Classification of Events in Distribution Networks using Autonomous Neural Models

André E. Lazzaretti, Vitor H. Ferreira, Hugo Vieira Neto, Rodrigo J. Riella and Julio Omori

***Abstract*—This paper presents a method for automatic classification of faults and events related to quality of service in electricity distribution networks. The method consists in preprocessing event oscillographies using the wavelet transform and then classifying them using autonomous neural models. In the preprocessing stage, the energy present in each sub-band of the wavelet domain is computed in order to compose input feature vectors for the classification stage. The classifiers investigated are based in Multi-Layer Perceptron (MLP) feed-forward artificial neural networks and Support Vector Machines (SVM), which automatically promote input selection and structure complexity control simultaneously. Experiments using simulated data show promising results for the proposed application.**

***Index Terms*—Quality of service event classification, wavelet transform, input selection, model complexity, Bayesian inference applied to Multi-Layer Perceptrons, Support Vector Machines.**

I. Introduction

DUE to the growing use of power electronics and industrial control devices, there is a raise in concerns – mainly on the side of power service providers – with events related to quality of service. These events are characterized by alterations in voltage or current waveforms, resulting from short-circuits of bad equipment operation in consumer installations [1]. Alterations in waveforms can sometimes result in damages to the consumers, forcing energy providers to monitor power provisions in order to ensure that it is adequately delivered [2] in the various stages of the electric power system.

Having this reality in mind, the Energy Company of Paraná (COPEL) and the Institute of Technology for Development (LACTEC) have developed a project under the Research and Development Program from the Brazilian Electricity Regulation Agency (ANEEL), entitled *Bar Feeder Quality of Service Monitoring System*, whose objective is to provide continuous monitoring of bar feeder voltages in power distribution substations, logging oscillographies of quality of service events [3].

The detection of events is done based on previously defined threshold values for voltage levels, harmonic distortion and voltage unbalance, i.e. every time that some voltage waveform exceeds some specified threshold, the referred waveform is logged for later analysis. The difficulty in this method concerns the large amounts of data logged by the monitoring systems, which turn out to be unfeasible to be manually analyzed by maintenance and protection engineers, resulting in a lack of adequate labeling of events in the majority of registered oscillographies.

The lack of ground-truth information (labeled data) ends up bringing difficulties to the development of automatic oscillography classifiers, for there is no way of using supervised learning algorithms without prior data classification for training, and this is the main reason for the simulation method employed in this work.

Among several events that happen in distribution networks, the class that has a higher impact in the context of power quality are faults in the electric power system [1], 70% of which happen in the distribution system [4]. Therefore, we are interested in this work in the diverse fault cases that may happen in distribution networks, as well as other quality of service events that typically happen in the particular substation chosen as the basis for simulations.

Initially, we built a simplified model of a distribution substation and its feeders using the Alternative Transient Program (ATP) [5] with the intent of generating a database of labeled oscillographies that could be used to train, validate and test our classification scheme. The use of ATP for this purpose is well-known and established, and we trust that the results obtained with simulated data will serve as a basis for the development of a classifier that will be able to work reliably on real data.

Regarding the classification process, the method chosen to be used in this work is already being discussed in several papers in the field [6]–[8] and consists in three classical steps: preprocessing, classification and post-processing [9]. In the preprocessing step, relevant features are extracted from the signals under analysis, using a transformation from time domain to frequency domain via Fourier transform, or to time-scale domain, via wavelet transform. Following that, the extracted features are fed to the classification step based on artificial neural networks, which is usually sensitive to specific characteristics of the event dataset and the preprocessing method. Finally, the post-processing stage usually involves fuzzy logic in order to reach a decision, in which a confidence level is attributed to the neural network output for each class of the problem.

In this work we decided to base the preprocessing on the discrete wavelet transform because of the resulting reduced feature vector size when compared with the feature vector size when using the Fourier transform. The idea here is to achieve efficient feature extraction without compromising classification performance [10]. Next, in the classification

A. E. Lazzaretti and R. J. Riella are with the Institute of Technology for Development (LACTEC), Avenida Comendador Franco 1341, Curitiba-PR, Brazil.

V. H. Ferreira is with the Federal University Fluminense (UFF), Rua Passo da Pátria 156, Niterói-RJ, Brazil.

H. Vieira Neto is with the Federal University of Technology – Paraná (UTFPR), Avenida Sete de Setembro 3165, Curitiba-PR, Brazil.

J. Omori is with the Energy Company of Paraná (COPEL), Rua José Izidoro Biazetto 158, Curitiba-PR, Brazil.

step, we compared two approaches based on artificial neural networks, Multi-Layer Perceptrons and Support Vector Machines, which include automatic input relevance selection and structure complexity control (number of hidden layer neurons for the MLP, and hyperparameters C and σ_i^2 for the SVM) [11], [12]. In our case, it was not necessary to use any post-processing step, as the classification methods themselves yield the desired output for the process as a whole.

The paper is divided as follows. In section II we present the theoretical aspects involved in our method. Section III describes our experimental setup, from modeling to classification. In section IV we present the results obtained in the classification of simulated events, and finally in section V we summarize our conclusions and propose future work.

II. Theoretical Aspects

A. Discrete Wavelet Transform

The discrete wavelet transform (DWT) consists in a time-scale signal representation through a decomposition in frequency sub-bands, using orthonormal bases obtained from digital filter banks. In the DWT, filters with different cut-off frequencies are used in order to analyze the signal in multiple scales – the signal is processed by a series of low-pass and high-pass filters. These filters separate low and high frequency components in different subspaces.

The construction of the filters is characterized by the properties of the wavelet function that is used, as defined in [13] and [14]. Therefore, orthonormal bases of discrete wavelet functions are associated not only to the mother wavelet function $\Psi(x)$, but also to the scale function $\phi(x)$, also known as father wavelet. The mother wavelet function is associated to details, or high-pass filters, whereas the scale function is related to approximations, or low-pass filters, forming an orthonormal basis.

Filtering operations are responsible for resolution analysis, while scale analysis is done using downsampling and upsampling operations during decomposition and reconstruction, respectively. The great advantage of the DWT is that the direct summation of approximation and detail coefficients, after filtering and upsampling, results in the original signal with very few losses.

Decomposition of the input signal into approximation and detail coefficients is the foundation of multiresolution analysis [14] and can be done using a pair of finite impulse response filters – a high-pass filter H and a low-pass filter L for the decomposition process, and their conjugates H' and L' for the reconstruction process. The wavelet decomposition procedure can be repeated, generating additional decomposition levels. At each level, the filtering and downsampling operations result in half the number of samples, consequently presenting half temporal resolution and double frequency resolution.

Once the signal is decomposed, the most prominent frequency components result in high amplitudes in the DWT sub-bands that include these particular frequencies, as it happens in the Fourier transform (FT). However, the main difference between the DWT and the FT is that the temporal localization of the frequency components is not lost in the former.

The temporal localization of a frequency component has a resolution that depends on its decomposition level – the higher the frequency component under analysis, the lower its decomposition level and, consequently, the higher its resolution in the time domain. Therefore, the wavelet decomposition procedure presents good time domain resolution for high frequency components and good frequency domain resolution for low frequency components. These properties not only make an alternative spectral representation to the one given by the FT possible, using nonlinearly spaced frequency sub-bands, but also the temporal localization of specific components of the signal under analysis.

B. Multi-Layer Perceptron

Once the structure of the MLP neural network is defined (i.e. number of hidden layers, number of neurons in each layer, and type of neuron activation function) and given the set of input-output pairs $D = \{X, Y\}$, $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, $Y = \{d_1, \ldots, d_N\}$, the goal of training the model according to a Bayesian inference approach consists in determining the parameter vector $\boldsymbol{w}$ that maximizes the *a posteriori* probability $p(\boldsymbol{w}|X, Y)$ given by:

$$p(\boldsymbol{w}|X, Y) = \frac{p(Y|\boldsymbol{w}, X)p(\boldsymbol{w}|X)}{p(Y|X)}, \tag{1}$$

where $p(\boldsymbol{w}|X)$ represents the *a priori* probability of $\boldsymbol{w}$, $p(Y|X)$ is a normalization factor and $p(Y|\boldsymbol{w}, X)$ is a likelihood function, which is related to the probability distribution of $\boldsymbol{x}_i$ being in a given class. Considering a dichotomy, the likelihood function is given by:

$$p(Y|\boldsymbol{w}, X) = \prod_{i=1}^{N} f(\boldsymbol{x}_i, \boldsymbol{w})^{d_i}[1 - f(\boldsymbol{x}_i, \boldsymbol{w})]^{1-d_i}, \tag{2}$$

where $d_i \in [0, 1]$, with $f(\boldsymbol{x}_i, \boldsymbol{w})$ corresponding to the MLP output as follows:

$$f(\boldsymbol{x}_i, \boldsymbol{w}) = \frac{1}{1 + \mathrm{e}^{\left[-\sum_{i=1}^{M} w_i \varphi_i \left(\sum_{j=1}^{N} w_{ij} x_j + b_i\right)\right]}}. \tag{3}$$

In equation (3), $\varphi_i(a) : \mathbb{R} \to \mathbb{R}$ is the activation function of the i^{th} hidden layer neuron, b_i is the neuron bias, M is the number of neurons in the hidden layer, and $\boldsymbol{w}$ is a vector including all model parameters, i.e. synaptic weights and bias.

In order to choose $p(\boldsymbol{w}|X)$, one must take into account the fact that it is expected that different sets of weights – for instance, weights connecting the input to the model and weights connecting hidden layer neurons to the output – present different behaviors during the estimation process. Having said that, it is reasonable to use specific distributions for each set of weights, defining the *a priori* distribution as follows:

$$p(\boldsymbol{w}|X) = \prod_{i=1}^{g} p(\boldsymbol{w}_i) \quad (4)$$
$$= \frac{1}{\prod_{i=1}^{g} \left(\frac{2\pi}{\alpha_i}\right)^{\frac{M_i}{2}}} \mathrm{e}^{\left(-\frac{1}{2}\sum_{i=1}^{g} \alpha_i \|\boldsymbol{w}_i\|^2\right)},$$

where $\boldsymbol{w}_i$ represents the set containing M_i parameters, with α_i corresponding to the hyperparameter given by the inverse of the zero-mean Gaussian distribution variance used for the *a priori* representation of $\boldsymbol{w}_i$, and g is related to the number of parameter sets. A specific set of model parameters, including a set of weights that connect the input to the model, originates an automatic input relevance estimation process [15].

Having the distributions $p(Y|\boldsymbol{w}, X)$ and $p(\boldsymbol{w}|X)$, the *a posteriori* probability $p(\boldsymbol{w}|X, Y)$ is given by:

$$p(\boldsymbol{w}|X, Y) = \frac{1}{Z_S} \mathrm{e}^{-S(\boldsymbol{w})}, \quad (5)$$

with

$$S(\boldsymbol{w}) = -\sum_{i=1}^{n} d_n \ln f(\boldsymbol{x}_i, \boldsymbol{w}) + (1 - d_n) \ln[1 - f(\boldsymbol{x}_i, \boldsymbol{w})] + \frac{1}{2}\sum_{i=1}^{g}\left(\alpha_i \sum_{j=1}^{M_i} w_{ij}^2\right), \quad (6)$$

where $Z_s = \int \mathrm{e}^{-S(\boldsymbol{w})} \mathrm{d}\boldsymbol{w}$ is a normalization factor. Hence, in order to maximize $p(\boldsymbol{w}|X, Y)$, it is necessary to minimize $S(\boldsymbol{w})$. For that we use an evidence approximation approach, which besides yielding estimates for $\boldsymbol{w}$ also yields estimates for α_i, which can be compared to an empirical relevance threshold for the identification and removal of irrelevant inputs. Further details about the estimation of $\boldsymbol{w}$ and α_i can be found in [15] and [16]. The empirical definition of relevance thresholds is explained in more details in [11] and [12].

Having the estimates of $\boldsymbol{w}$ and α_i, a model evidence maximization approach can be used for the analytic assessment of the goodness of fit of nested model sets, i.e. MLPs with different numbers of hidden layer neurons, to the data $D = \{X, Y\}$. Such assessment is computed using the model evidence, whose logarithm is given by the following expression:

$$\ln p(Y|H_h) = -S(\boldsymbol{w}) - \frac{1}{2}\ln|\mathbf{A}(\boldsymbol{w})| + \frac{1}{2}\sum_{i=1}^{g} M_i \alpha_i + 2\ln m + \ln m! + \frac{1}{2}\sum_{i=1}^{g} \ln\left(\frac{2}{\gamma_i}\right) + \frac{1}{2}\ln\left(\frac{2}{N-\gamma}\right), \quad (7)$$

where γ_i is the effective number of estimated parameters for the i^{th} set of weights $\boldsymbol{w}_i^* = [w_{i1}^*, \ldots, w_{iM}^*]^T$ and γ is the effective number of estimated parameters for the model, which are given by:

$$\gamma_i = \alpha_i \sum_{j=1}^{M_i} (w_{ij}^*)^2; \ \gamma = \sum_{i=1}^{g} \gamma_i. \quad (8)$$

C. Support Vector Machines

Support vector machines were originally developed to solve classification problems via the concept of an optimum separation hyperplane, which maximizes the separation margin ρ between classes. The motivation for maximizing ρ is based on a complexity measurement known as Vapnik-Chervonenkis (VC) dimension [17], whose upper limit is inversely proportional to ρ. Mathematically, the output of an SVM can be expressed as:

$$f(\boldsymbol{x}, \boldsymbol{W}, b) = \mathrm{sgn}[\boldsymbol{W}^T \boldsymbol{\Phi}(\boldsymbol{x}) + b]; \quad (9)$$
$$\boldsymbol{W} = [W_1, \ldots, W_N]^T;$$
$$\boldsymbol{\Phi}(\boldsymbol{x}) = [\phi_1(\boldsymbol{x}), \ldots, \phi_N(\boldsymbol{x})]^T,$$

where $\boldsymbol{\Phi}(\boldsymbol{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^N$ represents the nonlinear input mapping in feature space, with $\boldsymbol{W}$ is the parameter set of the model, b is the bias and sgn$[a]$ is the signal function:

$$\mathrm{sgn}[a] = \begin{cases} 1, \text{ if } a \geq 0 \\ 0, \text{ if } a < 0 \end{cases} \quad (10)$$

Maximization of the separation margin ρ can be formulated via the following restrict optimization problem:

$$\min_{\boldsymbol{W}, b, \boldsymbol{\xi}} E_s(\boldsymbol{W}) = \frac{1}{2}\boldsymbol{W}^T\boldsymbol{W} + C\sum_{i=1}^{N} \xi_i \quad (11)$$

s.t.

$$\begin{cases} d_i[\boldsymbol{W}^T \boldsymbol{\Phi}(\boldsymbol{x}) + b] \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}, i = 1, 2, \ldots, N$$

In equation (11), the first term of the objective function is responsible for the complexity control of the model by means of maximization of ρ. The second term is related to the classification error for the dataset, as for correctly classified data ξ_i is equal to zero. The hyperparameter C is responsible for the balance between model complexity and goodness of fit to the training data, and therefore is denominated regularization parameter [18].

The quadratic optimization problem in equation (11) can be solved by the method of Lagrange multipliers, whose dual formulation is given by:

$$\max_{\boldsymbol{\alpha}} \Psi(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_i d_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)\alpha_i\alpha_j \quad (12)$$

s.t.

$$\begin{cases} 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{N} \alpha_i d_i = 0 \end{cases}, i = 1, 2, \ldots, N,$$

where $\boldsymbol{\alpha}$ represents the set of Lagrange multipliers and $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ the dot product kernel in feature space, as follows:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = [\boldsymbol{\Phi}(\boldsymbol{x_i})]^T \boldsymbol{\Phi}(\boldsymbol{x_j}). \tag{13}$$

There are several types of kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, which must abide to the conditions of Mercer's theorem. In this work, we use the Gaussian kernel given by:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathrm{e}^{\left[-\sum_{l=1}^{N} \sigma_i^2 (x_{il} - x_{jl})^2\right]}, \tag{14}$$

where σ_i^2, $l = 1, 2, \ldots, N$ are the kernel hyperparameters.

At the optimal location of equation (12) not all α_i^* are nonzero. The vectors for which α_i^* do not equal to zero are the so called support vectors, which define the decision surface of the SVM as follows:

$$f(\boldsymbol{x}, \boldsymbol{W}, b) = \mathrm{sgn}\left[\sum_{i=1}^{N_S} \alpha_i d_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right], \tag{15}$$

where N_S is the number of support vectors.

Despite being concerned with the control of structure complexity in their formulation and yielding the model structure as a subproduct of the training algorithm through the number of support vectors, the SVM have some hyperparameters that must be specified by the user, such as the regularization constant C and the kernel hyperparameters σ_i^2. These parameters are commonly selected via cross-validation and in this work we select them by means of minimization of the upper limit of the estimated generalization error in a leave-one-out approach. This resampling method yields an almost non-biased estimate for the generalization ability of the model [19], but is computationally intensive. On the other hand, the upper limit used in this work was analytically developed in [20] and is conceptually founded on the span of support vectors, as given by the expression:

$$T[f(\boldsymbol{x}, \boldsymbol{W}, b)] = \sum_{i=1}^{N_S} \alpha_i S_i^2, \tag{16}$$

where S_i^2 represents the extension of the i^{th} support vector, given by:

$$S_i^2 = \frac{1}{(\tilde{\mathbf{K}}^{-1})_{ii}}, \tag{17}$$

where $(\tilde{\mathbf{K}}^{-1})_{ii}$ represents the i^{th} diagonal element of the inverse matrix of $\tilde{\mathbf{K}}$, as follows:

$$\tilde{\mathbf{K}} = \begin{bmatrix} \mathbf{K}_{N_S} & \boldsymbol{u} \\ \boldsymbol{u}^T & 0 \end{bmatrix}, \tag{18}$$

where $\mathbf{K}_{N_S}$ is the dot product kernel matrix for all support vectors and $\boldsymbol{u} \in \mathbb{R}^{N_S}$ is the unitary vector. Given the multimodal characteristics of $T[f(\boldsymbol{x}, \boldsymbol{W}, b)]$ [21], confirmed for regression problems in [11] and [12], in this work we use genetic algorithms [22] for minimization of $T[f(\boldsymbol{x}, \boldsymbol{W}, b)]$ and estimation of C and σ_i^2. The analysis of the estimated σ_i allows the implementation of a method for measuring the relevance of the inputs in the calculation of the output. The artificial proof variables insertion allows the estimation of evidence in significance empirical thresholds, which are used for identification and removal inputs irrelevant. More details can be found in [11].

III. Experimental Setup

As described earlier, the basic method used in this work consists in building a database of simulated events using ATP and then use it for the classification process, which is composed by a preprocessing stage and the classification stage itself. The criteria used in the modeling of events, as well as their preprocessing and classification are described as follows.

A. ATP Modeling

In the basic ATP model used for the generation of simulated events with different parameters, we have considered the basic elements of a distribution substation from COPEL, as well as the structures needed for the simulation of the events of interest. These elements are: substation transformer, capacitor bank, grounding transformer, four bar feeders, and the equivalent of the electric circuit up to the substation transformer. Only the main branch of the feeders was modeled, being segmented in several parts, respecting wire gauges in each part. All parameters used in the model were based on real data obtained from COPEL. We also inserted elements used to simulate events: single-line-to-ground faults (A, B and C), two-phase-line-to-ground faults (AB, BC and CA), three-phase-line-to-ground faults, two-phase faults, three-phase faults, feeder circuit breaker switch-off, automatic feeder reclosing, capacitor bank switching, and signals with no apparent disturbances. These are the most common events monitored by COPEL in the substation that was modeled.

The instant and the location of the events generated were varied and when concerning short-circuits, the fault resistance was also varied. Moreover, we considered feeder load cycles to simulate events for different load values. Finally, we considered the typical harmonic distortion of the substation in order to model the system operation as close to reality as possible. We emphasize that every event instance was automatically generated from a single ATP base file, adding up to 6480 cases (432 cases for each event type, for training and test datasets).

B. Preprocessing

Aiming to extract relevant features from the signals under analysis, our preprocessing stage computed the energy within several sub-bands of the DWT.

Initially, the voltage signals from the three substation phases were decomposed into ten DWT levels using (Daubechies mother wavelet – *'db8'*), originating ten detail components and the approximation component for each phase, adding up to 33 components in the wavelet domain. We decided to use ten decomposition levels because the sampling rate was 7680Hz – with this amount of levels we are able to represent most of the frequency content of the voltage signals with appropriate resolution.

After decomposing the input signals, we calculated the energy contents in each DWT level. This calculation is done based on the energy contents before and during the event, i.e. we obtain for each DWT level the energy ratio between the cycle where the fault occurred and the cycle immediately before. The choice for just one cycle in our method avoids the interference of any protection actuation in the distribution system, because protection actuation would take at least one cycle to become effective.

The energy ratio used in this work is given by the following expression:

$$E = \frac{1 + E_{DF}}{1 + E_{BF}} - 1, \tag{19}$$

where E_{DF} and E_{BF} are the wavelet energy levels of the cycles during and before the fault, respectively.

The formulation above is a modification of the direct energy ratio E_{DF}/E_{BF}, so that divisions by zero are avoided. We assume normalized energy levels, so that $E \in [-1, 1]$. Using this technique, it was possible to identify precisely in which DWT sub-bands there was an increase or decrease on energy contents, which characterize the events. This property was not observed when using the direct energy ratio E_{DF}/E_{BF}, which was very susceptible to noise. In order to use equation 19, it is necessary to normalize the sum of the energy contents in every DWT level to one.

C. Classification

Two methods were compared when using the MLP and SVM approaches in the classification stage, following the theoretical approaches previously described and using the 33 input values generated in the preprocessing stage. These two methods differ with respect to how they were trained and validated.

The first classification method operates directly on the 15 classes of simulated events, i.e. we measured the performance of the classifiers for training and test datasets using the 6480 samples from all 15 existing classes directly, for both MLP and SVM approaches. On the other hand, the second classification method operates in three stages. The first stage separates faults from the remaining types of event. From this initial separation, two other independent classifiers are used, one specifically to faults and the other to the remaining types of event. Each classifier was assessed using separate training and validation datasets, and after that, the whole three-stage classification system was globally assessed using a different test dataset. The second classification method was used also used for both MLP and SVM approaches.

For the datasets used in both classification methods, we varied the instant when the events occurred (from 0° to 180°), the load cycle (the most expressive loads for each feeder during the day were selected), the distance where the events occurred (near and far from the substation), and in the case of faults, we also varied the fault resistance (between 5Ω and 2000Ω in the training set, and between 20Ω and 1700Ω in the test set). The fault resistance values that were used are in the typical range of the events that happen in reality for the substation being simulated.

IV. Results and Discussions

Table I presents the classification results using the MLP approach for the single-stage classifier and the three-stage classifier. Model characteristics – number of hidden layer neurons and selected inputs – are also presented for each classifier stage. The partial accuracy represents the independent performance for each classifier independently of the classification process as a whole, while the total accuracy for stages 2 and 3 takes into account the classification errors introduced by stage 1.

TABLE I
MLP classifier validation results

Classifier	Partial Accuracy	Total Accuracy	Number of Neurons	Number of Inputs
Single (15 classes)	—	86.3 %	19	33
Stage 1 (2 classes)	—	96.1 %	15	33
Stage 2 (4 classes)	89.7 %	85.2 %	17	32
Stage 3 (11 classes)	93.1 %	92.1 %	18	33

The equivalent results for the SVM approach are presented in table II. In this case, model characteristics are represented by the average number of support vectors and selected inputs. Average values were used because SVM training is performed using pairs of classes and for each pair there is a resulting number of support vectors and selected inputs. This makes the presentation of detailed partial results unfeasible in this paper and so we use average values to give an approximate idea of the complexity of the SVM models in terms of support vectors (equivalent to the number of hidden layer neurons in the MLP approach) and relevant inputs for the classification process.

TABLE II
SVM classifier validation results

Classifier	Partial Accuracy	Total Accuracy	Average Number of Support Vectors	Average Number of Inputs
Single (15 classes)	—	89.5 %	68	22
Stage 1 (2 classes)	—	92.3 %	1191	21
Stage 2 (4 classes)	93.8 %	87.6 %	153	23
Stage 3 (11 classes)	86.1 %	78.6 %	55	26

The analysis of tables I and II summarizes the main characteristics of the methods for classification of events in distribution networks presented in this work. We emphasize the advantages of autonomous model selection taking into account the most relevant aspects of the data presented during training, making the classification process independent of human intervention. Besides that, selected inputs can be used as a basis for the analysis of which DWT sub-bands are the most relevant for event classification. It can be noted from table I that the MPL approach considered all DWT sub-bands relevant for classification, indicating that our preprocessing method is adequate to represent relevant features of events in distribution networks. The average number of inputs selected by the SVM approach in table II also confirms this hypothesis – approximately 2/3 of the inputs, in average, were considered relevant in this case.

Table III shows a summary of the test results for each approach that was investigated, considering the global performance of three-stage classifiers for all 15 classes of events.

TABLE III
MLP AND SVM CLASSIFIER TEST RESULTS

Classifier	Total Accuracy
Single-stage MLP	86.3 %
Three-stage MLP	87.1 %
Single-stage SVM	89.5 %
Three-stage SVM	91.8 %

It can be noticed that the use of three separate classification stages results in a marginal improvement in the performance of both MLP and SVM approaches. Moreover, the use of a first classification stage in order to separate faults from the remaining types of events offers more flexibility in the analysis of events of the same nature in the following classification stages.

V. CONCLUSION

This paper demonstrated the feasibility of employing autonomous neural models (MLP and SVM) for the task of event classification in power distribution networks. The autonomy of the models can be understood as the ability of automatically controlling model complexity while simultaneously selecting relevant inputs. These features allow systems that use such approaches to operate almost independently from prior steps in the classification process, such as the preprocessing stage and the definition of training and test datasets, which depends on the distribution substation in which data acquisition is to be performed.

Our work also shows that the subdivision of the main classifier in three stages yields an improvement in classification accuracy as a whole. Moreover, the use of a dedicated classifier stage for faults allows the implementation of fault localization techniques based on the oscillographies used in the classification, providing useful information for maintenance teams of the power service provider. In future work we intend to use SVM input selection information in order to investigate the relevance of DWT sub-bands for signal representation of particular events.

REFERENCES

[1] J. Morais, Y. Pires, C. Cardoso, and A. Klautau, "Data mining applied to the electric power industry: Classification of short-circuit faults in transmission lines," in *Proceedings of the 7th International Conference on Intelligent Systems Design and Application*, October 2007, pp. 943–948.

[2] ANEEL, *Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional – PRODIST, Módulo 8 – Qualidade de Energia Elétrica*, Agência Nacional de Energia Elétrica, 2007, in Portuguese.

[3] R. J. Riella, V. P. Ferrari, G. Paulillo, M. R. Ortega, and J. G. Pereira, "Desenvolvimento de um sistema de monitoramento contínuo da qualidade da energia elétrica para subestações de distribuição," in *Anais do XVIII Seminário Nacional de Distribuição de Energia Elétrica*, October 2008, in Portuguese.

[4] R. C. Dugan, M. F. McGranaghan, S. Santoso, and H. W. Beaty, *Electrical Power System Quality*, 2nd ed. McGraw-Hill, 2002.

[5] EMTP, *Alternative Transients Program (ATP) Rule Book*, Can Am EMTP User's Group, 1995.

[6] K. M. Silva, B. A. Souza, and N. S. D. Brito, "Fault detection and classification in transmission lines based on wavelet transform and ANN," *IEEE Transactions on Power Delivery*, vol. 21, no. 4, pp. 2058–2063, October 2006.

[7] O. Dag and C. Ucak, "Fault classification for power distribution systems via a combined wavelet-neural approach," in *Proceedings of the 2006 International Conference on Power System Technology*, vol. 2, November 2004, pp. 1309–1314.

[8] H. Zheng-You, C. Xiaoqing, and L. Guoming, "Wavelet entropy measure definition and its application for transmission line fault detection and identification," in *Proceedings of the 2006 International Conference on Power System Technology*, October 2006, pp. 1–6.

[9] J. Arrillaga, B. C. Smith, N. R. Watson, and A. R. Wood, *Power System Harmonics*. Wiley, 2003.

[10] O. Delmont Filho, "Um algoritmo para detecção, localização e classificação de distúrbios em qualidade da energia elétrica utilizando a transformada wavelet," Ph.D. dissertation, EESC/USP, São Carlos, 2007, in Portuguese.

[11] V. H. Ferreira and A. P. A. da Silva, "Toward estimating autonomous neural network load forecasters," *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 1554–1562, November 2007.

[12] V. H. Ferreira, "Desenvolvimento de modelos neurais autônomos para previsão de carga elétrica," Ph.D. dissertation, COPPE/UFRJ, Rio de Janeiro, 2008, in Portuguese.

[13] I. Daubechies, *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

[14] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Academic Press, 1999.

[15] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[16] D. J. C. Mackay, "Bayesian methods for adaptive models," Ph.D. dissertation, Caltech, Pasadena, 1992.

[17] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[18] V. Cherkassky and F. F. Mulier, *Learning from Data: Concepts, Theory and Methods*. Wiley, 1998.

[19] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2001.

[20] V. Vapnik and O. Chapelle, "Bounds on error expectation for support vector machines," *Neural Computation*, vol. 12, no. 9, pp. 2013–2036, September 2000.

[21] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 131–159, January 2002.

[22] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Professional, 1989.