004.

[4] Güler; I. ; Übeyli, E. D. "A modified mixture of experts network structure for ECG beats classification with diverse features." Engineering Applications of Artificial Intelligence, vol. 18, pp. 845-856, 2005.

[5] Guler, I., Kiymik, M. K., Akin, M., Alkan, A. "AR spectral analysis of EEG signals by using maximum likelihood estimation." Computers in Biology and Medicine, vol. 31, pp. 441–450, 2001.

[6] Kara; S.; Okandan, M. "Atrial fibrillation classification with artificial neural networks." Pattern Recognition, vol. 40, pp. 2967 – 2973, 2007.

[7] Lima, C. A. M., Villanueva, W. J. P., dos Santos, E. P. and Von Zuben, F. J. "A multistage ensemble of support vector machine variants." In: Procs. of the 5th International Conference on Recent Advances in Soft Computing, Nottingham, pp. 670-675, 2004.

[8] Mackay, D. J. C. "The evidence framework applied to classification networks." Neural Computation, vol. 4, no. 5, pp. 720-736, 1992

[9] Nabney, I. T. "Efficient training of RBF networks for classification." In Proc. 9th ICANN, 1999, vol. 1, pp. 210-215.

[10] Petrosian, A., Prokhorov, D., Homan, R., Dashei, R., Wunsch, D. "Recurrent neural network based prediction of epileptic seizures in intraand extracranial EEG." Neurocomputing, vol. 30, pp. 201–218, 2000.

[11] Pradhan, N., Sadasivan, P. K., Arunodaya, G. R. "Detection of seizure activity in EEG by an artificial neural network: A preliminary study." Computers and Biomedical Research, vol. 29, pp. 303–313, 1996.

[12] Revett, K. Jahankhani, P.; Kodogiannis, V. "EEG Signal Classification Using Wavelet Feature Extraction and Neural Networks." IEEE John Vincent Atanasoff 2006 International Symposium on Volume, pp. 120 – 124, 2006.

[13] Shawe-Taylor, J, Cristianini, N. "An Introduction to Support Vector Machines", Cambridge. Press, 2000.

[14] Subasi, "A. EEG signal classification using wavelet feature extraction and a mixture of expert model." Expert Systems with Applications, vol. 32, pp.1084-1093, 2007.

[15] Tipping, M. "Sparse bayesian learning and the relevance vector machine." Journal of Machine Learning Research, pp. 211-244, 2001.

[16] Tipping, M. "The Relevance Vector Machine." Advances in Neural Information Proceeding systems 12. Cambreidge, Mass: MIT Press, 2000.

[17] Übeyli, E. D. "Wavelet/mixture of experts network structure for EEG signals classification." Expert Systems with Applications, vol. 34, pp. 1954-1962, 2008.

[18] Vapnik, V. N. "The Nature of Statistical Learning Theory", Springer Verlag, 1995.

[19] Weng, W., Khorasani, K. "An adaptive structure neural network with application to EEG automatic sei-

zure detection." Neural Networks, vol. 9, pp. 1223–1240, 1996.

[20] Zoubir, M., Boashash, B. 'Seizure detection of newborn EEG using a model approach". IEEE Transactions on Biomedical Engineering, vol. 45, pp. 673–685, 1998.

Table 1 - Comparative analysis: For each quadruple <features vector, model type, kernel type, kernel parameter>

Feature vector	Kernel Type	Kernel Parameter		SVM		RVM	
		σ	d	Error	SV	Error	RV
Raw EEG Data	ERBF	0.5	-	0.575 ± 0.0226	180 ± 0.0000	0.500 ± 0.0333	103.2 ± 0.4422
		1	-	0.385 ±0.0699	180 ± 0.0000	0.485 ± 0.0325	96.3 ± 0.6675
		2	-	0.005 ± 0.0050	151.2 ± 0.5333	0.480 ± 0.0326	88.3 ± 0.6675
		4	-	0.000 ± 0.0000	66.6 ± 0.6863	0.015 ± 0.0106	57.5 ± 5.6711
	Poly	-	1	0.290 ± 0.0296	127.8 ± 1.3317	0.250 ±0.0307	5.4 ± 0.2666
		-	2	0.230 ± 0.0416	130.3 ± 2.0279	0.460 ± 0.0305	8.2 ± 0.1333
		-	3	0.395 ± 0.0404	105.0 ± 2.0817	0.365 ± 0.0799	4.1 ± 0.1000
		-	4	0.340 ± 0.0400	19.5 ± 0.9098	0.500 ± 0.0333	8.0 ± 0.2582
	RBF	0.5	-	0.580 ± 0.0200	180 ± 0.0000	0.500 ± 0.0333	180 ± 0.0000
		1	-	0.575 ± 0.0226	180 ± 0.0000	0.500 ± 0.0333	180 ± 0.0000
		2	-	0.575 ± 0.0226	180 ± 0.0000	0.500 ± 0.0333	23.7 ± 0.55877
		4	-	0.295 ± 0.0479	180 ± 0.0000	0.440 ± 0.0286	100.5 ± 10.896
W_Avg	ERBF	0.5	-	0.105 ± 0.0273	154.8 ± 0.87939	0.125 ± 0.0300	35.2 ± 4.7394
		1	-	0.115 ± 0.0307	101.5 ± 0.95743	0.115 ± 0.0247	12.1 ± 0.27689
		2	-	0.115 ± 0.0307	98.7 ± 0.9434	0.125 ± 0.0271	18.6 ± 4.1317
		4	-	0.115 ± 0.0307	97.1 ± 0.72188	0.130 ± 0.0249	79 ± 9.3178
	Poly	-	1	0.455 ± 0.0283	168.9 ± 1.8586	0.560 ± 0.0266	1.9 ± 0.27689
		-	2	0.120 ± 0.0226	30.3 ± 0.68394	0.110 ± 0.0221	10.5 ± 0.37268
		-	3	0.235 ± 0.0258	51.2 ± 1.0414	0.270 ± 0.0606	11.3 ± 0.81718
		-	4	0.220 ± 0.0226	47.1 ± 1.2949	0.365 ± 0.0703	9.9 ± 0.9481
	RBF	0.5	-	0.100 ± 0.0197	120.1 ± 1.1874	0.250 ± 0.0428	52.6 ± 1.3515
		1	-	0.130 ± 0.0249	44.8 ± 0.8	0.150 ± 0.0341	28.1 ± 2.6476
		2	-	0.120 ± 0.0290	45.5 ± 1.0138	0.125 ± 0.0250	13.3 ± 1.3337
		4	-	0.130 ± 0.0260	65.1 ± 1.0588	0.100 ±0.0210	94.8 ± 8.4074
W_Std	ERBF	0.5	-	0.005 ± 0.0050	87.6 ± 0.7774	0.005 ± 0.0050	96.7 ± 6.2004
		1	-	0.000 ± 0.0000	33 ± 0.42164	0.005 ± 0.0050	17.9 ± 6.5599
		2	-	0.000 ± 0.0000	18.4 ± 0.49889	0.005 ± 0.0050	71.1 ± 12.469
		4	-	0.000 ± 0.0000	16.5 ± 0.45338	0.005 ± 0.0050	133.8 ± 9.4349
	Poly	-	1	0.005 ± 0.0050	4.2 ± 0.13333	0.010 ± 0.0066	1.9 ± 0.1000
		-	2	0.005 ± 0.0050	6.7 ± 0.335	0.005 ± 0.0050	2.2 ± 0.1333
		-	3	0.000 ± 0.0000	3.6 ± 0.33993	0.000 ± 0.0000	2 ± 0.0000
		-	4	0.010 ± 0.0066	6.5 ± 0.26874	0.005 ± 0.0050	3.6 ± 0.2666
	RBF	0.5	-	0.005 ± 0.0050	63.5 ± 0.79232	0.005 ± 0.0050	67.1 ± 6.6507
		1	-	0.000 ± 0.0000	33.6 ± 0.68638	0.010 ± 0.0066	32.5 ± 8.9446
		2	-	0.000 ± 0.0000	5.5 ± 0.70317	0.005 ± 0.0050	65.2 ± 14.3045
		4	-	0.000 ± 0.0000	5.4+-0.22111	0.005 ± 0.0050	96 ± 11.1744
W_Max	ERBF	0.5	-	0.005 ± 0.0050	90.9 ± 0.58595	0.005 ± 0.0050	106.3 ± 4.96
		1	-	0.000 ± 0.0000	28.6 ± 0.6532	0.000 ± 0.0000	26.1 ± 5.5165
		2	-	0.000 ± 0.0000	13.4 ± 0.4761	0.005 ± 0.0050	97.9 ± 6.6857
		4	-	0.000 ± 0.0000	12.3 ± 0.36667	0.000 ± 0.0000	131.2 ± 4.2395
	Poly	-	1	0.010 ± 0.0066	5.4 ± 0.2211	0.010 ± 0.0066	1.9 ± 0.1000
		-	2	0.015 ± 0.0106	8.4 ± 0.4000	0.040 ± 0.0348	2.8 ± 0.2000
		-	3	0.005 ± 0.0050	6.3 ± 0.2134	0.005 ± 0.0050	2 ± 0.0000
		-	4	0.020 ± 0.0081	10.8 ± 0.3887	0.135 ± 0.0781	21.6 ± 17.605
	RBF	0.5	-	0.000 ± 0.0000	70.5 ± 0.5426	0.005 ± 0.0050	61.3 ± 10.127
		1	-	0.005 ± 0.0050	9.3 ± 2.2214	0.015 ± 0.0076	62.2 ± 8.7519
		2	-	0.005 ± 0.0050	5.4 ± 0.5206	0.010 ± 0.0066	40.6 ± 11.6411
		4	-	0.010 ± 0.0066	6.3 ± 0.1527	0.005 ± 0.0050	118.2 ± 2.8316