Randomized Pattern Classifiers for Epileptic Seizure Detection: A Critical Assessment

Natanael R. Silva†, Júlio P. Silva Júnior†
Guilherme A. Barreto†, and Rodrigo B. Souza‡

†Federal University of Ceará (UFC), Department of Teleinformatics Engineering
Campus of Pici, Center of Technology, Fortaleza, Ceará, Brazil
Emails: natanrs21@gmail.com, juliojrps@gmail.com, gbarreto@ufc.br

‡Federal University of Ceará (UFC), Hospital Walter Cantídio
Department of Clinical Medicine, Fortaleza, Ceará, Brazil
Email: rodrigo.becco@ebserh.gov.br

Resumo. In this paper we evaluate the performances of randomized pattern classifiers in the task of EEG-based epileptic seizures detection. Our goal is to investigate if these new class of machine learning methods actually outperform powerful nonlinear classifiers, such as the MLP and SVM, in complex pattern recognition tasks. The rationale for the current work comes from the observation that the recent wave of applications involving randomized classifiers tend to report only positive reports, in which these networks always achieve equivalent or better performances than non-randomized nonlinear classifiers. A comprehensive performance evaluation is carried out, with the results strongly corroborate our hypothesis that randomized classifiers usually do not perform better than well trained standard nonlinear classifiers. Additionally, the performances of randomized classifiers are more dependent on the feature extraction method than non-randomized ones.

Keywords: epileptic seizures, randomized classifiers, Welch's periodogram, LPC coefficients, ROC curves.

1 Introduction

There has been an ever growing interest on randomized machine learning algorithms for complex pattern recognition tasks. A few examples of such algorithms are the random vector functional link (RVFL) [1,2], the extreme learning machine (ELM) [3], the no-prop network [4], random forests [5], and random kitchen sinks [6]. All this interest seems to be primarily motivated by the very fast way they are designed, without resorting to a long learning process across several training epochs, as required by standard learning algorithms such the backpropagation algorithm [7].

Fast classifier design is achieved (e.g. for the neural network based classifiers) simply by randomizing the input-to-hidden layer weights/biases. Only the hidden-to-output layer weights (output weights, for short) are computed, which

can be carried out by means of any standard parameter estimation technique for linear systems (e.g. ordinary least squares).

A direct consequence of all the hype around randomized classifiers is that the papers on this issue almost always report positive results, in which the performances of the randomized classifiers are equivalent or better than those achieved by powerful nonlinear classifiers, such as the MLP and SVM. Critical evaluations are very rare or inexistent. Thus, a more experienced reader can clearly identify confirmation bias in the reported results, where the randomized networks are built to their best performances, while more traditional classifiers are not designed with the same enthusiasm.

From the exposed, in this paper we aim at filling a gap in the literature of randomized pattern classifiers. For this purpose, we selected a very challenging task of detecting epileptic seizures from EEG signals [8]. The epileptic seizure is a state in which there is an abnormal, excessive, synchronous discharge of neurons located basically in the cerebral cortex. This abnormal activity is intermittent and usually self-limiting, lasting from a few seconds to a few minutes and affects millions of people worldwide (around 1% of the world population) [9].

The EEG signal is highly noisy, nonlinear and nonstationary [10]. Such features characterize EEG processing (e.g. for seizure detection) as a very challenging task, despite all the developments in nonlinear feature extraction and machine learning methods. The literature on the application of standard nonlinear classifiers in seizure detection/classification, such as the MLP and kernel machines (SVM and LSSVM), is extensive [11]. As expected, there is an increasing interest in applying randomized classifiers for the same tasks [12,13,14,15].

The contribution of this paper is twofold. First, we provide a comprehensive investigation of the performances of randomized classifiers in EEG-based seizure detection for different feature selection techniques. Then, we compare the performances of these randomized classifiers to those of standard linear/nonlinear classifiers. The obtained results seem to corroborate our hypothesis that randomized classifiers often do not perform better than well trained standard nonlinear classifiers. Furthermore, for the task we are interested in, the performances of randomized classifiers are highly dependent on the feature extraction method.

The remainder of the paper is organized as follows. In Section 2, we describe the randomized classifiers evaluated in this paper. in Section 3 we briefly present the two feature extraction methods used in the simulations and describe how to build the feature vectors for classification purposes. The results are presented and discussed in Section 4. The paper is concluded in Section 5.

2 Evaluated Classifiers

Let us assume that we have already collected N data pairs $\{(\mathbf{x}_n, d_n)\}_{n=1}^{N}$ for building and evaluating the model, where $\mathbf{x}_n \in \mathbb{R}^p$ is the n-th p-dimensional input pattern and $d_n \in \{-1, +1\}$ is the corresponding target class label. Then, let us randomly select N_1 ($N_1 < N$) training input-output pairs from the available data pool and arrange the input vectors along the columns of the matrix $\mathbf{X}$

$(p \times N_1)$, while the target values are stacked into the column-vector $\mathbf{d}$ $(N_1 \times 1)$:

$$\mathbf{X} = [\mathbf{x}_1 \mid \mathbf{x}_2 \mid \cdots \mid \mathbf{x}_{N_1}] \quad \text{and} \quad \mathbf{d} = [d_1 \; d_2 \; \cdots \; d_{N_1}]^T, \tag{1}$$

where the superscript T denotes the transpose of a vector/matrix.

2.1 The Random Vector Functional Link Network (RVFL)

The RVFL [1,2] is a randomized SLFN with two pathways for processing information from input units to output neurons. These pathways are then added to form the network's output. The first pathway is a linear one, which directly connects the input units to the output neuron. Mathematically, we get

$$y_n^{(1)} = \mathbf{w}_1^T \mathbf{x}_n, \tag{2}$$

where $\mathbf{w}_1 \in \mathbb{R}^p$ is the corresponding weight vector[1]. The second pathway processes the input vectors through a hidden layer of q $(q \geq 1)$ nonlinear neurons; that is,

$$y_n^{(2)} = \mathbf{w}_2^T \mathbf{h}_n, \tag{3}$$

where $\mathbf{w}_2 \in \mathbb{R}^q$ is the corresponding weight vector and $\mathbf{h}_n \in \mathbb{R}^q$ is the hidden activation vector, i.e. the vector containing the outputs of the hidden neurons in response to the current input vector $\mathbf{x}_n$. The vector $\mathbf{h}_n$ is computed as

$$\mathbf{h}_n = \phi(\mathbf{M}\mathbf{x}_n) = [\phi(\mathbf{m}_1^T \mathbf{x}_n + b_1), \ldots, \phi(\mathbf{m}_q^T \mathbf{x}_n + b_q)]^T, \tag{4}$$

where $\phi(\cdot)$ is a nonlinear (e.g. sigmoidal) activation function operating at each component of its argument vector, $\mathbf{M}$ is a $q \times p$ weight matrix, and b_j, $j = 1, .., q$, denotes the bias of the j-th hidden neuron. It should be noted that the weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ are estimated from data, while the entries of the matrix $\mathbf{M}$ and the biases b_j are randomly sampled either from a uniform or a normal distribution.

If we add the outputs of both pathways, we get

$$y_n = y_n^{(1)} + y_n^{(2)} = \mathbf{w}_1^T \mathbf{x}_n + \mathbf{w}_2^T \mathbf{h}_n = [\mathbf{w}_1^T \mid \mathbf{w}_2^T] \begin{bmatrix} \mathbf{x}_n \\ - \\ \mathbf{h}_n \end{bmatrix} = \mathbf{w}^T \mathbf{z}_n, \tag{5}$$

where $\mathbf{w} = [\mathbf{w}_1^T \mid \mathbf{w}_2^T]^T$ is the $(p+q) \times 1$ vector obtained from the concatenation of the weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$. By the same token, $\mathbf{z}_n$ is the $(p+q) \times 1$ vector formed from the concatenation of the current input vector $\mathbf{x}_n$ and the current hidden activation vector $\mathbf{h}_n$.

The weight vector $\mathbf{w}$ can be readily estimated via the ordinary least squares (OLS) method by means of the following expression:

$$\mathbf{w} = (\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{d}, \tag{6}$$

[1] We assume that all vectors are column-vectors, unless stated otherwise.

where $\mathbf{Z} = [\mathbf{z}_1 \,|\, \mathbf{z}_2 \,|\, \cdots \,|\, \mathbf{z}_{N_1}]$ is a $(p+q) \times N_1$ matrix whose N_1 columns are the augmented vectors $\mathbf{z}_n = [\mathbf{x}_n^T \mid \mathbf{h}_n^T]^T \in \mathbb{R}^{p+q}$, $n = 1, \ldots, N_1$, where N_1 is the number of available training input patterns. The vector $\mathbf{d}$ is defined in Eq. (1). To avoid numerical problems, a regularized version of Eq. (6) is commonly used, which is given by

$$\mathbf{w} = (\mathbf{Z}\mathbf{Z}^T + \lambda \mathbf{I})^{-1}\mathbf{Z}\mathbf{d}, \tag{7}$$

where the constant $\lambda > 0$ is the regularization parameter.

2.2 The Extreme Learning Machine (ELM)

The ELM network is a recent randomized SLFN introduced by Huang *et al.* [3], whose weights from the inputs to the hidden neurons are randomly chosen, while only the weights from the hidden neurons to the output are analytically determined. Consequently, ELM offers significant advantages such as fast learning speed, ease of implementation, and less human intervention when compared to more traditional SLFNs, such as the Multilayer Perceptron (MLP) and RBF networks.

From an architectural point of view, the ELM network can be understood as a simplified version of the RVFL in which the direct linear path is removed. Thus, the equations of the ELM are easily obtained as follows:

Output computation: From Eq. (5), once we remove the direct linear pathway, we get

$$y_n = y_n^{(2)} = \mathbf{w}_2^T \mathbf{h}_n, \tag{8}$$

where $\mathbf{h}_n$ is defined as in Eq. (4).

Estimation of $\mathbf{w}_2$: In this case, the expression of the OLS estimate in Eq. (6) reduces to

$$\mathbf{w}_2 = (\mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{d}, \tag{9}$$

where $\mathbf{H} = [\mathbf{h}_1 \,|\, \mathbf{h}_2 \,|\, \cdots \,|\, \mathbf{h}_{N_1}]$ be a $q \times N_1$ matrix whose N_1 columns are the hidden activation vectors $\mathbf{h}_n \in \mathbb{R}^q$, $n = 1, \ldots, N_1$, where N_1 is the number of available training input patterns.

The former two randomized classifiers (i.e. RVFL and ELM) originated in the neural network field. The RKS classifier [16], in its turn, has origin within the field of kernel machines.

2.3 Random Kitchen Sinks (RKS)

Kernel methods provide an elegant, theoretically well-founded, and powerful approach to solving many learning problems. Since traditional algorithms require the computation of a full $N \times N$ pairwise kernel matrix to solve learning problems on N input vectors, however, scaling these methods to large-scale datasets containing more than thousands of data points has proved challenging.

Rahimi and Recht[17,16] triggered interest in one very attractive approach: approximating a continuous shift-invariant kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ by

$$k(\mathbf{x}, \mathbf{y}) \approx z(\mathbf{x})^T z(\mathbf{y}), \tag{10}$$

where $z : \mathcal{X} \rightarrow \mathbb{R}^D$. Then primal methods in $\mathbb{R}^D$ can be used, allowing most learning problems to be solved in $O(N)$ time [18].

To approximate the RKHS induced by the Gaussian kernel

$$k(x, y) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2), \tag{11}$$

it suffices to sample the weights $\mathbf{m}$ from the inverse Fourier transform of $k(\cdot, \cdot)$, which is just a multivariate Gaussian density with mean vector $\mathbf{0}$ and covariance matrix $2\gamma\mathbf{I}$. A direct consequence of this approach is that the RKS classifier becomes very similar in structure to the ELM network.

Thus, for each input vector $\mathbf{x}_n$, we compute the approximated feature mapping as

$$\mathbf{h}_n = \phi(\mathbf{M}\mathbf{x}_n) = [\phi(\mathbf{m}_1^T \mathbf{x}_n + b_1), \ldots, \phi(\mathbf{m}_q^T \mathbf{x}_n + b_q)]^T, \tag{12}$$

with $\mathbf{m}_i \sim \text{Normal}(\mathbf{0}, 2\gamma\mathbf{I})$ and $b_i \sim \text{Uniform}(-\pi, +\pi)$. The output vectors are then computed exactly as shown in Eq. (9).

For the sake of completeness, we included in the comparative analysis the standard least squares (LS) linear classifier, and also two powerful supervised nonlinear classifiers, namely the multilayer perceptron (MLP) network and the Vapnik's support vector Machine (SVM).

3 Feature Extraction Methods

The problem of interest - epileptic seizure detection - is usually treated a binary classification problem. In this regard, the classifier has to differentiate between periods of either normal or convulsive epileptic activity (henceforth called convulsive seizure or simply seizure). The data used in this study were collected at the Children's Hospital of Boston, Massachusetts (MIT). The database *CHB-MIT Scalp EEG Database* is available for public download[2]. It is comprised of 24 patients already diagnosed with epilepsy, where video-EEG monitoring was performed after the withdrawal of the anti-convulsive medication [19,20]. The files have a sample rate of 256 Hz, the electrode assembly follows the standard 10-20 and the number of channels vary from 23 to 32. In total, 197 seizures were cataloged in 141 files, totaling 196 minutes of intervals containing epileptic seizures for all patients.

Since EEG is a complex time series, with aperiodic behavior and highly nonstationary, it is necessary to reparametrize the available time series in order to extracting attributes that preserve relevant information contained in the original signals. Two of the most commonly used methods for this purpose are the Welch's periodogram [21] or linear predictive coding (LPC) coefficients [22]. It should be

[2] CHB-MIT Scalp EEG Database, `http://www.physionet.org/pn6/chbmit/`

noted that, however, both methods assume stationarity of the time series. To deal with the high non-stationarity of an EEG signal, the original sequence is segmented in smaller subsequences, which are then assumed stationary.

Welch's Periodogram - This method for estimating the PSD of a stochastic signal combines windowing and averaging in order to obtain a smooth spectrum estimation without random fluctuations resulting from the estimation process itself. The original data sequence of each channel is divided into a number K of possible overlapping segments. A window v_n is defined over each of these segments and the corresponding periodograms are computed and then averaged. If $x_n^{(k)}$ represents the sample x_n of the k^{th} data subsequence (of length L), then the modified periodogram for that subsequence is computed as

$$\hat{P}_x^{(k)}(\omega) = \frac{1}{L}\left|\sum_{n=0}^{L-1} v_n x_n^{(k)} e^{-j\omega n}\right|^2, \qquad k = 1, \ldots, K \tag{13}$$

where $\omega = 2\pi f$ (in rad/s) is the angular frequency, and the window v should obey the following (normalization) property: $(1/L)\sum_{n=0}^{L-1} v_n^2 = 1$. Then the estimate of the PSD of the signal, for each frequency ω, is taken as

$$\hat{S}_x(\omega) = \frac{1}{K}\sum_{k=1}^{K} \hat{P}_x^{(k)}(\omega) \tag{14}$$

LPC coefficients - This method assumes that the dynamics of the *k-th* data subsequence can be described by an autoregressive model of order p, $AR(p)$:

$$x_n^{(k)} = a_{k,1}x_{n-1}^{(k)} + a_{k,2}x_{n-2}^{(k)} + ... + a_{k,p}x_{n-p}^{(k)} + \epsilon_n, \tag{15}$$

where $a_{k,1}$, ..., $a_{k,p}$ are the coefficients of the AR(p) model of the *k-th* subsequence, and ϵ_n is an additive white Gaussian noise (AWGN) process i.e. $\epsilon_n \sim N(0, \sigma_\epsilon^2)$. The coefficients $\{a_{k,l}\}_{l=1}^{p}$ associated to each subsequence $k = 1, ..., K$, can be estimated by several methods, such as well-known Yule-Walker (YW) equations or the ordinary least squares (OLS) [22].

3.1 Building the Feature Vectors

Let us consider a set of $N = 23$ channels from which we collect 23 EEG signals of a certain duration for each patient. For the dataset we are working with, each EEG signal lasts 1 hour, sampled at a rate of 256Hz. For a given EEG channel, feature vectors are built every two seconds. Successive two seconds long segments (called EEG epochs) are processed by a time window of duration $L = 2$ seconds. For a sampling rate of 256Hz, each segment contains 512 samples.

▷ **Welch's Periodogram** - Following the methodology proposed in [20], feature vectors extracted using the Welch's method for one subject are built according to the following steps (see Fig. 1):

- **Step 1** - For the current two-second long EEG epoch, apply the Welch's periodogram method. Repeat it for all the N EEG channels.
- **Step 2** - Apply a logarithmic scale to the resulting PSD values to convert them to decibels (dB).
- **Step 3** - Segment the resulting PSD (in dB) in $M = 8$ frequency bands covering the range from 0.5 to 25Hz and then compute the spectral energy falling within each band. For channel k, this procedure leads to the computation $M = 8$ feature values $x_{1,k}$, $x_{2,k}$, ..., $x_{M,k}$.
- **Step 4** - Within each two second EEG epoch at time $t = T$, concatenate the $M = 8$ spectral energies extracted from each of $N = 23$ EEG channels. This process forms a feature vector $\mathbf{X}_T$ of dimension $M \times N = 184$, defined as

$$\mathbf{X}_T = [x_{1,1} \;\; x_{2,1} \;\; \cdots \;\; x_{M,1} \mid \cdots \mid x_{1,N} \;\; x_{2,N} \;\; \cdots \;\; x_{M,N}]^\top \tag{16}$$

- **Step 5** - Build a stacked feature vector that is the result of concatenating the feature vectors from $W = 3$ contiguous, but non-overlapping two second epochs. The augmented feature vector $\mathcal{X}_T$ is defined as

$$\mathcal{X}_T = [\mathbf{X}_T \;\; \mathbf{X}_{T-L} \;\; \cdots \;\; \mathbf{X}_{T-(W-1)L}]^\top, \tag{17}$$

  and has dimension $W \times M \times N = 552$.

A word on the Step 5 is necessary. Electroencephalographers[3] require an EEG reading that looks abnormal to persist for a minimum of 6 to 10 seconds before considering it an actual seizure [23]. To incorporate this domain knowledge, we set $W = 3$ so that the evaluated classifiers take into consideration the evolution of feature vectors over at least a period of 6 seconds.

▷ **LPC coefficients** - The feature vector design using LPC coefficients also evolves in segments of two-second long EEG epochs. However, instead of specifying the number M of frequency bands over which we compute the energy per EEG epoch, we need to specify beforehand the order p of the AR model (p). After some experimentation, we set $p = 4$. Larger values did not improve considerably the classification accuracies, while lower values led to a degradation in performance.

Thus, feature vectors built by means of the LPC method for one subject are obtained as follows:

- **Step 1** - For the current two-second long EEG epoch, apply the Yule-Walker equation in order to estimate the corresponding p coefficients of the AR model. Repeat it for all the N EEG channels.
- **Step 2** - Within each two second EEG epoch at time $t = T$, concatenate the $p = 4$ coefficients estimated for each of the $N = 23$ EEG channels. This process forms a feature vector $\mathbf{X}_T$ of dimension $p \times N = 92$, defined as

$$\mathbf{X}_T = [a_{1,1} \;\; a_{1,2} \;\; \cdots \;\; a_{1,p} \mid \cdots \mid a_{N,1} \;\; a_{N,2} \;\; \cdots \;\; a_{N,p}]^\top \tag{18}$$

[3] A person who specializes in reading and interpreting electroencephalography.

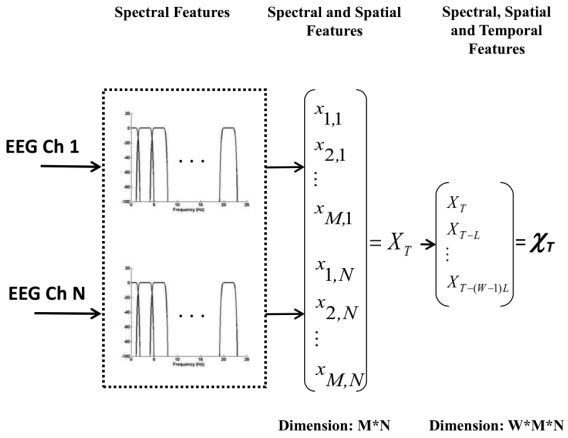


Fig. 1. Building the feature vectors for epileptic seizure detection from the N EEG channels for one patient [20].

- **Step 3** - Build a stacked feature vector that is the result of concatenating the feature vectors from $W = 3$ contiguous, but non-overlapping two second epochs. The augmented feature vector $\mathcal{X}_T$ has dimension $W \times p \times N = 276$:

$$\mathcal{X}_T = [\mathbf{X}_T \;\; \mathbf{X}_{T-L} \;\; \cdots \;\; \mathbf{X}_{T-(W-1)L}]^\top \tag{19}$$

It should be noted that due to the very nature of the task, there are many more feature vectors labeled as *normal* (-1, negative class) than as *seizure* (+1, positive class). In fact, only 2% of the analyzed EEG signals correspond to intervals containing seizures. This implies that the seizure detection tasks is highly unbalanced.

4 Results and Discussion

In this section we report the results of the performance analysis we carried out. To handle the unbalanced categories, we purposely equalized the ratio of positive to negative instances per patient. Then, we randomly divided the available instances per patient in 3 subgroups: training (70%), validation (20%) and testing (10%). The following classifiers were evaluated in this paper, namely: LS, RVFL, RKS, ELM, MLP and SVM classifiers. The two feature extraction techniques and all the classifiers were implemented in Matlab.

The classifiers' hyperparameters are listed in Table 1. These values were chosen after 100 independent runs of 5-fold cross-validation over the training and validation sets. The numerical results reported in the following tables come from two different individuals (represented by P1 and P2), whose numerical results on test sets are typical among those collected across the whole set of individuals.

Table 1. Hyperparameters of the evaluated classifiers. q is the number of hidden neurons, η is the learning rate, α is momentum factor, γ is the scaling parameter of the Gaussian kernel in SVM and RKS, D is the dimension of the mapping in RKS.

Classifier	**Selected Hyperparameters**		
MLP	$q = 250$	$\eta = 0.05$	$\alpha = 0.75$
RKS	$D = 300$	$\gamma = 0.005$	-
ELM	$q = 280$	-	-
SVM	$C = 1000$	$\gamma = 0.005$	-
RVFL	$q = 250$	-	-

Table 2. Scenarios defined for the experiments

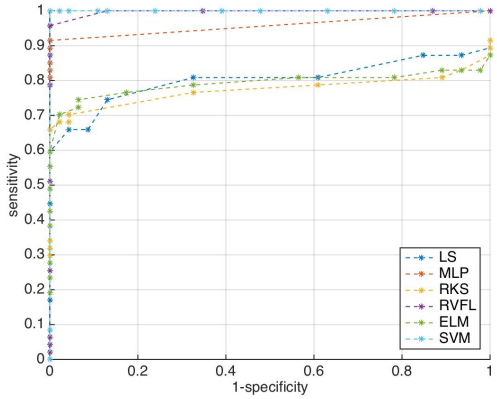
Patient	**Features**	**Classifier**					
		LS	**MLP**	**RKS**	**RVFL**	**ELM**	**SVM**
1	**Welch**	A	B	C	D	E	F
	LPC	G	H	I	J	K	L
2	**Welch**	M	N	O	P	Q	R
	LPC	S	T	U	V	W	X

We defined 24 scenarios (combination of feature extraction method and classifier) for the purpose of comparison of the results, as shown in Table 2. For each patient, 100 independent runs were performed for the 24 scenarios. At the end of each test phase, the ROC curve was constructed, and the accuracy, sensitivity, specificity and Matthews correlation coefficient (MCC) were computed from the corresponding confusion matrix.

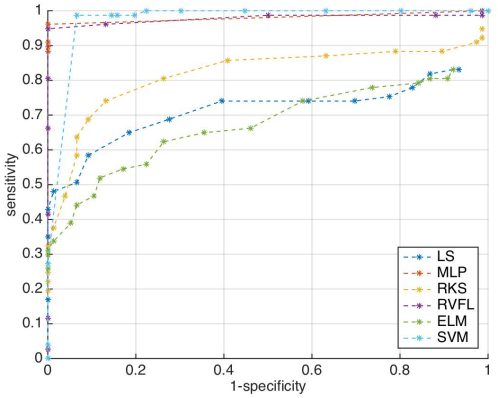
Table 3. Results of Patients 1 and 2 for the 24 simulation scenarios.

Patient 1												
Measures	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
Accuracy	0.806	0.899	0.814	0.973	0.785	0.988	0.990	0.992	0.981	0.994	0.981	0.996
Sensitivity	0.648	0.802	0.667	0.948	0.631	0.994	0.994	0.992	0.996	0.992	0.996	0.992
Specificity	0.972	1,000	0.967	1,000	0.946	0.983	0.987	0.991	0.965	0.996	0.965	1,000
MCC	0.654	0.819	0.665	0.949	0.606	0.977	0.981	0.984	0.963	0.988	0.963	0.992

Patient 2												
Measures	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**
Accuracy	0.734	0.940	0.809	0.970	0.823	0.934	0.981	0.977	0.951	0.994	0.931	0.991
Sensitivity	0.647	0.881	0.719	0.950	0.713	0.997	0.991	0.954	0.940	0.992	0.905	0.999
Specificity	0.822	1,000	0.901	0.991	0.937	0.870	0.970	1,000	0.963	0.995	0.958	0.983
MCC	0.478	0.887	0.633	0.942	0.666	0.877	0.962	0.955	0.904	0.987	0.865	0.982



(a) ROC curves (P1, Welch)



(b) ROC curves (P2, Welch)

Fig. 2. ROC curves for all classifiers using Welch's features for the two patients.

In Table 3 the average values of the evaluation measures for the two patients in the 24 scenarios are reported. It can be easily observed that, for all evaluation measures, the use of the LPC coefficients led to better performances of the classifiers (columns G to L for Patient 1, and S to X for Patient 2) when compared to the results achieved by the use of the Welch's periodogram.

A better way to visualize the differences in performance of the several classifiers for the two feature extraction methods and for the two patients is by means of the corresponding ROC curves. These curves are shown in Fig. 2 for the Welch's periodogram features and in Fig. 3 for LPC features.

We can infer from these figures several important conclusions. (*i*) First of all, the MLP and SVM classifiers always perform well, independent of the chosen feature extraction methods. (*ii*) The only randomized classifier whose performance

is equivalent to those of the MLP and SVM classifiers is the RVFL network. (*iii*) The performances of the LS, RKS and ELM classifiers are always inferior to those of the MLP, SVM and RVFL classifiers, no matter which feature extraction method is used. (*iv*) The performances of all classifiers (including the ELM/RKS/LS trio) improve when the LPC features are used for both patients (compare Figs. 3a and 3b).

As a general conclusion, we can state that we cannot take it for granted that the performances of randomized networks are always equivalent or superior to those of nonlinear classifiers. As we have shown the MLP and SVM classifiers performed quite well for the problem of interest. Among the randomized classifiers there were high variability among the results, with the RVFL classifier clearly achieving the best performance. in reality, the performance of the RVFL classifier was comparable to those presented by the MLP/SVM classifiers. With respect to these two classifiers, the RVFL classifier offers the additional advantage of very fast training, being this way a good alternative to them.

5 Conclusions and Further Work

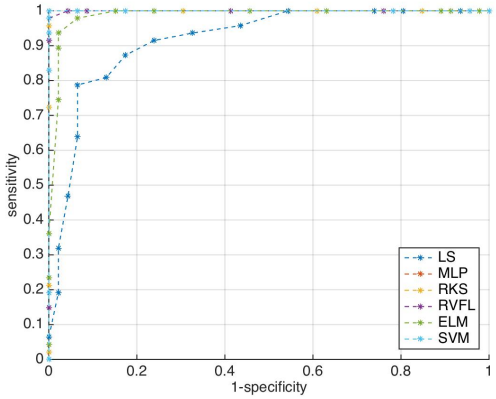
In this paper we have compared the performances of randomized classifiers with those achieved by standard nonlinear classifiers (SVM and MLP). the chosen task was the detection of epileptic seizures from EEG signals. We have shown that only one (the RVFL network) out of three of the most widely used randomized classifiers achieved performances comparable to those provided by the SVM and MLP classifiers. The additional advantage of fast training of the former makes it a respectable alternative to the latter ones.

In the current paper we evaluated only standalone classifiers. Currently, we are investigating the performances of the randomized classifiers evaluated in this paper in ensemble-like structures in order to compare their performances with that of the random forest classifier on a fair basis. We are also evaluating their performances using wavelet features.

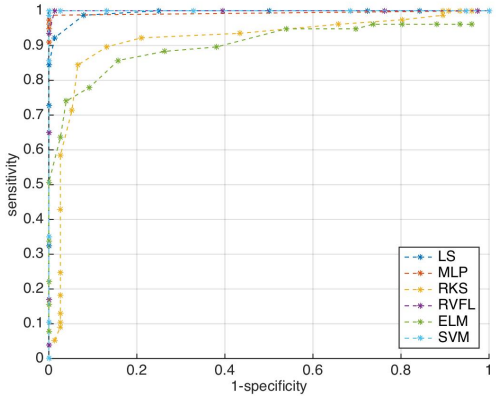
Acknowledgments: The authors thank CNPq (grant 309451/2015-9), CAPES and NUTEC for the financial support.

References

1. G.-H. P. Y.-H. Pao and D. J. Sobajic, "Learning and generalization characteristics of the random vector functional-link net," *Neurocomputing*, vol. 6, pp. 163–180, 1994.
2. L. Zhang and P. N. Suganthan, "A comprehensive evaluation of random vector functional link networks," *Information Sciences*, vol. 367–368, pp. 1094–1105, 2016.
3. G. Huang, G.-B. Huang, S. Song, and K. You, "Trends in extreme learning machines: A review," *Neural Networks*, vol. 61, no. 1, pp. 32–48, 2015.
4. B. Widrow, A. Greenblatt, Y. Kim, and D. Park, "The No-Prop algorithm: A new learning algorithm for multilayer neural networks," *Neural Networks*, vol. 37, pp. 182–188, 2013.



(a) ROC curves (P1, LPC)



(b) ROC curves (P2, LPC)

Fig. 3. ROC curves for all classifiers using LPC features for the two patients.

5. T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.
6. A. Rahimi and B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 1313–1320.
7. L. Zhang and P. N. Suganthan, "A survey of randomized algorithms for training neural networks," *Information Sciences*, vol. 364–365, pp. 146–155, 2016.
8. H. Adeli and S. Ghosh-Dastidar, *Automated EEG-Based diagnosis od neurological disorders: Inventing the Future of Neurology.* New York: CRC Press, 2010.
9. K. Lehnertz, F. Mormann, T. Kreuz, R. Andrzejak, C. Rieke, P. David, and C. Elger, "Seizure prediction by nonlinear EEG analysis," *IEEE Engineering in*

Medicine and Biology Magazine, vol. 22, no. 1, pp. 57–63, 2003.
10. D. P. Subha, P. K. Joseph, R. Acharya, and C. M. Lim, "Eeg signal analysis: A survey," *Journal of Medical Systems*, vol. 34, no. 2, pp. 195–212, 2010.
11. T. N. Alotaiby, S. A. Alshebeili, T. Alshawi, I. Ahmad, and F. E. A. El-Samie, "EEG seizure detection and prediction algorithms: a survey," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 183, 2014.
12. Y. Wang, Z. Li, L. Feng, C. Zheng, and W. Zhang, "Automatic detection of epilepsy and seizure using multiclass sparse extreme learning machine classification," *Computational and Mathematical Methods in Medicine*, vol. 2017, no. ID 6849360, pp. 1–10, 2017.
13. S. Ding, N. Zhang, X. Xu, L. Guo, and J. Zhang, "Deep extreme learning machine and its application in EEG classification," *Mathematical Problems in Engineering*, vol. 2015, no. ID 129021, pp. 1–11, 2015.
14. H. Zhao, X. Guo, M. Wang, T. Li, C. Pang, and D. Georgakopoulos, "Analyze EEG signals with extreme learning machine based on PMIS feature selection," *International Journal of Machine Learning and Cybernetics*, pp. 1–7, 2015.
15. C. Donos, M. Dumpelmann, and A. Schulze-Bonhage, "Early seizure detection algorithm based on intracranial EEG and random forest classification," *International Journal of Neural Systems*, vol. 25, no. 5, pp. 1–11, 2015.
16. A. Rahimi and B. Recht, "Uniform approximation of functions with random bases," in *Proceedings of the 46th Annual Allerton Conference on Communication, Control, and Computing*, 2008, pp. 555–561.
17. ——, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 1177–1184.
18. D. J. Sutherland and J. Schneider, "On the error of random fourier features," in *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI'2015)*, 2015, pp. 862–871.
19. A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, Harvard University–MIT Division of Health Sciences and Technology, 2009.
20. A. Shoeb and J. Guttag, "Application of machine learning to epileptic seizure detection," in *Proceedings of the 27th International Conference on Machine Learning (ICML'2010)*, 2010, pp. 1–8.
21. P. D. Welch, "The use of the fast fourier transform for the estimation of power spectra," *IEEE Transactions on Audio Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.
22. C. W. Therrien, *Discrete Random Signals ans Statistical Signal Processing.* New Jersey: Prentice-Hall, 1992.
23. C. Logar, B. Walzl, and H. Lechner, "Seizure prediction by nonlinear EEG analysis," *Role of long-term EEG monitoring in diagnosis and treatment of epilepsy*, vol. 34, no. Suppl 1, pp. 29–32, 1994.