Fault Classification in Power Distribution Systems using PMU Data and Machine Learning

Flávio Lori Grando, André Eugenio Lazzaretti, Miguel Moreto, and Heitor Silvério Lopes
Technological University of Paraná (UTFPR) – Curitiba – PR – Brazil
flaviogrando@alunos.utfpr.edu.br, lazzaretti@utfpr.edu.br, miguel.moreto@ufsc.br, hslopes@utfpr.edu.br

***Abstract*—This work presents the analysis of machine learning methods for fault (short-circuit) classification in electrical distribution networks using data from PMUs (Phasor Measurement Units) installed along the network. The Alternative Transient Program (ATP) was used to simulate 26,928 different instances distributed into 33 types of faults – single and multi-phase, including or not the ground and different wire breakages – and one normal condition of the system. The IEEE 123-bus distribution system was used as the test system. We compared five machine learning methods for classification: Linear Discriminant Analysis (LDA), k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Decision Trees (DTs). The best result was achieved by the SVM with Gaussian kernel and ANN. The input data (feature extraction) was also varied, testing data from one or several PMUs, ABC sequence phasors and symmetrical sequence phasors. We obtained slightly better results for symmetrical components and multiple PMUs in the network. Finally, classes of the same short-circuit with different wire breakages were grouped, raising the overall classification accuracy. Overall conclusion is that the proposed approach is feasible for fault classification using PMU-data in a distribution network.**

Keywords*—*Distribution Systems, Fault Classification, Machine Learning, Phasors, PMU.

I. Introduction

The synchronized phasor measurement technology has opened a new paradigm in the observability of electrical systems, allowing to trace in real-time the dynamics of the system through synchronized data with high precision and resolution [1]. The synchronization of the measurements is obtained through Global Positioning System (GPS). Thus, Phasor Measurement Units (PMUs) extract measurements of synchronous phasors (synchrophasors) and frequency of sinusoidal signals at different points along an electric power system. Then, the information is sent out to a Phasor Data Center (PDC), as shown in Fig. 1.

Conventional synchronized phasor measurement systems are based on PMUs that extract measurements of potential and current transformers at power substations. Because they are based on measurements at substations, they are unable to record local dynamics along the distribution system [2]. On the other hand, previous works suggest several applications for synchrophasorial measurement at distribution level [3], [4], [5]. Due to the observability that the PMU can provide and the large amount of information generated, the synchrophasor technology has been largely applied with machine learning-based approaches to detect, classify, and locate events [6], [7], [8].

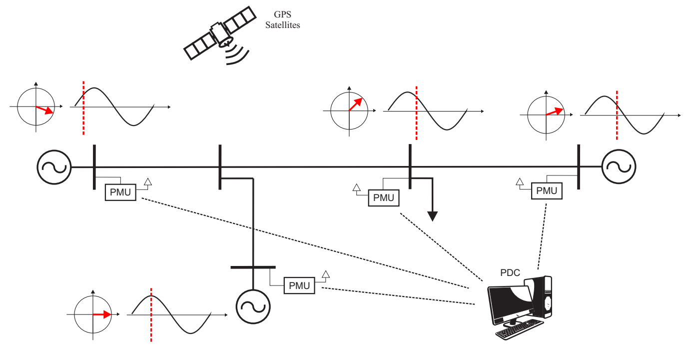

Figure 1. Representation of a Synchronized Phasor Measurement System.

Several common events of a distribution system can be observed by PMUs, such as the load and equipment switching-transients, as well as short-circuits (faults) [9]. These two group of events are the major interest, since they usually result in blackouts and high costs for power utilities. In this sense, several works have been conducted, mainly motivated by the expansion of the system and the growth of smart grids [10], [11].

In [12], for instance, a PMU data-driven framework was proposed to distinguish a malfunctioned capacitor bank switching and a malfunctioned regulator on-load tap changer switching from two normal operating events, using the IEEE 123-bus. For different noise levels and number of PMUs, the authors showed the feasibility of using PMU data to satisfactory classify those events.

A similar data-driven approach was proposed in [13]. Authors classified power quality events using real-data collected during 15 days from two micro-PMUs installed on a real distribution feeder. The power quality events include the detection of internal phase imbalance in a 900 kVAR capacitor bank as well as a potential malfunction in its Volt/VAR controller. On the other hand, [14] presented the application of a micro-phasor measurement unit for power distribution network monitoring. Particularly, the authors discuss the detection of abnormal events, that is, transients in voltage and current waveforms that may be caused by faults, topology changes, load behavior, and source dynamics, without however, discriminate among types of faults.

However, in the particular case of faults along the distribution system, a detailed analysis of different types of faults, taking into account different types of wire breakage, is still underexplored in the literature, especially using PMU-data. Therefore, this work aims at applying machine learning methods for fault classification in electric power distribution networks, using data from different PMUs installed along the system. To do that, the IEEE 123-bus modeled in the ATP was used. We compared different classification methods (kNN, LDA, SVM, ANN, and DT), different PMU locations, and different feature extraction procedures. Additionally, 33 types of faults were considered and 3 types of wire breakage (load, source and normal) are simulated, resulting in a detailed analysis of fault situations in power distribution systems.

This paper is structured as follows. In Section II, the proposed method is detailed, including the simulated system and cases, feature extraction, and classification procedure. Section III discusses the experimental setup and results obtained. Finally, conclusions are drawn and future work is outlined in Section IV.

II. Proposed Method

The study was performed in a simulation environment, where the distribution network was simulated in the Alternative Transient Program (ATP). The ATP output data corresponds to voltage and current waveforms (oscillographs), which were imported into the MATLAB software for processing and application of classification methods. In MATLAB, a phasor estimation algorithm based on the Discrete Fourier Transform performs the PMU function, providing phasors according to the IEEE C7.118.1-2011 standard [15]. In the following subsections, each stage of the proposed method is presented (summarized in Fig. 2).

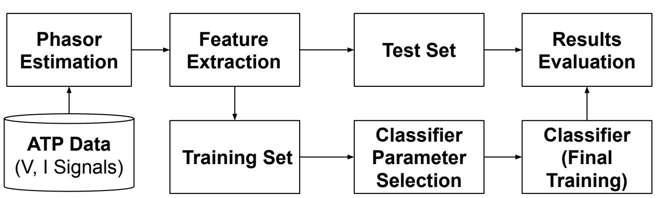

Figure 2. Overview of the proposed method.

A. Simulation of the Distribution System

The simulated system corresponds to a publicly available IEEE 123-bus feeder created by the IEEE Power and Energy Society for test purposes [16]. This system operates at a nominal voltage of 4.16 kV and it is characterized by overhead and underground lines; unbalanced loads with constant current, impedance and power; voltage regulators; capacitor banks; and multiple switches. A physical representation of this circuit is shown in Fig. 3, along with the fault locations and the selected PMU monitoring points. A total of 8 fault locations were distributed along the main branches of the circuit. The PMUs were located at the points where there are normally closed switches, which are considered strategic points from the operation point of view, as suggested in [12].

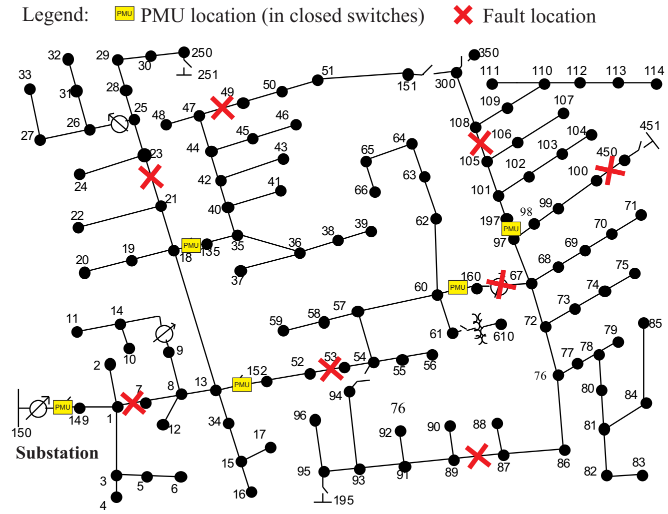

Figure 3. IEEE 123-bus test system.

In this study, some simplifications were considered in the original IEEE system, such as:

- Mutual inductances were excluded;
- Voltage regulators were removed;
- The connection of the loads was considered of Y-type;
- All loads were simplified to constant impedance.

Such simplifications do not have impact on the type of analysis proposed in this work – particularly for PMU data, as discussed in [17]. For simulation purposes, voltage and current waveforms of normally closed switches are exported and processed in MATLAB using the following procedure:

- Insertion of Additive White Gaussian Noise (AWGN) with 40 dB;
- Decimation procedure, reducing from 20,000 points to 500 points per cycle (simulating sampling);
- Saturation of the signal, simulating the full-scale of the analog-to-digital converter;
- Estimation of the phasors by applying the Discrete Fourier Transform.

The phasor estimation complies with the IEEE C37.118.1 standard format. This means that one phasor is calculated for each signal cycle in the fundamental frequency component, that is, 60 phasors per second. Thus, two databases were generated, one containing 98 GB of oscillographs and the other containing 630 MB of phasors registers. More details of the data are described in the next subsection, whilst the simulated cases are described later.

B. Feature Extraction

Since electromagnetic transients present short duration in time (less than one cycle), the characteristic transient signature is not recorded by PMU data. Therefore, we decided to extend the observation analysis including one cycle before and one cycle after the transient event, as shown in Fig. 4.

Thus, two sets of attributes were created. The first set was obtained by the difference between the pre and post-transient

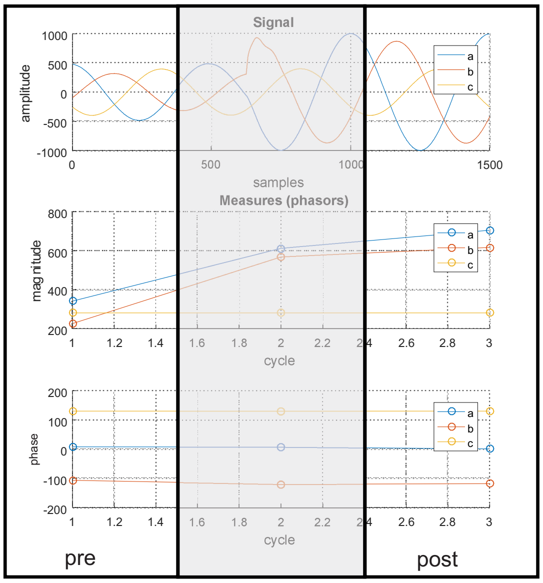

Figure 4. Waveform of the phasor data.

values, for each parameter (magnitude and angle) and each phase (A, B and C). The second set was composed of the difference of the parameters between each phase in the post-transient moment. That is, an analysis of the system imbalance during the fault. These sets result 12 attributes, 6 attributes per signal (voltage and current).

Additionally, the symmetrical components for the phasor set of sequence ABC were calculated, generating a new group of phasor registers. Using the zero, positive, and negative sequence components (sequence 012) for phase A (reference phase), the same feature extraction procedure was applied, that is, system state variation and unbalance between phasors during the fault. With that, the number of features resulted in 24 attributes per PMU. Finally, the four groups of data were organized as follows:

1) PMU-1 (subestation), data from sequence ABC (24 features);
2) All PMUs, data from sequence ABC (120 features);
3) PMU-1 (subestation), data from sequence 012 (24 features);
4) All PMUs, data from sequence 012 (120 features);

Therefore, we decided to present a comparison between the data from one PMU, located at the feeder output (substation), and additional PMUs installed along the power grid. Additionally, we compare ABC sequence data and symmetrical sequence (012).

C. Simulated Cases

This work considers conventional types of single-phase, two-phase and three-phase faults, with short-circuits including individual phases and ground. A distinction was also done for phases (A, B or C) in each fault condition. Three-phase and three-phase-to-ground faults are separated into two classes. In addition, fault types can be combined with the cable break condition, which may be determinant for power flow direction when the line is in fault conditions. In this sense, there are three possibilities, summarized in Fig. 5. Notice that, due to three-phase transformer connections, the *load* condition can also be observed in distribution systems.

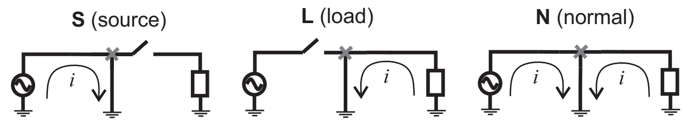

Figure 5. Representing diagram w.r.t. cable break condition.

Simulations were performed with various conditions of the distribution system, according to the combination presented in Table I. The first situation is the load profile, represented as a percentage of typical values, where 100% corresponds to 3.5 MW and 1.9 MVAr. The second situation corresponds to the fault impedance, expressed in Ohms. The third is the fault location, referenced by the line (inter-node) of the circuit in Fig. 3. The fourth situation refers to the direction of the power flow at the fault instant. Finally, the last situation corresponds to the fault type. The letters indicate the phases (A, B and C) and the ground (G). Thus, a single-phase-ground fault in phase A, for instance, is represented by '*AG*'. A two-phase fault, in phases B and C, is represented by '*BC*', and a three-phase-to-ground fault is represented by '*ABCG*'.

Table I. SIMULATED CASES AND CORRESPONDING PARAMETERS.

Simulation	**Load Condition (%)**	**Fault Resistance (Ohm)**	**Fault Location** (line)	**Power Flow direction**	**Fault Type**
	10	1	1-7	N (normal)	AG
	20	5	21-23	S (source)	BG
	30	10	47-49	L (load)	CG
	40	20	52-53	–	AB
	50	30	160-67	–	BC
Conditions	60	40	87-89 –	–	CA
	70	50	100-450	–	ABG
	80	100	105-108	–	BCG
	90	200	–	–	CAG
	100	–	–	–	ABC
	110	–	–	–	ABCG
Quantity	11	9	8	3	11

The combination of test conditions resulted in 26,928 cases, including an additional class defined as normal (without any fault). The classes were categorized by the combination of fault types with the power flow direction and the normal operating class of the system, totaling 34 classes.

D. Machine Learning Classifiers

For the evaluation of the four test groups, different classification methods were used, detailed as follows:

- K-Nearest-Neighbors: in this approach, the feature vector classification is performed according to the previous classified feature vectors, associating it to the one which presents the most similar characteristics (the closest in terms of the Euclidean distance) [18]. Since the classification is simply based on distances related to a training set, this method may be considered one of the simplest machine learning algorithms, easily implemented. The main disadvantage is that it requires the storage of all the training data, which may cause problems to embed such classifier.

- Decision Trees: possibly, DTs are the most common method used for fault classification found in the literature [18]. In a (binary) DT, different binary classifications are performed, considering different input features. These classifications are concatenated in a tree structure, in which each node concerns the test of a variable according to its possible range of values. In the end, a combination of different evaluations is performed in order to obtain the final class label.
- Support Vector Machines: they were originally developed to solve classification problems using 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 the Vapnik-Chervonenkis (VC) dimension [19], whose upper limit is inversely proportional to ρ. Additionally, it is possible to include a nonlinear input mapping, by replacing the dot product in the original formulation by the kernel product in feature space. There are several types of kernel, which must abide to the conditions of Mercer's theorem – in this work, we use the Gaussian kernel.
- Artificial Neural Network: consists of a parallel distributed signal processor, made up of simple processing units (neurons), capable of storing knowledge (in the synaptic weights) using a learning algorithm, making their knowledge available for future use. In this work, we used the Multi-Layer Perceptron (MLP) network, which is a feedforward architecture, with just a single hidden layer, an input layer and the output layer that corresponds to the assigned class (label). The learning process is based on the backpropagation algorithm, which basically consists of a method that estimates the gradient of the training error cost function along the layers of the network. Finally, a gradient descent-based method is used to optimize and estimate the parameters – synaptic weights.
- Linear Discriminant Analysis: this method reduces high-dimensional data to a lower dimensional space, maximizing the separation between classes. This is done in order to reduce its complexity and the required computational effort, as well as to avoid the possibility of overfitting.

The algorithms were implemented in MATLAB, using the classifiers provided by the software through the Statistics and Machine Learning Toolbox.

III. Results and Discussion

The accuracy of the methods was calculated as the average of the test sets using a 5-fold cross-validation procedure. The results are presented in Table II. In the case of ANNs, the second column represents the number of neurons of the hidden layer, or more than one layer separated by hyphen.

Table II. Accuracy considering different methods and test groups.

Measurement		**PMU-1**		**All PMUs**	
Num. of Feats		**24**	**24**	**120**	**120**
Classifier	**Type**	**Seq. ABC**	**Seq. 012**	**Seq. ABC**	**Seq. 012**
LDA	-	0.28	0.26	0.35	0.36
k-NN	Eucl. Dist.	0.31	0.36	0.34	0.40
DT	All comb.	0.38	0.37	0.38	0.42
SVM	Gauss. Kernel	0.46	**0.52**	0.43	**0.55**
RNA	35	0.37	0.36	0.41	0.37
RNA	70	0.40	0.37	0.44	0.44
RNA	105	0.39	0.38	0.45	0.45
RNA	140	0.39	0.39	0.45	0.45
RNA	35-35	0.44	0.45	0.46	0.44
RNA	35-35-35	0.47	0.45	0.45	0.46
RNA	70-70	0.45	0.50	0.45	**0.54**
RNA	70-70-70	0.50	**0.54**	**0.52**	0.50

The best classifiers were: the ANNs with more than one hidden layer and 70 neurons, and the SVM with Gaussian kernel. It is worth noting that the apparently low accuracy is justified by the presence of 34 classes. Neural networks with only one hidden layer did not perform well even with an increased number of neurons in the hidden layer. Additionally, simpler methods such as DTs and kNN presented comparable performance.

Concerning the test groups, it can be observed that the use of more PMUs improves the overall classification performance. However, the precision gain is still low when compared to the increased complexity of the classifier, which has 4x more attributes. In this case, there is, probably, a redundancy in the information used that may even impair the performance when used in excess. One solution to this problem would be the use algorithms for optimal allocation of PMUs in the network in order to maximize the performance of the classifier.

The group of attributes based on symmetrical components was slightly better than the group of ABC components. In the worst case, the accuracy was similar, while in the best results the accuracy was significantly higher.

Analyzing the classes, it is possible to identify that the SVM classifier, despite presenting an equivalent global accuracy of the multilayer neural network, had accumulated errors in class AG, confusing it with all classes. Therefore, a result was selected with the RNA 70-70 classifier to present the confusion matrix in Fig. 6. A color scale was used to facilitate the analysis, where the highest values are in red, mean values in yellow and lowest values (0) in green. The class **X** at the end of the confusion matrix represents the normal operating class of the system. On the right side there is a column with the accuracy for each class, with the color scale of the worst result in red and best result in green.

An analysis of the matrix leads to the following conclusions:

- A hit tendency for each class, observed by the quantities of the main diagonal;
- Observing the diagonals of groups N, S and L, there is a tendency to obtain the correct classification for the types of fault (AG, AB, etc.) with some confusion between these groups, especially between N and S;

CLASSIFICATIONS

		N (normal)											S (source)											L (load)											X	Accuracy		
TARGETS	Type	AG	BG	CG	AB	BC	CA	ABG	BCG	CAG	ABC	ABCG	AG	BG	CG	AB	BC	CA	ABG	BCG	CAG	ABC	ABCG	AG	BG	CG	AB	BC	CA	ABG	BCG	CAG	ABC	ABCG	X			
N (normal)	AG	134	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.94	AG	N
N (normal)	BG	0	157	0	0	0	0	1	0	0	0	0	0	2	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.97	BG	N
N (normal)	CG	0	0	149	0	0	0	0	0	2	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.93	CG	N
N (normal)	AB	0	0	0	151	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.99	AB	N
N (normal)	BC	0	0	0	0	163	0	0	0	0	0	0	0	0	0	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.97	BC	N
N (normal)	CA	0	0	0	0	0	152	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.93	CA	N
N (normal)	ABG	0	0	0	0	0	0	145	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	ABG	N
N (normal)	BCG	0	0	1	0	0	0	0	151	0	4	0	0	0	0	0	0	0	0	13	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0.87	BCG	N
N (normal)	CAG	0	0	0	0	0	0	0	0	157	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	CAG	N
N (normal)	ABC	0	0	0	0	0	0	0	0	0	72	61	0	0	0	0	0	0	0	0	0	8	5	0	0	0	0	0	0	0	0	0	0	0	1	0.49	ABC	N
N (normal)	ABCG	0	0	0	0	0	0	0	0	0	73	78	0	0	0	0	0	0	0	0	0	7	8	0	0	0	0	0	0	0	0	0	0	0	0	0.47	ABCG	N
S (source)	AG	84	0	0	0	0	0	0	0	0	0	0	56	0	0	0	0	1	1	0	0	0	0	3	0	1	0	0	0	0	1	0	0	0	2	0.38	AG	S
S (source)	BG	0	102	0	0	0	0	3	0	0	0	0	0	32	0	10	0	0	0	1	0	1	0	0	7	0	1	0	2	0	0	1	0	0	0	0.20	BG	S
S (source)	CG	0	0	77	0	0	0	0	0	1	0	0	0	0	63	0	2	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	2	0.42	CG	S
S (source)	AB	0	0	0	27	0	0	0	0	0	0	0	0	0	0	132	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0.81	AB	S
S (source)	BC	0	0	0	0	96	0	0	0	0	0	0	0	0	4	0	20	1	0	31	0	0	0	0	0	1	0	2	1	0	1	0	0	0	0	0.13	BC	S
S (source)	CA	0	0	0	0	0	44	0	0	0	0	0	0	0	14	0	0	99	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	0	0	0.61	CA	S
S (source)	ABG	0	2	0	1	0	0	95	0	0	1	1	0	0	0	24	0	0	26	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0.17	ABG	S
S (source)	BCG	0	0	0	0	0	0	0	92	0	1	0	0	0	0	0	2	0	0	56	0	3	2	0	0	0	0	4	0	0	0	2	0	0	1	0.34	BCG	S
S (source)	CAG	0	0	1	0	0	0	0	0	108	0	0	0	0	17	0	0	8	0	0	39	1	0	0	0	0	0	0	2	0	0	1	0	0	4	0.22	CAG	S
S (source)	ABC	0	0	0	0	0	0	0	0	0	22	4	0	0	0	0	0	0	0	18	0	88	19	0	0	0	0	0	0	0	2	0	4	0	2	0.55	ABC	S
S (source)	ABCG	1	0	0	0	0	0	0	0	0	17	10	0	0	0	0	0	0	1	23	0	83	29	0	0	0	0	0	0	0	0	0	0	0	1	0.18	ABCG	S
L (load)	AG	20	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	91	0	0	10	0	0	0	0	0	0	8	27	0.13	AG	L
L (load)	BG	0	5	0	2	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	63	0	27	0	1	1	0	0	0	0	57	0.40	BG	L
L (load)	CG	0	0	3	0	1	0	0	0	0	0	0	0	0	9	0	2	0	0	0	0	0	0	0	0	83	0	0	6	0	0	1	0	0	49	0.54	CG	L
L (load)	AB	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	19	1	0	71	0	3	23	0	0	0	7	13	0.50	AB	L
L (load)	BC	0	0	0	0	6	0	0	0	0	0	0	1	0	0	0	6	0	0	4	0	0	0	0	0	13	0	34	4	0	42	0	0	3	35	0.23	BC	L
L (load)	CA	0	0	0	0	1	1	0	0	0	0	0	0	0	3	0	1	5	0	0	0	0	0	3	0	9	9	2	44	0	0	30	0	14	19	0.31	CA	L
L (load)	ABG	0	0	0	3	1	0	4	0	0	0	0	0	0	0	3	0	0	1	0	0	1	0	16	1	0	51	0	1	44	0	0	0	8	22	0.28	ABG	L
L (load)	BCG	0	0	0	0	1	0	0	4	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	19	0	53	0	0	48	0	0	2	45	0.27	BCG	L
L (load)	CAG	0	0	0	0	0	0	0	0	8	0	0	3	0	3	0	0	0	0	0	0	0	0	0	0	12	7	3	22	0	0	57	0	18	17	0.38	CAG	L
L (load)	ABC	0	0	0	0	0	0	1	0	0	0	0	1	0	0	4	1	0	0	9	0	9	3	5	0	0	16	2	3	0	5	9	26	55	24	0.15	ABC	L
L (load)	ABCG	0	0	0	2	0	0	1	0	0	3	0	0	0	0	4	0	0	1	2	0	7	2	3	0	1	5	1	4	0	3	9	7	80	23	0.51	ABCG	L
X	X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	161	1.00	X	X
																																				0.55	Total	

Figure 6. Confusion matrix with the accuracy for each fault class.

- The group with the smallest error among fault types is N, with average performance for S, and the worst for L;
- In addition to the internal error (fault types) of the L group, there is a great deal of confusion with class X, the normal operating class of the system;
- Confusion between three-phase (ABC) and three-phase earth fault (ABCG), as expected;
- Overall performance is negatively affected mainly by the confusion between groups N and S, and between group L and class X.

Taking into account that the simulated system is radial, with no distributed generation and Y-grounded loads, the power flow is unidirectional, from the substation to the load, even in fault conditions. Therefore, the L group is best characterized as line loss (or load loss), where there is no fault current. Under this same analysis, the confusion between groups N and S is justified, because the direction of the power flow is the same for both cases. The difference between the two is that in relation to the S group, there is a loss of line simultaneously with the fault.

The confusion between group L and class X occurs mainly by two aspects. First, remote faults turn off a small amount of loads from the system, causing insensitive variations. Second, with the network operating at low load (for example, 10%), the loss of a portion of the system also generates less noticeable variations.

With the analysis described so far and comparing the performance and complexity of the classifiers, it is convenient to make some groupings between the classes:

- New class 'Faults', grouping the classes S and N;
- New class 'Line Loss', grouping the types of faults of the group L;
- New class 'ABC' grouping both types of three-phase faults.

Thus, the new confusion matrix is shown in Figure 7. In this case, there is an accuracy improvement from 55% to 73%, now with 19 classes. The performance of the individual groups was 100% for the normal operating class, 92% for the fault group, and 51% for line loss. Therefore, the latter is mainly responsible for the error rate. It is important to highlight that this grouping was performed based on the previous result, that is, it was not necessary a new training and validation process.

It is important to point out that extreme cases were simulated, especially those with high impedance, such as 50, 100 and 200 Ohms. These cases are difficult to detect due to the small variation of the resulting current, being likely the factor of greatest responsibility for the misclassification of the fault group.

IV. Conclusions and Future Work

This work presented a machine learning-based method to classify faults in distribution systems. The IEEE 123-bus distribution system was simulated in the ATP software in order to generate approximately 27,000 cases considering different load conditions, fault impedances, fault locations, and fault types. The waveforms generated in the ATP software were imported into the MATLAB software, where noise insertion, decimation and saturation processes were applied, generating an oscillographic and a phasor dataset. From the phasor data, different classifiers were applied to discriminate among faults,

		CLASSIFICATIONS																			
		Fault										Line loss							Normal	Accuracy	
	Type	AG	BG	CG	AB	BC	CA	ABG	BCG	CAG	ABC	AG	BG	CG	AB	BC	CA	ABC	X		Type
Fault	AG	283	0	0	0	0	1	1	0	0	0	3	0	1	0	1	0	0	2	0.97	AG
	BG	0	293	1	10	0	0	4	1	0	2	0	7	0	1	0	3	0	0	0.91	BG
	CG	0	0	299	0	2	1	0	0	3	0	0	0	3	0	0	0	0	2	0.96	CG
	AB	0	0	0	311	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0.99	AB
	BC	0	0	4	0	283	1	1	31	0	0	0	0	1	0	3	1	0	0	0.87	BC
	CA	0	0	14	0	0	307	0	0	0	0	0	0	0	0	0	4	0	0	0.94	CA
	ABG	0	2	0	25	0	0	266	0	0	3	0	1	0	1	0	0	0	0	0.89	ABG
	BCG	0	0	1	0	2	0	0	312	0	12	2	0	0	0	4	2	0	1	0.93	BCG
	CAG	0	0	18	0	0	8	0	0	304	1	0	0	0	0	0	3	0	4	0.90	CAG
	ABC	1	0	0	0	0	0	1	41	0	292	0	0	0	0	2	0	4	3	0.85	ABC
Line loss	A	23	0	0	0	0	0	0	0	0	0	91	0	0	10	0	0	8	27	0.57	A
	B	0	5	0	2	3	0	0	0	0	0	0	63	0	28	0	1	0	57	0.40	B
	C	0	0	12	0	3	0	0	0	0	0	0	0	83	0	0	7	0	49	0.54	C
	AB	0	0	0	9	1	0	6	0	0	1	35	2	0	189	0	4	15	35	0.64	AB
	BC	1	0	0	0	13	0	0	10	0	1	0	0	32	0	177	4	5	80	0.55	BC
	CA	3	0	6	0	2	6	0	0	8	0	3	0	21	16	5	153	32	36	0.53	CA
	ABC	1	0	0	10	1	0	3	11	0	24	8	0	1	21	11	25	84	47	0.34	ABC
Normal	X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	161	1.00	X
																				0.73	Global

Figure 7. New confusion matrix with the new groupings, resulting in 19 classes.

such as: DT, LDA, kNN, SVM, and ANN. The methods were evaluated for four different data groups: with 1 and 5 PMUs, with normal sequence, and symmetrical components.

In general, the symmetrical components were more effective in classification, compared to ABC components. The use of PMUs along the network has improved overall results, but a high increase in complexity versus gain in terms of accuracy. Among the classifiers, simple methods such as kNN and DTs had competitive performance w.r.t. RNAss with a single hidden layer. On the other hand, RNAs with more than one layer and SVM with Gaussian kernel presented the highest accuracy rates. In this case, performance was above 50% with 34 classes and 70% with groupings (19 classes), where the main misclassification factor was related to high impedance faults.

Future work includes expanding simulated events, expanding the distribution network, including distributed generation, and optimizing feature extraction and PMU location, in order to maximize the classification performance.

References

[1] E. M. Stewart and A. von Meier, *Phasor Measurements for Distribution System Applications*. Wiley, 2016, pp. 1–10.

[2] S. Brahma, R. Kavasseri, H. Cao, N. R. Chaudhuri, T. Alexopoulos, and Y. Cui, "Real-Time Identification of Dynamic Events in Power Systems Using PMU Data, and Potential Applications-Models, Promises, and Challenges," *IEEE Transactions on Power Delivery*, vol. 32, no. 1, pp. 294–301, 2017.

[3] M. Wache and D. C. Murray, "Application of synchrophasor measurements for distribution networks," in *Proc. IEEE Power and Energy Society General Meeting*, July 2011, pp. 1–4.

[4] J. Sexauer, P. Javanbakht, and S. Mohagheghi, "Phasor measurement units for the distribution grid: Necessity and benefits," in *Proc. IEEE PES Innovative Smart Grid Technologies (ISGT)*, Feb 2013, pp. 1–6.

[5] A. V. Meier, D. Culler, A. McEachern, and R. Arghandeh, "Micro-synchrophasors for distribution systems," in *2014 IEEE PES Innovative Smart Grid Technologies Conference (ISGT)*, Feb 2014, pp. 1–5.

[6] X. Liang, S. Wallace, and D. Nguyen, "Rule-based data-driven analytics for wide-area fault detection using synchrophasor data," *IEEE Transactions on Industry Applications*, vol. PP, no. 99, pp. 1–1, 2016.

[7] T. Guo and J. V. Milanovic, "Online Identification of Power System Dynamic Signature Using PMU Measurements and Data Mining," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 1760–1768, 2016.

[8] O. P. Dahal, S. M. Brahma, and H. Cao, "Comprehensive clustering of disturbance events recorded by phasor measurement units," *IEEE Transactions on Power Delivery*, vol. 29, no. 3, pp. 1390–1397, Jun 2014.

[9] R. Yadav, S. Raj, and A. K. Pradham, "Real-time Event Classification in Power System with Renewables using Kernel Density Estimation and Deep Neural Network," *IEEE Transactions on Smart Grid*, pp. 1—-8, 2019.

[10] M. A. Karim, M. Chenine, K. Zhu, L. Nordstrom, and L. Nordström, "Synchrophasor-based data mining for power system fault analysis," in *Proc. 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, Oct 2012, pp. 1–8.

[11] K. Chen, C. Huang, and J. He, "Fault detection, classification and location for transmission lines and distribution systems: a review on the methods," *High Voltage*, vol. 1, no. 1, pp. 25–33, 2016.

[12] I. Niazazari and H. Livani, "A PMU-data-driven disruptive event classification in distribution systems," *Electric Power Systems Research*, vol. 157, no. 1, pp. 251–260, 2018.

[13] A. Shahsavari, M. Farajollahi, E. Stewart, E. Cortez, and H. Mohsenian-Rad, "Situational Awareness in Distribution Grid Using Micro-PMU Data: A Machine Learning Approach," *IEEE Transactions on Smart Grid*, pp. 1–8, 2019.

[14] Y. Zhou, R. Arghandeh, I. Konstantakopoulos, S. Abdullah, A. Von Meier, and C. J. Spanos, "Abnormal event detection with high resolution micro-PMU data," in *Proc. 19th Power Systems Computation Conference*. Power Systems Computation Conference, 2016, pp. 1–7.

[15] S. A. IEEE, "IEEE Standard for Synchrophasor Measurements for Power Systems," *IEEE Std C37.118.1-2011 (Revision of IEEE Std C37.118-2005)*, pp. 1–61, Dec. 2011.

[16] IEEE Power and Energy Society. (2014) Distribution test feeders, 123 – bus feeder. [Online]. Available: https://ewh.ieee.org/soc/pes/dsacom/testfeeders/

[17] F. L. Grando, C. Gobbato, G. W. Denardin, and M. Moreto, "Phasor and frequency measurements in power systems: Hardware strategy to improve accuracy in estimation algorithms," in *12th IEEE/IAS International Conference on Industry Applications*, 2016.

[18] V. S. Cherkassky and F. Mulier, *Learning from Data: Concepts, Theory, and Methods*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1998.

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