Journal of Control, Automation and Electrical Systems (2020) 31:990–1000
https://doi.org/10.1007/s40313-020-00597-6

Sensibility Analysis of a Fault Location Method Based on ANN, WPT and Decision Tree in Distribution Systems

André Luís da Silva Pessoa[1] · Mário Oleskovicz[1] · Paulo Estevão Teixeira Martins[1]

Received: 5 August 2019 / Revised: 9 March 2020 / Accepted: 13 April 2020 / Published online: 7 May 2020

Abstract
Given the possibilities provided by smart grids in terms of communication infrastructure and information acquisition, there are new options on how to use the signals coming from meters to locate short circuits that occur in the system. This paper presents a framework for fault location in radial distribution systems based on machine learning algorithms and a multistage approach. A methodology in order to segment the system for proper identification of the outage region is presented. Studies are carried out involving the variation of the fault impedance, the fault incidence angle and the number and position of the meters. The IEEE 34-bus bar distribution feeder was considered for the tests. The results so far are promising, attesting and validating the presented methodology.

Keywords Machine learning · Meters positioning · Multistage approach · Fault location · Smart grids

1 Introduction

The current operation and management status of the electric power systems, with little or no automatism, represents an obstacle hindering a quick restoration of the power supply when facing certain adverse situations, such as interruptions caused by short circuits. In this sense, the development and application of concepts that permeate smart grids (SG) can be addressed as part of the solution (Goel and Agarwal 2015).

In the outlined context, SG creates an environment that is favorable to the development of several tools to improve the operation and management of the PS (power system) as a whole. The algorithms are one of these tools that are focused on the precise (physical) fault location.

It is worth emphasizing that the fault location algorithm is a supplementary step to the protection system (Saha et al. 2009). Therefore, fault locators must provide a quick and accurate fault location.

✉ André Luís da Silva Pessoa
alsp@usp.br

Mário Oleskovicz
olesk@sc.usp.br

Paulo Estevão Teixeira Martins
pauloetm@usp.br

[1] Department of Electrical and Computer Engineering, University of São Paulo, USP, São Carlos, Brazil

It is also worth noting that, among the main problems and challenges in fault location, especially for the distribution systems, the focus of this research is the existence of non-homogeneous lines; the presence of lateral branches (which may incur in multiple locations of faults), transformers with distinct relations, and the presence of single-phase and three-phase loads; limitations in measurements; dynamic topology; and the effect of the fault resistance, which must also be considered as shown in Bahmanyar et al. (2017).

For the distribution systems, fault location algorithms are generally divided into two basic groups, which are the algorithms that estimate the fault distance from a reference point in the PS, for example, the power substation, and methods that identify the outage area, pointing in which part of the system the fault occurred (Bahmanyar et al. 2017).

Regardless of the group, various articles have proposed alternatives for fault location based on intelligent techniques, among which Adewole et al. (2016), Rafinia and Moshtagh (2014), Dehghani and Nezami (2013), Zayandehroodi et al. (2013), Ray et al. (2015), Farias et al. (2016), Lovisolo et al. (2012), Lout and Aggarwal (2013), Zapata-Tapasco et al. (2014) and Pérez and Vásquez (2016) can be cited.

The research reported in Adewole et al. (2016) presents an approach to identify in which section of the system a fault occurred. It uses artificial neural networks (ANNs) to estimate the distance of the fault from the substation and extracts fault signal features by using the discrete wavelet transform

(DWT). In Rafinia and Moshtagh (2014), the authors also used an approach involving ANN, but in conjunction with fuzzy logic, to determine the fault type and estimate the distance of the fault from the substation. DWT was also used to extract the required features. In Dehghani and Nezami (2013), ANN was trained to estimate the fault distance from the substation. Zayandehroodi et al. (2013) and Ray et al. (2015) used ANN with radial basis function to estimate the fault distance. On the other hand, Farias et al. (2016) used a hybrid approach in which ANN was used to estimate parameters of an analytical model for estimating fault distance. ANN was used in Lovisolo et al. (2012) to estimate the areas where the fault may have occurred. The research reported in Lout and Aggarwal (2013) also used an ANN to identify in which region of the system the fault occurred, and a second ANN to estimate the fault distance. In Zapata-Tapasco et al. (2014), decision trees were used to identify the area of the fault. Pérez and Vásquez (2016) used support vector machines to identify areas where the fault occurred.

This paper proposes a multistage approach to identify the region of the fault occurrence, to develop a methodology that is as robust as possible when there are changes in the fault profile (location, impedance and incidence angle) and in the placement of the meters. For this purpose, the generalization capacity of the ANN is used to estimate the fault distance. Then, the response is used as one of the inputs of the decision trees that will estimate the outage region. As for the preprocessing of fault signal information, the wavelet packet transform (WPT) was used, which provides greater accessibility to the information on these signals when compared to DWT.

In addition to the fault location, tests were released changing the number and positions of meters.

2 Wavelet Packet Transform

The discrete wavelet transform (Saha et al. 2009) enables multiresolution wavelet analysis, in which the signal is decomposed into multiple frequency bands. At each step of the wavelet analysis, the signal is divided into two components. These components are obtained by low-pass and high-pass filters and by sampling the signal with a degree of two. In this multiresolution analysis, only the low frequency bands are processed. Unlike DWT, the wavelet packet transform (WPT) considers the processing of high-frequency signals as well. Figure 1 illustrates the approximations and details that are obtained at each WPT decomposition level. Thus, for example, for the fourth level of decomposition, there are 16 frequency ranges between approximations and details.

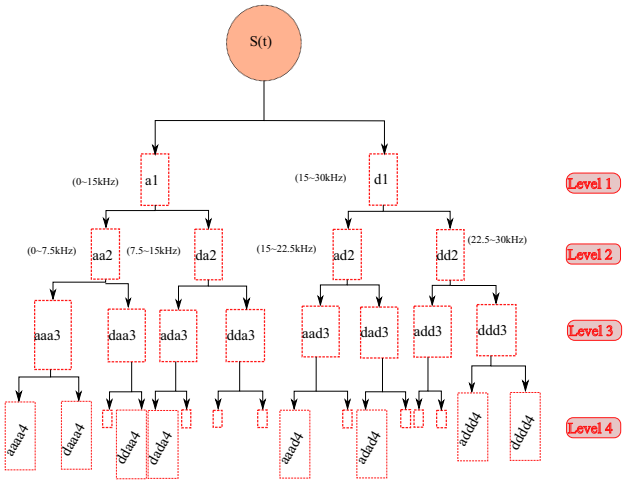

Fig. 1 Decomposing tree of orthogonal wavelet package

3 Artificial Neural Networks and Decision Trees

In this research, multilayer perceptron ANN was used through the toolbox available in the MATLAB software, known as *nntool*, with training by Levenberg–Marquardt (Adewole et al. 2016), to estimate the distance of the fault from the substation.

The decision tree is a good alternative to solve classification problems, as it allows to infer to which class a particular set of features is most correlated. To obtain the desired decision tree, binary divisions of the search space are recursively performed in order to segment it. The *Gini index* (James et al. 2013) is calculated to define the best binary division. It can be defined as:

$$G = \sum_{k=1}^{K} \hat{p}_{mk} \left(1 - \hat{p}_{mk}\right). \tag{1}$$

where K is the number of classes, and $\hat{p}_{mk}$ is the proportion of observations of the training set in region mth of the kth class. When the values of $\hat{p}_{mk}$ are close to 0 or 1, the *Gini index* tends to be small. Thus, when it is evident that a particular binary division has a small value of *Gini index*, it means that this division indicates groups with predominantly a single class. The decision trees, by their training process, enable the identification of the parameter (or parameters) that is the most relevant for the identification of classes in a given problem. Therefore, in this article, the decision trees were applied to infer which parameters/magnitudes are actually more relevant for solving the problem. The toolbox *Statistics and Machine Learning*, also available in MATLAB, was used to implement the decision trees. Decision trees classify the instances from the root of the tree to some node (leaf), which provides the classification of this instance (Mitchell

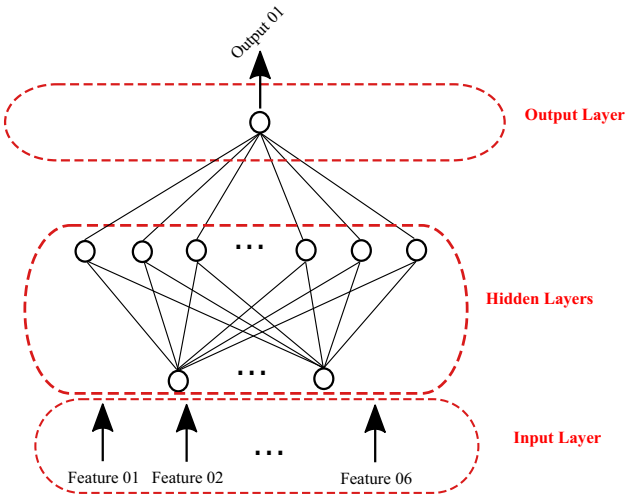

Fig. 2 ANN topology used in this paper

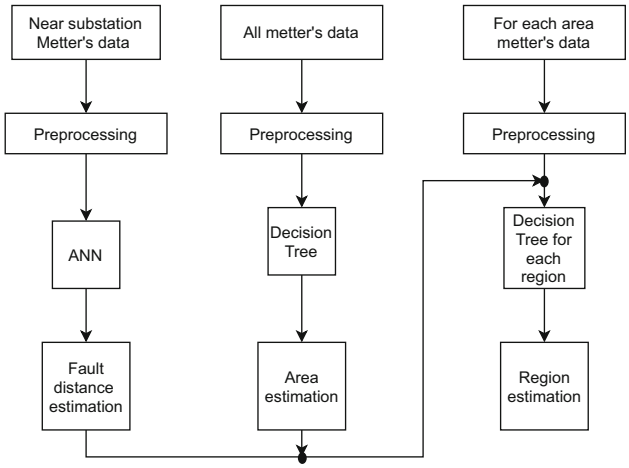

Fig. 3 Flowchart representing the main steps of the implemented methodology

et al 1997; James et al. 2013). Like most ANNs, decision trees also have supervised training (James et al. 2013). Therefore, the training database must contain the input parameters (signal information) and the desired outputs (area or region of the fault occurrence) for each input pattern. At the end of the training, the tree is formed by a series of IF–THEN conditions (disjunctive expressions) that lead to the area or region of the fault occurrence.

By way of illustration, ANNs with 6 inputs, 2 hidden layers and one output were used in this article. Figure 2 presents the ANN topology used in this paper. As shown below, the 6 input parameters of the ANNs are the RMS values of the three-phase voltage and current signals.

4 Algorithm for Fault Location

Figure 3 presents the general methodology of the fault location algorithm.

This framework is divided into four main stages. The first stage is the processing of the voltage and current signals (of the second post-fault cycle) that is used for locating the fault. In this preprocessing, a total of 114 parameters are obtained. The parameters were obtained based on the second post-fault cycle of the three-phase voltage and current signals by calculating the RMS value of these signals (6 parameters), by extracting the magnitude and phase angle of the fundamental component of these signals (12 parameters), extracting the approximations and details of the fourth WPT decomposition level using 4 db as the mother wavelet (plus 96 parameters: 6 signals, with 16 coefficients for the fourth decomposition level each). Thus, for the signals obtained from each of the meters present in the system, 114 parameters are obtained.

After processing the fault signals and obtaining the required parameters, the fault distance is estimated through multilayer perceptron networks.

The area of the fault occurrence (bigger delimited portion) is identified by decision trees (training algorithm—*classification and regression trees*). Finally, the parameters obtained by the processing of the signals, as well as the estimation of the fault distance and the identification of the fault area are used to identify in which region (more precise or better delimited portion) the fault occurred. All these stages are presented in this section.

4.1 Post-fault Signal Processing

This methodology is based on the context of SG and considers, therefore, the presence of intelligent meters installed in the system, and an entire communication infrastructure that enables the access to recorded voltage and current signals. In addition, sending these voltage and current signals to an operation center, where the fault location algorithms is executed, is also considered.

When the instant of the fault situation is properly detected and classified in one of its 11 possible types (single-phase, two-phase, two-phase-ground, three-phase and three-phase-ground faults), the three-phase voltage and current signals of the substation and intelligent meters are preprocessed. The detection and classification modules were not implemented in this research. Therefore, their information is considered as known.

For the preprocessing of the signals, the discrete Fourier transform (DFT) is used to obtain the amplitude and phase of the fundamental component. The calculation of the root mean square (RMS) value and the application of the WPT (considering Daubechies with support 4 as mother wavelet) are used to calculate the energy of the coefficients of the fourth level

of decomposition. Thus, in total, each measurement point enables the extraction of 114 parameters (6 RMS values, 12 amplitudes and phases of the fundamental frequency component, and 96 energies of the WPT coefficients of the 4th level of decomposition). A sample rate of 256 samples per cycle was used to process the fault signals. Only the second post-fault cycle was considered. It is worth noting that the first post-fault cycle was not used, as it has a higher presence of transients and exponential decay direct current component. In addition, the fault location is a complementary function to the protection of the electrical system. Therefore, as in most practical implementations, before activating the fault location process, this must be detected and spotted by the associated protection system. Then, in the case of permanent faults, the oscillographic records can only be retrieved and accessed for the fault location process after the protection system is activated.

The presence of measurement errors and/or lack of synchronism, although not considered in the sensitivity analysis presented in this article, are seen as elements that should be analyzed in future studies related to the context of smart grids.

4.2 Estimation of the Fault Distance Using ANN

After preprocessing the signals, the fault distance is estimated through multilayer perceptron neural networks using a Levenberg–Marquardt-based back propagation algorithm. It is worth mentioning that this distance estimation is used later in the segmentation algorithm to better define the regions of the system under analysis.

For the estimation of the fault distance, only the RMS values of the signals from the meter near the substation and the data of the phases directly involved in the disturbances are used, since the noninvolved phases are influenced by the load variation of the system. Thus, for example, for faults involving phase *A* with connection to ground, only the RMS values of the voltage and current signals of phase *A* are used as input parameter of the ANN.

Eleven (11) ANNs (one for each fault type) are used in total to estimate the fault distance. In addition, both input and output parameters were normalized within the range of -1 to 1. The activation functions used in the hidden layers and in the output layer were, respectively, the hyperbolic tangent and the linear function. Another important point was the use of an algorithm for an optimized search of the best topology for each of the 11 ANNs trained. This algorithm was based on an evaluation of ANN performance by varying the number of neurons in their hidden layers. Each topology was evaluated based on its error.

Each type of fault considered has a specific ANN associated. The topology of each of these ANNs was obtained by the following procedure:

1. Each initial topology has 2 hidden layers with 5 neurons each;
2. The ANN is trained with the training dataset and validated with the test set. Three different training sessions are performed with this topology, and the 95th percentile (P95) of the error obtained is stored for each training;
3. The lowest P95 value obtained in 3 trainings is stored;
4. Five neurons are kept in the first hidden layer, 2 neurons are sequentially added to the second hidden layer, up to a maximum of 25 neurons. Steps (2) and (3) are repeated every time two neurons are added;
5. Next, 2 neurons are added to the first hidden layer, and steps (2), (3) and (4) are repeated. The number of neurons in the first layer remains fixed;
6. The algorithm ends when the last topology with 25 neurons in the first hidden layer and 25 neurons in the second hidden layer are evaluated;
7. The topology with the lowest P95 value among all evaluated architectures is used.

It is worth noting that these ANNs will be trained only once. However, if there are considerable changes in the representative standards of the electrical system under consideration, a new training and testing/validation process addressing the new characteristics is probably necessary. Therefore, using them after their training does not require a high computational effort.

4.3 Identification of the Area and Region of the Fault

For a better understanding of the implemented methodology, two terms are initially defined. The first term is the area in which the fault occurred, and the second is the region of the fault occurrence. The area is a larger portion on the system in which the fault occurred. The region is a more physically delimited (better defined) portion, so that the physical location of the fault is more precise.

4.3.1 Defining the Areas

The areas have been defined according to the placement of the smart meters, which in this case are only considered in the main feeder. Figure 4 illustrates the definition of areas for an allocation of 3 m in the IEEE 34-bus bar distribution feeder.

4.3.2 Defining the Regions

The methodology characterizes the regions of a distribution system according to two criteria:

(i) each lateral branch is considered as a region; and

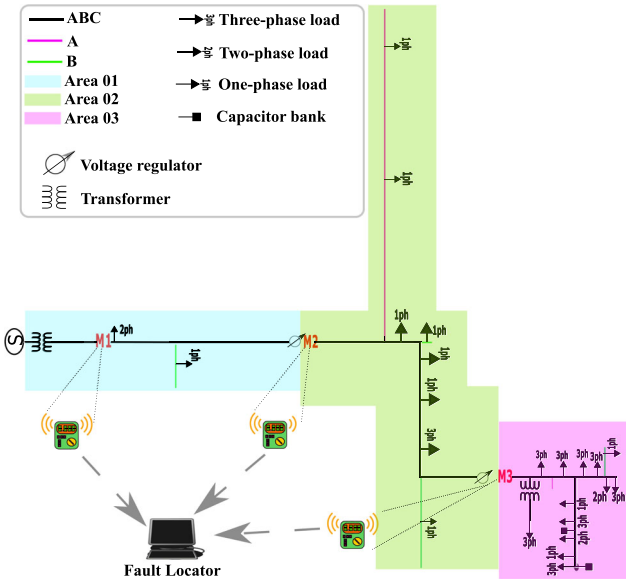

Fig. 4 IEEE 34-bus bar distribution feeder with 3 m installed

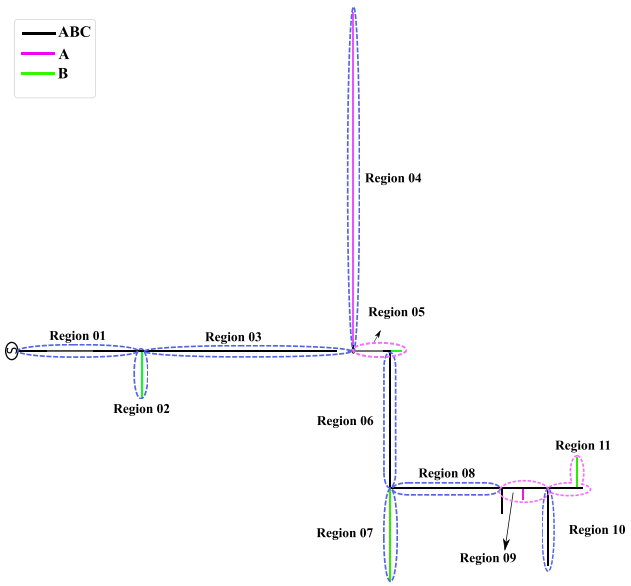

Fig. 5 IEEE 34-bus bar distribution feeder with the region definition

(ii) all portions of the main branch before a lateral branch (or between lateral branches) define a region.

According to the two criteria presented, the main element for the definition of the regions is the presence of lateral branches. However, to avoid an excessive number of regions, which may be small, regions much smaller than their neighbors are grouped. The clusters of these small regions are based on da Silva Pessoa et al. (2018).

Figure 5 presents the IEEE 34-bus bar distribution feeder with the defined regions. Eleven regions were considered for the fault location.

4.3.3 Decision Trees Applied in the Identification of the Area and Region of the Fault Occurrence

The method extracts all 114 parameters from the three-phase voltage and current signals and uses them both to estimate the area of occurrence of the fault and to determine in which region the fault occurred.

Regarding the configurations of the decision trees, those that will estimate the area as well as those estimating the region are identical. Therefore, the difference is basically in the input and output data, that is, in the training and validation database.

The constitution of the database used in this stage is presented hereinafter.

- **Database to identify the area of the fault occurrence:** For each type of fault, only one decision tree is designed to identify in which area of the system the fault occurred. To do so, the database for training and validation consists of the 114 parameters resulting from the signals of each meter installed in the system.
- **Database to identify the region of the fault occurrence:** This method requires one decision tree for each meter installed, in order to identify the region of the fault occurrence on the distribution system. The training and validation database consists of the 114 parameters resulting from the processing of the three-phase voltage and current signals. The information on the occurrence area of the fault and the estimation of the fault distance is then added. Hence, for each of these decision trees, there are 116 input parameters.

As shown, for each type of fault, regardless of the number of meters, a single decision tree is used to identify the outage area. For the identification of the outage region, the same number of decision trees as the number of meters installed in the system will be used.

Each of the decision trees that identifies the outage region reports on which region of the system the fault occurred. If most of these decision trees present the same response, the response of the majority is considered as final. If there is no consensus on the response presented by the trees, the response of the tree belonging to the area in which the fault occurred is adopted as the final response.

5 Case Study

In order to estimate the fault distance, a second fault cycle of three-phase voltage and current signals was considered.

Table 1 Number of simulated short circuits for the training and test database for each of the 11 fault types

Fault type	Training dataset	Test dataset
A-ground	6936	1248
B-ground	6768	1278
For each of the other fault types	5712	1038

As for the fault positions, these were applied in distribution system branches with less than 4000 m, at 1%, 25%, 50%, 75% and 99 % of their length.

The short-circuit situations were applied at every 1000 m for lines longer than 4000 m. In addition to the fault position variation, fault impedances were also considered with values of 0.0001 Ω, 10 Ω, 20 Ω, 30 Ω, 40 Ω and 50 Ω, and incidence angles of 0, 30, 60 and 90 degrees to generate a representative dataset to train the algorithms.

To generate a test dataset, faults applied at every 2,000 m, with fault impedance values of 5 Ω, 25 Ω and 45 Ω, and incidence angles of 45° and 75° were used.

Based on the methodology used for short-circuit simulations, Table 1 shows the number of simulations performed to form the training and test databases for each of the 11 types of faults simulated.

As shown in Fig. 6, to evaluate the performance of the proposed methodology, based on the results of the allocation of the meters presented in Gomes et al. (2016), three scenarios were considered. The situations of absences before meter 1 were applied to verify the performance of the proposed methodology against different information not used in the learning phase.

Scenario 1 considered 6 m installed at positions related to **M1**, **M2**, **M3**, **M4**, **M5** and **M6**.
Scenario 2 considered 3 m allocated in the positions **M1**, **M3** and **M6**.
Scenario 3 also considered 3 m allocated in the system, but in the positions **M2**, **M4** and **M5**.

In this way, the sensitivity of the framework presented is evaluated considering a variation of the number and position of the meters.

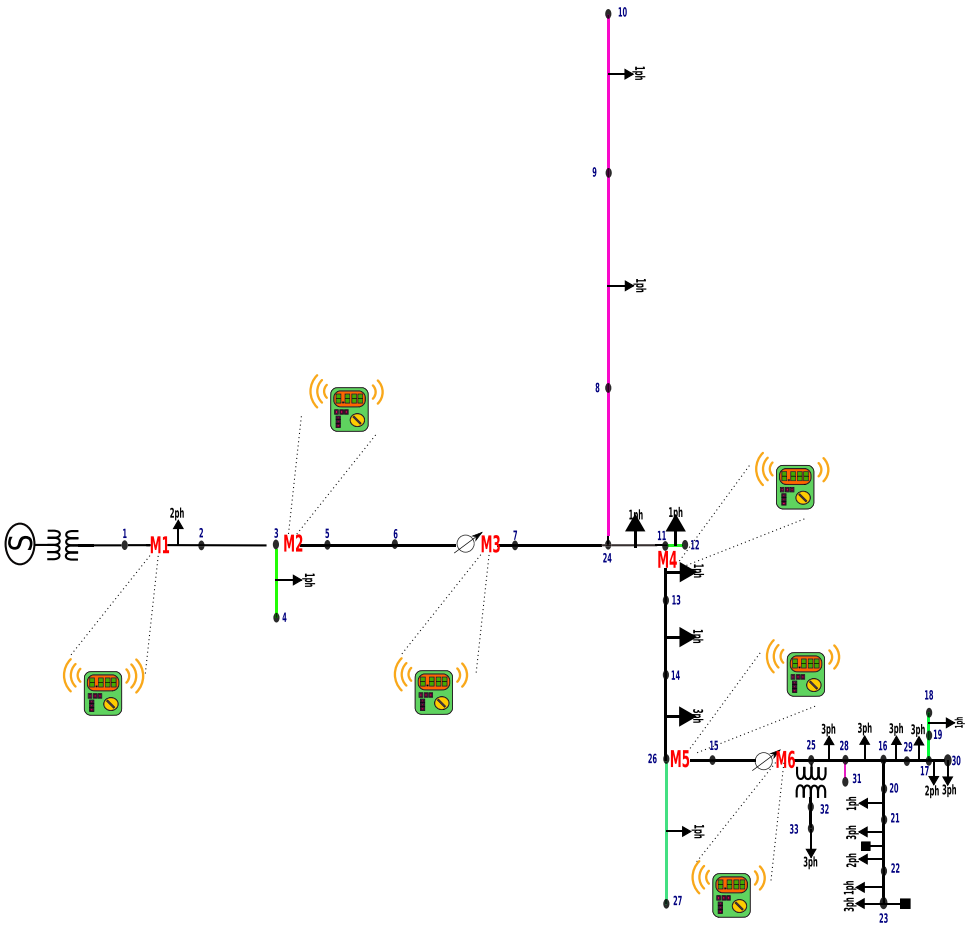

Fig. 6 Indication of the 6 potential positions of the meters used in this research

5.1 ANN Performance

Figure 7 shows the performance of the ANN for the location of the faults in phase A involving ground. In this figure, the estimation by the ANN of the distance of phase *A*-ground faults is represented in red. The real distance of the faults considered is represented in black. Since the data used for the ANNs training were from meter M01, and the faults applied before M01 have an atypical profile in relation to the entire database used in the training process, it is justified that the faults near the substation have a high percentual error.

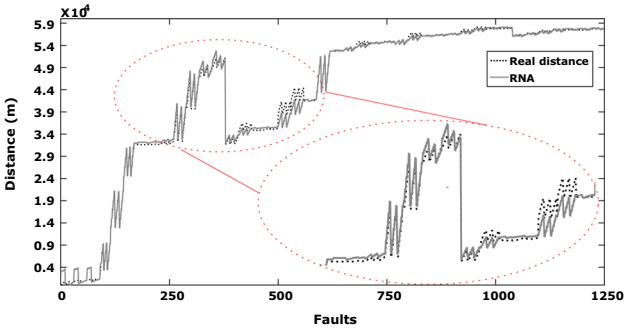

Fig. 7 Comparison of ANN performance in relation to the expected distance

The ANNs were satisfactory for the estimation of the fault distances. Tables 2 and 3 evaluate the performance of the ANN better.

Table 2 evaluates the percentage error (%) for three types of fault (*A*-ground, *AB*-ground and *ABC*-ground faults) by varying the fault resistance, the fault distance and keeping the fault incidence angle fixed at 45° radians.

Table 3 evaluates the percentage error (%) for the same three types of fault, but varying the angle of incidence of the fault, the fault distance and keeping the resistance of the fault fixed at 5 Ω.

The percentage error presented was calculated by:

$$E(\%) = 100 \cdot \left| \frac{D_{\text{real}} - D_{\text{estimated}}}{L} \right|. \tag{2}$$

where D_{real} is the actual fault distance, $D_{\text{estimated}}$ is the distance estimated by the ANNs, and L is the largest length observed in the system from the substation.

According to these tables, although the fault distances were estimated from the substation, no increase in the percentage error is observed when the actual fault distance is

Table 2 Percent error (%) in the estimation of the fault distance, varying the resistance of the fault and keeping the incidence angle of the fault fixed at 45°

Real fault distance	A-ground error		AB-ground		ABC-ground	
	$R_f = 5\,\Omega$	$R_f = 45\,\Omega$	$R_f = 5\,\Omega$	$R_f = 45\,\Omega$	$R_f = 5\,\Omega$	$R_f = 45\,\Omega$
1.05 km (Region 01)	4.68	0.02	0.11	1.23	3.60	0.37
9.31 km (Region 01)	0.84	0.08	0.53	0.71	0.31	0.40
19.14 km (Region 03)	0.28	0.35	0.02	0.32	1.19	0.40
31.74 km (Region 05)	0.56	0.98	1.07	0.94	0.02	0.15
34.84 km (Region 06)	0.82	0.87	1.33	1.25	0.58	1.48
43.48 km (Region 08)	0.40	0.59	1.25	1.30	0.68	0.75
52.7 km (Region 09)	0.01	0.29	0.87	1.10	0.31	0.43
55.98 km (Region 10)	0.16	0.11	0.24	0.39	0.47	0.31
55.99 km (Region 11)	0.16	0.25	0.24	0.53	0.47	0.46

Table 3 Percent error (%) in the estimation of the fault distance, varying the resistance of the fault and keeping the resistance of the fault fixed at 5 Ω

Real fault distance	A-ground		AB-ground		ABC-ground	
	$\phi = 45°$	$\phi = 75°$	$\phi = 45°$	$\phi = 75°$	$\phi = 45°$	$\phi = 75°$
1.05 km (Region 01)	4.68	4.46	0.11	0.59	3.60	4.03
9.31 km (Region 01)	0.84	0.62	0.53	1.16	0.31	0.42
19.14 km (Region 03)	0.28	0.15	0.02	0.35	1.19	1.20
31.74 km (Region 05)	0.56	0.65	1.07	0.89	0.02	0.03
34.84 km (Region 06)	0.82	0.89	1.33	1.20	0.58	0.61
43.48 km (Region 08)	0.40	0.41	1.25	1.19	0.68	0.69
52.7 km (Region 09)	0.01	0.02	0.87	0.84	0.31	0.33
55.98 km (Region 10)	0.16	0.01	0.24	0.08	0.47	0.35
55.99 km (Region 11)	0.16	0.13	0.24	0.22	0.47	0.49

increased, but there is a tendency of percentage error decrease when the distance increase is observed. Moreover, there was no correlation between the percentage error and the fault resistance. This also occurred with the incidence angle of the fault, where no correlation between the variation of this element and the percentage error was evidenced.

A key point in identifying the faulty region is to use the estimation of the distance indicated by the ANN as input of the decision trees that will identify the faulty region.

5.2 Decision Trees Feature Selection

The decision tree favors the identification of parameters that are the most significant for the problem under analysis. The *estimate predictor importance values* (EPIV) of the decision trees used were evaluated in order to estimate, among all the parameters used, which are in fact the most significant for the identification of the faulty area and region. EPIV was calculated using the predictor importance method of the MATLAB Statistics and Machine Learning toolbox. According to the MATLAB documentation, the method estimates the predictor importance by summing changes in the mean squared error (MSE) due to splits on every predictor and dividing the sum by the number of branch nodes. If the tree is grown without surrogate splits, this sum is taken over best splits found at each branch node. If the tree is grown with surrogate splits, this sum is taken over all splits at each branch node including surrogate splits. The predictor importance values have one element for each input predictor in the data used to train this tree. At each node, MSE is estimated as node error weighted by the node probability. Variable importance associated with this split is computed as the difference between MSE for the parent node and the total MSE for the two children.

The following analyses allow the identification of the most significant parameters for the identification of the area and region, considering all types of fault applied. Thus, two types of graphs will be shown. The first graph identifies which are the most significant parameters, and the second one (a histogram) identifies how these parameters are distributed in terms of EPIVs.

In the analyses performed on Figs. 8, 9, 10 and 11, all EPIVs were normalized for better interpretation. In Fig. 8, it is evidenced that most of the parameters used—333 of the 342 (based on a scenario with 3 m present in the system)—are not so relevant for the identification of the area, and that only 3 have an EPIV value above 0.5. These parameters (Fig. 9) are: *RMS* value of the voltage signal of phase *A* of meter 01; *RMS* value of the voltage signal of phase *B* of meter 01; and *RMS* value of the current signal of phase *A* of meter 03.

For the identification of the faulty region (by Fig. 10), it is also observed that only a minority of the considered parameters is really relevant for the decision trees to carry out the classification process. Also, only 1 parameter has EPIV

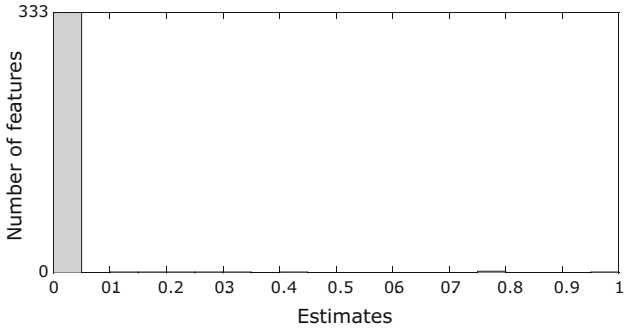

Fig. 8 EPIV histogram of parameters for area identification decision trees

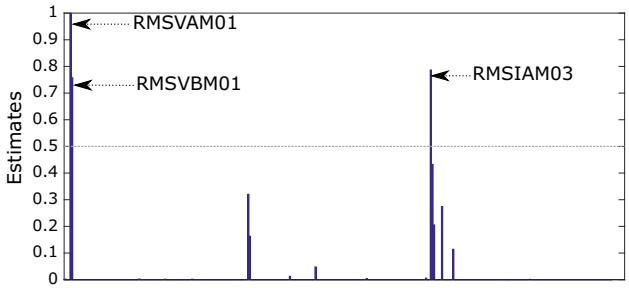

Fig. 9 EPIV of parameters for area identification decision trees

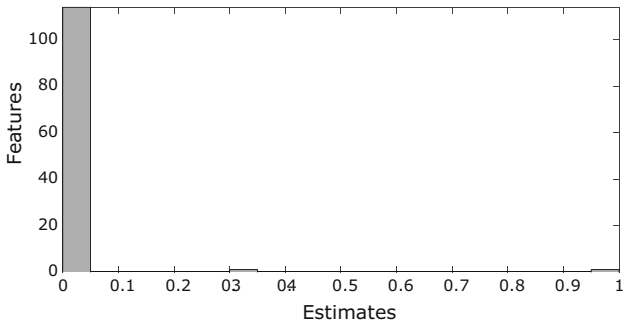

Fig. 10 EPIV histogram of parameters for region identification decision trees

greater than 0.5. In Fig. 11, it is evidenced that the most relevant parameter for the identification of the faulty region is the estimation of the distance of the fault, followed by the area in which the fault occurred. Still in Fig. 11 of the parameter group with low EPIV value, the group formed by the values of *RMS*, amplitude, and phase angle of the three-phase voltage and current signals stands out when compared to the parameters formed by the energy of the WPT coefficients.

5.3 The Performance of the Decision Trees

Tables 4 and 5 present, respectively, the results for the estimation of the area and region of faults occurrence for each of the 3 scenarios evaluated.

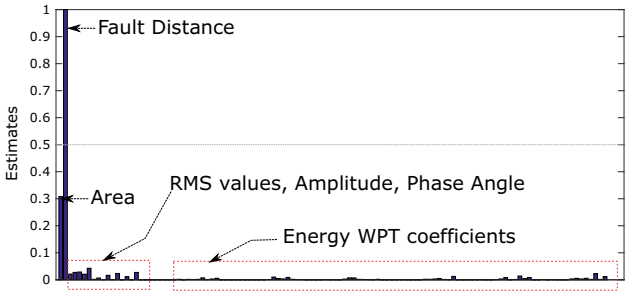

Fig. 11 EPIV of parameters for region identification decision trees

Table 4 Hit rate (%) for the identification of the fault occurrence area for each of the 11 types of faults evaluated

Fault type	Scenario 01	Scenario 02	Scenario 03
A-ground	99.52	100	99.52
B-ground	98.28	100	98.12
C-ground	98.17	100	98.5
AB	99.42	100	99.7
BC	99.71	100	99.7
AC	99.42	100	99.4
AB-ground	98.46	100	98.1
BC-ground	98.94	100	99.2
AC-ground	98.65	100	99.4
ABC	99.42	100	100
ABC-ground	98.65	100	98.9

Table 5 Hit rate (%) for the identification of the fault occurrence region for each of the 11 types of faults evaluated

Fault type	Scenario 01	Scenario 02	Scenario 03
A-ground	91.1	89.0	92.5
B-ground	87.7	80.2	82.4
C-ground	88.6	82.6	79.4
AB	95.1	98.6	93.6
BC	96.0	97.7	95.7
AC	97.4	96.8	96.8
AB-ground	86.3	84.7	87.8
BC-ground	91.0	88.9	89.8
AC-ground	94.0	93.5	94.5
ABC	94.8	94.2	94.5
ABC-ground	90.4	88.8	93.4

For the identification of the outage area, it is evident that, regardless of the number of meters or their position, the hit rate varied around 99%. This proved that increasing the number of meters not necessarily results in a significant gain of result.

Regarding the general performance for the identification of the outage region, the best result was obtained for scenario 01 (highest number of meters). However, the difference between scenarios 2 and 3 was not very significant (with exception to phase *C*-ground faults). Thus, there was little sensitivity of the algorithm for a variation of the number of meters and their positioning.

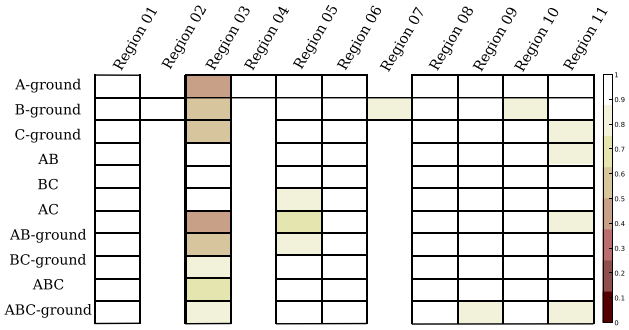

Fig. 12 Overall performance for the identification of the outage region considering scenario 1

Figure 12 shows a global view on the performance of all types of fault. This figure shows an indication of how well each type of fault was identified for each evaluated region. The lighter colors in the figure indicate a higher hit rate, and regions where a specific type of fault cannot occur are empty. For example, only the *B*-ground faults can be considered in regions 2 and 7. And in region 4, only the *A*-ground faults can be considered.

Thus, for the meter allocation considered, region 3 demonstrated greater difficulty to locate the faults through the application of decision trees.

The confusion matrix for the location of *A*-ground faults in scenario 1 is presented in Fig. 13. Among the 114 fault cases that actually occurred in region 3, 60 of them were improperly indicated as if occurring in region 5. This confusion matrix shows that in all cases with a location error, they were pointed out as if occurring in a region neighboring the region in which the fault actually occurred.

When scenarios 2 and 3 (Figs. 14, 15), in which the number of meters was reduced from 6 to 3, were evaluated, region 3 continued to have the greatest difficulty in locating faults, and region 6 had difficulties in locating the respective faults.

It is observed that even in these two scenarios, where the meters were installed in different positions, the algorithm performance changed very little. This suggests a tolerance of this fault location approach based on machine learning algorithms for the variations in the distribution of the meters.

It is also clear that it is necessary to have an optimized meter allocation that is focused on fault location.

Another highlight is the possibility of allocating meters in the system based on the return of this framework to the location of faults.

Region of reference \ Estimated Region	1	3	4	5	6	8	9	10	11	
1	**114** 9.1%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	100% 0.0%
3	**0** 0.0%	**54** 4.3%	**0** 0.0%	**60** 4.8%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	47.4% 52.6%
4	**0** 0.0%	**0** 0.0%	**137** 11.0%	**13** 1.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	91.3% 8.7%
5	**0** 0.0%	**0** 0.0%	**2** 0.2%	**55** 4.4%	**3** 0.2%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	91.7% 8.3%
6	**0** 0.0%	**0** 0.0%	**0** 0.0%	**3** 0.2%	**147** 11.8%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	98.0% 2.0%
8	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**30** 2.4%	**0** 0.0%	**0** 0.0%	**0** 0.0%	100% 0.0%
9	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**197** 15.8%	**4** 0.3%	**9** 0.7%	93.8% 6.2%
10	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**198** 15.9%	**12** 1.0%	94.3% 5.7%
11	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**0** 0.0%	**4** 0.3%	**1** 0.1%	**205** 16.4%	97.6% 2.4%
	100% 0.0%	100% 0.0%	98.6% 1.4%	42.0% 58.0%	98.0% 2.0%	100% 0.0%	98.0% 2.0%	97.5% 2.5%	90.7% 9.3%	**91.1%** **8.9%**

Fig. 13 Confusion matrix concerning the location of phase *A*-ground faults in scenario 1

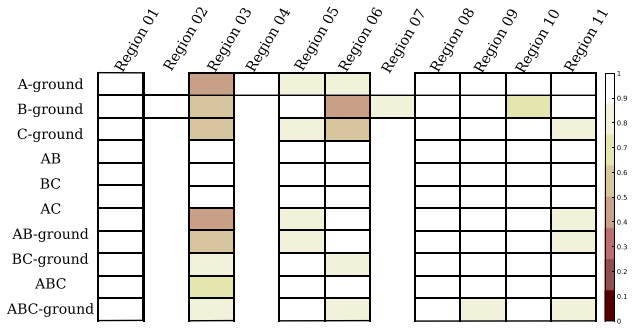

Fig. 14 Overall performance for the identification of the outage region considering scenario 2

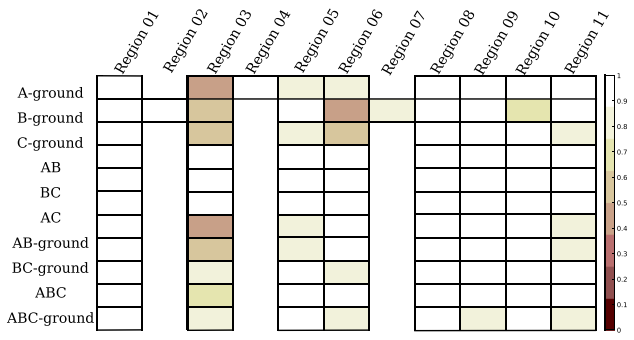

Fig. 15 Overall performance for the identification of the outage region considering scenario 3

6 Conclusions

This paper presented a framework based on machine learning for fault location in radial distribution systems that meets the needs and trends of intelligent electrical networks.

In this approach, ANNs were used to estimate the distance of the fault from the power substation, and decision trees were used to identify the area (larger delimited portion) and the region (better delimited portion) of the system where the fault occurred.

Percentage errors between 0.01 and 4.68% were observed in the evaluated situations for the estimation of the fault distance by ANNs.

The problem of multiple estimation of the fault was also addressed in the research. For the location of the faults, a multistage approach was considered in which some inferences obtained by the algorithm itself were used in its search process of the fault location. In order to identify the most specific location (region) of the fault occurrence, a previous identification of a more general region of the fault occurrence was used. The more general identification of the faults (area) presented a hit rate above 99%. This was a relevant information to aid the algorithm in identifying the region of fault occurrence.

A methodology for the segmentation of the system was presented for the identification of the outage region of the system. This segmentation considered the ability of the ANNs to understand the nature of faults. Therefore, small and close regions were added. Hit rates ranging from 79 to 98.6% for the identification of the outage region were observed. It is worth noting that in cases with an error in the location, the fault was pointed out as if occurring in neighboring regions. Hence, the identification of the region of the fault by the proposed methodology is a good indication for the utility to find the outage region in the system even when there is an error.

A study was also presented evaluating the impact of the variation of the resistance and the incidence angle of the fault, of the number of meters and their position in the system. In view of the situations and scenarios evaluated, it is concluded that this framework was not very sensitive to the evaluated aspects.

It is worth mentioning that the methodology presented has not yet been validated in relation to other radial distribution systems, as well as in the presence of distributed generation. However, the knowledge obtained about the test system used in this research allows to assume that the methodology, as presented, can be applied in new topologies, lacking a new set of computational simulations to generate a representative database. The studies will be in this direction.

Acknowledgements The authors acknowledge the Department of Electrical and Computing Engineering from São Carlos School of Engineering and the University of São Paulo (Brazil) for the research facilities provided to conduct this Project. Our thanks also extend to the financial support received from CAPES and CNPq (Governmental Brazilian Institutions).

References

Adewole, A. C., Tzoneva, R., & Behardien, S. (2016). Distribution network fault section identification and fault location using wavelet entropy and neural networks. *Applied Soft Computing*, *46*(2016), 296–306.

Bahmanyar, A., Jamali, S., Estebsari, A., & Bompard, E. (2017). A comparison framework for distribution system outage and fault location methods. *Electric Power Systems Research*, *145*(2017), 19–34.

da Silva Pessoa, A. L., Oleskovicz, M., & Martins, P. E. T. (2018). A multi-stage methodology for fault location in radial distribution systems. In *2018 18th international conference on harmonics and quality of power (ICHQP)* (pp. 1–6). New York: IEEE.

Dehghani, F., & Nezami, H. (2013). A new fault location technique on radial distribution systems using artificial neural network. In *22nd international conference and exhibition on electricity distribution (CIRED 2013)* (pp. 1–4). New York: IET.

Farias, P. E., de Morais, A. P., Junior, G. C., & Rossini, J. P. (2016). Fault location in distribution systems: A method considering the parameter estimation using a RNA online. *IEEE Latin America Transactions*, *14*(12), 4741–4749.

Goel, N., & Agarwal, M. (2015). Smart grid networks: A state of the art review. In *2015 international conference on signal processing and communication (ICSC)* (pp. 122–126). New York: IEEE.

Gomes, D. P. S., Oleskovicz, M., Kempner, T. R., & Filho, J. R. L. (2016). A generalized coverage matrix method for power quality monitor allocation utilizing genetic algorithm. In *International conference on renewable energy and power quality* (No. 14).

James, G., Witten, D., & Hastie, T.,& Tibshirani, R., (2013). *An introduction to statistical learning* (Vol. 6, p. 2013). Berlin: Springer.

Lout, K., & Aggarwal, R. K. (2013). Current transients based phase selection and fault location in active distribution networks with spurs using artificial intelligence. In *2013 IEEE power and energy society general meeting* (pp. 1–5.). New York: IEEE.

Lovisolo, L., Neto, J. M., Figueiredo, K., De Laporte, L. M., & Roch, J. D. S. (2012). Location of faults generating short-duration voltage variations in distribution systems regions from records captured at one point and decomposed into damped sinusoids. *IET Generation, Transmission and Distribution*, *6*(12), 1225–1234.

Mitchell, T. M., et al. (1997). *Machine learning*. New York: WCB.

Pérez, R., & Vásquez, C. (2016). Fault location in distribution systems with distributed generation using support vector machines and smart meters. In *Ecuador technical chapters meeting (ETCM), IEEE* (Vol. 1, pp. 1–6). New York: IEEE.

Rafinia, A., & Moshtagh, J. (2014). A new approach to fault location in three-phase underground distribution system using combination of wavelet analysis with ANN and FLS. *International Journal of Electrical Power and Energy Systems*, *55*(2014), 261–274.

Ray, P., Mishra, D. P., & Panda, D. D. (2015). Hybrid technique for fault location of a distribution line. In *2015 annual IEEE India conference (INDICON)* (pp. 1–6). New York: IEEE.

Saha, M. M., Izykowski, J. J., & Rosolowski, E. (2009). *Fault location on power networks*. Berlin: Springer.

Zapata-Tapasco, A., Mora-Flórez, J., & de Almeida, M. C. (2014). Fault location in power distribution systems using a learning approach based on decision trees. In *Transmission and distribution conference and exposition-Latin America (PES T&D-LA), 2014 IEEE PES* (pp. 1–6). New York: IEEE.

Zayandehroodi, H., Mohamed, A., Farhoodnea, M., & Heidari, A. (2013). New training strategies for RBF neural networks to determine fault location in a distribution network with DG units. In *2013 IEEE 7th international on power engineering and optimization conference (PEOCO)* (pp. 450–454). New York: IEEE.