Identification of Contradictory Patterns in Experimental Datasets for the Development of Models for Electrical Cables Diagnostics

Piero Baraldi, M. Compare, Enrico Zio, M. de Nigris, G. Rizzi

Identification of Contradictory Patterns in Experimental Datasets for the Development of Models for Electrical Cables Diagnostics

P. Baraldi[1], M. Compare[1], E. Zio[1,2,*], M. de Nigris[3], G. Rizzi[3]

[1]*Energy Department, Politecnico di Milano, Milan, Italy*
[2]*Ecole Centrale Paris-Supelec, Paris, France*
[3]*Enea Ricerca sul Sistema Elettrico – ERSE, Milan, Italy*
**Corresponding Author*

Abstract: The health state of an electrical cable may be difficult to know, without destructive or very expensive tests. To overcome this, Partial Discharge (PD) measurements have been proposed as a relatively economic and simple-to-apply experimental technique for retrieving information on the health state of an electrical cable. The retrieval is based on a relationship between PD measurements and the health state of the cable. Given the difficulties in capturing such relationship by analytical models, empirical modeling techniques based on experimental data have been propounded. In this view, a set of PD measurements have been collected by Enea Ricerca sul Sistema Elettrico-ERSE during past campaigns, for building a diagnostic system of electrical cable health state. These experimental data may contain contradictory information which remarkably reduce the performance of the state classifier, if not a priori identified and possibly corrected. In the present paper, a novel technique based on the Adaboost algorithm is proposed for identifying contradictory PD patterns within an a priori analysis aimed at improving the diagnostic performance. Adaboost is a bootstrap-inspired, ensemble-based algorithm which has been effectively used for addressing classification problems.

1. Introduction

A correct diagnosis of the state of electrical equipment is fundamental for the effective management of power networks since it allows reliable estimation of times to failure and optimal maintenance planning. However, the state of components such as electrical cables may be difficult to diagnose unless destructive or very expensive tests are used. To overcome this, attempts have been made to modeling the relationship between informative, non-destructive measurements of equipment operation and the related state. Given the difficulties of doing this with analytical models, empirical classification techniques based on experimental data are preferred for the estimation of the mapping function between the observed parameters and the discrete set of predefined classes representing health states. The estimate of the function (mathematically also referred to as hypothesis) is determined on the basis of a set of observations for which the corresponding classes are known (these data compose the so called training set) and is then used for classifying new observations (usually referred to as test patterns).

In this work, Partial Discharge (PD) measurements are considered as indicators of localized defects in electrical cables. During past experimental campaigns, Enea Ricerca sul Sistema Elettrico (ERSE) has built a database containing the values of the PD measurements and the corresponding health state of the cable. This database contains thousands of PD patterns recorded by a software tool that processes the PD measurements when these

are performed and classified by experts on the basis of both their experience and ERSE guidelines; for a small number (43) of them, the classification of the degradation state is guaranteed based on visual inspection. This is, in fact, a very expensive task which can be performed only occasionally since it entails the extraction of the cable from the ground and it implies the unavailability of the corresponding electrical line for several hours.

Based on these observed data, an empirical classifier can be developed for relating the PD measurements (input) with the health state of the cable (output, categorized into two classes: 'Bad' and 'Good').

On the other hand, errors can occur in data collection, when processing data (e.g., transcribing, transmission, omission errors, etc. [11]), when acquiring measurements (e.g., bad sensors, operator errors, data transferring errors etc.), when handling databases, etc. Any erroneous data could undermine the informational content of a database and recorded incoherent patterns could bias the mapping function created by training the classification algorithm, thus affecting its performance.

Incoherence in the present work is considered due to two main sources:

1. In some cases, information relevant for cable classification is missing. This leads to the fact that, without knowledge of the missing information, it is possible that two input patterns with the same values can be associated to both 'Good' and 'Bad' cables in the dataset. For example, in the real dataset object of the analysis of the case study of Section 4, the information regarding the insulation material of the tested cables has not been reported in the database, although it is relevant for cable classification since it influences the sensitivity of the cable to the discharges. Figure 1 gives a sketch of this type of incoherence. For the sake of simplicity, a mono-dimensional input space is considered that is, a generic parameter representative of the PD measurement is reported in abscissa in arbitrary units. For representation purposes, it is also assumed that cables whose insulation material is oil-paper are characterized by a value of the parameter smaller than a threshold (square marker in Figure 1) in case they are of class 'Good' (crosses), and larger in case they are of class 'Bad' (circles). If all the empirical patterns available were from oil-paper cables, it would be possible to build a classifier specific for this family of cables. However, ERSE experts believe that there are few patterns in the database that refer to cables whose insulation material is Cross Linked Poly-Ethylene (XLPE) or Ethylene Propylene Rubber (EPR). These families of cables are characterized by different relationships between the input parameters and the class; in the example of Figure 1, this situation is represented by assuming that the threshold that distinguishes between "Good" and "Bad" cables has a lower value than in the case of oil-paper cables. Notice that if the information regarding the insulation material is not reported, then the PD patterns of XLPE and EPR cables are assumed to be 'oil-paper' and thus may result as incoherent: Figure 1 shows that this situation of missing information results in the projection of patterns at different ordinates (missing information) on the ordinate of the oil-paper with the consequent introduction of some incoherent patterns.
2. There are degradation mechanisms that are not reflected in the PD measurements, i.e., some cable defects cannot be diagnosed using PD measurements,

but require other investigation techniques. This results in the presence in the database of cables classified as 'Bad' on the basis of visual inspection, although the obtained PD measurements do not reveal any local defect and thus are typical of "Good" cables.

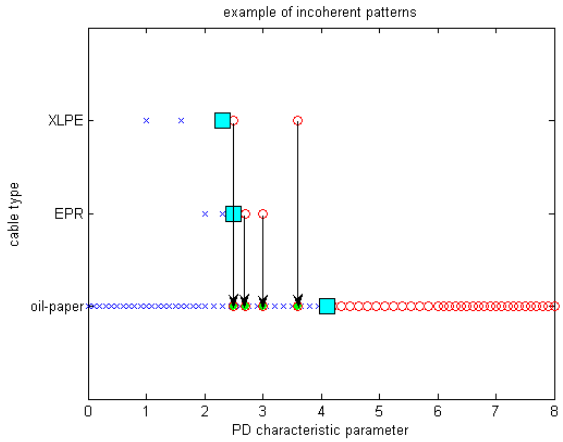

Figure 1: Example of Source of Incoherent Patterns.

Both situations described above relate to missing information in the database, i.e., variables related to the PD measurements which are available at the moment of the test but not recorded (case 1) or values of signals different from those measured in the experimental tests (case 2). The lack of such information renders some patterns of the dataset incoherent and contradictory, in the sense that identical or very similar values of the input signals are associated to different classes. The contradictory patterns in the dataset are obviously harmful for the development of empirical models that are trained on the basis of input-output data.

The objective of the present work is to propose a methodology (based on the Adaboost technique [4]) which allows identifying the contradictory patterns in a dataset, so that they can be eliminated or corrected before use for training of the classifier. In the example of Figure 1, the EPR and XLPE contradictory patterns would be identified and removed from the dataset, thereby allowing the development of a classifier of oil-paper cables.

It seems worth emphasizing that the problem here tackled differs from that of developing and using empirical classifiers in case of datasets containing missing or corrupted values, which has been largely discussed in the literature ([1], [9], [10]). Approaches to these "missing feature/missing data problems" exist, which involve estimating the values of the missing data by exploiting the presence of some patterns complete with measurements of the signals missing in other patterns. On the contrary, in the problem of interest in this work, measurements of the missing signals are not available in any pattern of the dataset, so that it is not possible to empirically infer them from available complete patterns. Thus, the solution inevitably adopted entails the removal from the dataset of the patterns that cannot be univocally classified, due to the missing information. This cleaning of the dataset is expected to improve the performance of the algorithm in the diagnosis of the health

state of those cables which can be coherently classified on the basis of the available information.

The paper is organized as follows: in Section 2, a brief outline on the characteristics of PD measurements is provided; in Section 3, the key ideas underlying the Adaboost technique are explained, and a description of the proposed methodology for the identification of the anomalous PD patterns is provided. In Section 4, the methodology is applied on the available PD measurement dataset. Finally, Section 5 concludes the paper with some considerations.

2 PD Measurement

The PD measurement is acquired through an off-line process. First, the cable under inspection is de-energized and disconnected from any source or load from all terminals. The PD measurement system, equipped with its own generator, is then connected to energize the cable with damped oscillatory voltages of frequency in the range 150 to 250Hz. In terms of signal acquisition, the system captures the high frequency signal generated by the partial discharges activity and conveys it to a signal conditioner unit through the coupling capacitor. The signal conditioner reduces the overall bandwidth of the acquired wave and amplifies it, thus enhancing its signal-to-noise ratio.

Schematically, the PD measurement is performed in three successive phases corresponding to three energizing voltage levels; in the first phase, the PD inception voltage U_i (i.e., the voltage value at which the PDs start) is reached through a stepwise or continuous increase of the voltage applied to the cable; at this voltage level, the following two parameters are collected: the PD value, i.e. the value of the discharge expressed in pC and the dispersion index, i.e. the length of the cable in which the discharge activity is localized.

In the following two phases, the cable is tested at the nominal voltage level U_0 and the maximum voltage value U_{max}, respectively, and the values of the above two parameters are measured. For convenience of data manipulation, the values of U_i, U_0 and U_{max} are normalized with respect to U_0.

The three triplets of values (normalized voltage level, PD value and dispersion index) corresponding to the three different values of voltage, constitute a pattern of 9 features in which the normalized nominal voltage value U_0/U_0 is always equal to 1; thus, this feature is non-discriminating and thus it is not considered in the diagnostic analysis. A PD measurement pattern is then made up of 8 feature.

During past campaigns, a set of 43 PD measurement patterns has been collected by ERSE. The dataset contains 16 patterns of class ‘Bad’ and 27 patterns of class ‘Good’. The classification of the health state (diagnosis) is based on visual inspections made by ERSE experts who extracted from the ground and cut up a cable section after acquiring the PD measurement patterns. For these 43 patterns, paper reports have been prepared by the experts, containing photos and further information about the tested cables and the electrical line which they belong to.

A first attempt to build a diagnostic system based on the ERSE dataset has been performed by using the Adaboost algorithm [1]. Only 28 patterns of the ERSE database are correctly classified (65%). Such low classification performance of the diagnostic system can be considered as a symptom of the presence of contradictory patterns in the database used for building the classifier.

3 Identification of Contradictory Patterns

Let x_k, k=1,…, n, be a pattern of an empirical dataset S. The information available for each pattern are the values of the f features and its class c_k, i.e. $x_k=(x_{1k}, x_{2k},..., x_{fk}, c_k)$, c_k=1,2, …, Ω (in the present case study Ω=2). Let us assume that in the dataset S there are an unknown number nc of contradictory patterns, i.e. patterns of different classes with very similar input values.

The methodology here proposed for the identification of the contradictory patterns is based on the assumption that for an empirical classification algorithm it is difficult to learn the relationships between the input signals and the class of the patterns in those zones of the input space characterized by the presence of contradictory patterns.

Different types of classification algorithms give different warnings of their difficulties in learning the training dataset S. This Section investigates the behavior of the Adaboost classification approach in the case in which a dataset S containing contradictory patterns is used to train the classifiers. Section 3.1 summarizes the main concepts of the Adaboost algorithm, whereas in Section 3.2 an indicator of the degree of contradictoriness of the patterns is introduced. In Section 3.3, the capability of the proposed indicator in distinguishing contradictory patterns is verified on artificial case studies and finally in Section 3.4 the performance of an empirical classifier trained on a dataset which does not contain the patterns identified as contradictory is analyzed. Section 3.5 analyses the sensitivity of the performances of the proposed methodology to the main parameters which it depends on.

3.1 Adaboost Algorithm

Adaboost Algorithm is one of the most influential classification algorithms in recent history of computational intelligence [1]; in particular, it is one of the best known ensemble-based algorithms. The underlying idea of ensemble algorithms is derived from daily life decision making: in the hope to making a more informed decision, a number of individual opinions are usually sought and then opportunely weighted and combined to elaborate the ultimate decision [1]. In the health state classification task of interest here, a number of diverse classifiers provide different mapping functions (corresponding to the opinions of the daily life example) whose combination may provide a superior mapping function than that provided by any single classifier. A fundamental issue for the success of an ensemble is the negative correlation of these classifiers, i.e., their capability of erring on different sub-regions of the input space [12]. In this respect, notice that only a negative correlated ensemble allows reducing the variance and increasing the confidence of the decision with respect to a single classifier. In fact, there are random aspects in classification (due to training data, initializations, etc.) which may lead to substantially varying decisions. Then, combining the outputs of several such classifiers can reduce the risk of an unfortunate selection of a poorly performing classifier.

Effective ensemble-based classification algorithms have been recently developed by resorting to the bootstrap method. Bootstrapping [2] is a computer-intensive re-sampling method whose key idea is to treat the available dataset, $S=\{x_1,x_2,...,x_n\}$, as if it were the entire population and then to create an opportune number, B, of alternative versions of S (S_b*, b=1,2,..., B) by randomly sampling from it with replacement (i.e., every sample is returned to S after sampling so that a particular data point could appear multiple times in a bootstrap sample). In the bootstrap-inspired ensemble-based classification algorithms, a number B of classifiers are generated by training with bootstrap samples S_b*.

The main characteristics of the Adaboost algorithm are suggested by its name which stands for Adaptive Boosting. It is a boosting algorithm: a sequence of B classifiers, C_b, b=1,2,…B, is created by training a classifier algorithm on different bootstrap samples S_b*, b=1,2,…, B. The probability mass distribution, $D_b=\{p_b(1), p_b(2),\ldots, p_b(n)\}$, whose generic element $p_b(k)$ gives the probability of drawing pattern x_k from S in the bootstrap sample S_b*, is opportunely altered after building a classifier in order to ensure that more informative points are drawn into the next dataset used for building the successive classifier. In this sense, Adaboost is adaptive because it updates the distribution D such that after a classifier C_b is built, the subsequent classifier, C_{b+1}, pays more attention to training patterns that were misclassified by C_b. In particular, if pattern x_k is misclassified by the generic classifier C_b, then the probability p_{b+1} (k) that x_k is drawn when building the next data training set (S_{b+1}^*) is increased with respect to $p_b(k)$; on the contrary, sampling probabilities of the points correctly classified are reduced. In this way, the probability that S_{b+1}^* will contain a larger number of patterns x_k increases and this gives the classifier C_{b+1} more chances to correctly classify x_k. In case x_k is again misclassified, then $p_{b+2}(k)$ will be further enhanced. The increasing behavior of the sampling probability associated to x_k ends when a classifier that is able to correctly classify x_k is built. In this way, subsequent classifiers are tweaked in favor of those patterns misclassified by previous classifiers and thus tend to have higher performance on these difficult patterns.

The classifiers are then combined through weight majority voting to obtain the final classification. The voting weight of a classifier is strictly dependent on its performance: the larger the number of patterns of S correctly classified the larger its vote.

With respect to the choice of the classification algorithm, weak learner algorithms characterized by a performance slightly more accurate than the random guessing are used within the Adaboost scheme. Typically, a weak learner is obtained by applying a short training session to empirical algorithms such as Neural Networks (NN), Multi-Layer Perceptron (MLP), Radial Basis Function (RBF), naïve Bayesian decision trees and k-Nearest Neighbors (KNN) [13]. In this work, the Evolutionary Fuzzy C-Means (EFCM, see [7]) has been considered as base classifier of the Adaboost algorithm. The use of this algorithm in combination with the Adaboost scheme is a novelty. EFCM is a supervised classification algorithm that uses the knowledge of the class of the patterns for finding for each class an optimal Mahalanobis metric that defines a geometric cluster as close as possible to the a priori known class. The Mahalanobis metrics are defined by the matrices whose elements are identified by the supervised evolutionary algorithm so as to minimize the distances between the patterns of each class and the center (also referred to as cluster prototype) of the corresponding cluster. Since the EFCM iteratively searches for an optimal Mahalanobis, it is possible to obtain a weak learner classifier characterized by a short training time by reducing the number of iteration performed during the search.

More details about the Adaboost algorithm and its pseudo-code are given in Appendix 2, whereas further details on the EFCM algorithm can be found in [7] and [8].

3.2 Degree of Contradictoriness

Within an Adaboost classification approach, contradictory patterns are expected to be among the patterns that are misclassified by the ensemble classifiers and thus with an associated high value of the probability mass functions $p_b(k)$, $b=1, 2,\ldots, B$. The idea is thus

to consider as indicator of the degree of contradictoriness of pattern x_k, k=1, 2,…, n, the quantity:

$$w_k = \frac{\sum_{b=1}^{B} p_b(k)}{B} \quad (1)$$

Given the updating dynamics of the distribution D, a pattern x_k which is correctly classified by all classifiers C_b, b=1,…,B, is associated to low values of $p_b(k)$, b=1,…,B. On the contrary, the probability masses $p_b(k)$ associated to patterns which are difficult to be classified have the oscillating behavior described in Section 3.1; thus, the mean value w_k of the probability masses associated to these patterns tends to be larger: the contradictory patterns of S are then expected to occupy the first positions of the ranking of the values $w_1, w_2,\ldots,w_n$.

3.3 Simulation Results

The capability of the proposed indicator w_k in identifying the contradictory patterns is firstly verified with respect to artificial datasets. The artificial dataset A considered is formed by n=43 patterns of 2 classes (16 of class "Bad" and 27 of class "Good"), in analogy to the PD measurement dataset. Patterns are randomly drawn from 3-dimensional Gaussian distributions whose centers and standard deviations are reported in Table 1. Three features instead of the 8 features of the PD dataset have been considered, in order to allow visualization. A dataset S containing nc contradictory patterns is obtained by randomly selecting from A nc patterns that are associated to the wrong class, i.e., if the pattern selected is classified as "Bad" in A, then it is forced to "Good" in S and vice-versa. The set of nc contradictory patterns will be called S_c. The number of nc=5 contradictory patterns (approximately 10% of the entire dataset) has been suggested by ERSE experts. In order to compute the degree of contradictoriness of the patterns, S has been used as training dataset for the Adaboost algorithm with a number of classifiers B=30. The computational time required by the Adaboost algorithm has been of about 15 minutes, on an Intel Centrino® CPU, 1.20 GHz, with 1GB RAM. Since the identification of the outliers using the Adaboost algorithm is performed off-line before the development of the final classification, the computational time is acceptable.

Table 1: Clusters Centers and Standard Deviations.

	Bad	Good
Center	[0.7 0.65 0.9]	[-0.3 0.3 0.3]
Std. Deviation	[0.3 0.2 0.25]	[0.7 0.25 0.25]

Table 2: Methodology Performances.

ni/nc	Σ
0.71	0.17

Once the mean values of w_k, k=*1,2,..,n*, of the degrees of contradictoriness of (Equation 1) have been computed, the contradictory patterns are identified by fixing a threshold value T for w_k: all the nr patterns with w_k larger than T form the set $\hat{S}_c=\{k \mid w_k \geq T\}$ and are removed from the dataset. The threshold T has been defined by:

$$T = W + V$$

being W and V the mean and standard deviation of the vector $\{w_1, w_2, \ldots, w_n\}$, respectively. In this way, the patterns with very high degrees of contradictoriness with respect to those of the other patterns of A are the most plausible candidates to be contradictory.

In order to cross-validate the results, 50 different datasets randomly obtained from the Gaussian distribution of Table 1 have been generated. In each dataset, 5 contradictory patterns have been randomly selected. Table 2 reports the mean fraction *ni/nc* of the contradictory patterns correctly identified by the proposed criterion and the standard deviation σ of the 50 collected values of *ni/nc*. In practice, almost 70% of the contradictory patterns are correctly identified.

To further delve into the proposed methodology, a detailed analysis of its application on one of the 50 considered Gaussian datasets is illustrated in details. Figure 2 shows the position of the patterns in the input space, the set S_c of the contradictory patterns and the set $\hat{S}_c$ of the patterns identified as contradictory by applying the selection criterion. This identifies 3 of the 5 contradictory patterns of S_c; the remaining 2 contradictory patterns of S_c which have not been identified are at the border between the classes 'Good' and 'Bad' in a zone of the input space with a low density of patterns. Thus, when training a classification algorithm on these data, a mapping function which assigns these points to the class 'Bad' could be built, as well as one which assigns them to the class 'Good'. Notice that, the considered criterion identifies as contradictory a pattern in the top right part of Figure 3 which does not belong to S_c but is close to a pattern of S_c. This pattern of S_c and some patterns of class 'Good' that are in this zone of the input space are difficult to classify; the Adaboost algorithm tends to assign to them an high value of the probability $p_b(k)$ of being drawn in the bootstrap samples S_b*. Thus, the corresponding classifier C_b is forced to build a mapping function that assigns the class 'Good' also to the neighbors of these patterns that are 'Bad' and thus tend to misclassify them; in the next iteration of the Adaboost algorithm, a larger value of $p_{b+1}(k)$ is assigned to the misclassified patterns with the consequent oscillating behavior of the probabilities $p_b(k)$ associated to the patterns of this zone, which results in large values of the degree of contradictoriness.

3.4 Developing a Classifier on Coherent Patterns Only

For illustration purposes, it is assumed that errors are introduced when collecting data, so that the training datasets contain contradictory patterns; if no action of removing or correcting the contradictory patterns is performed, then all the patterns are used to train the corresponding classifiers. In general, this results in building classifiers with low performances since all the contradictory information concurs in creating biased mapping functions.

On the contrary, the methodology here proposed adjusts the given dataset by removing the *nr* patterns of the set $\hat{S}_c$ identified according to the selection criterion. In this respect, the evaluation of the performances of a classifier trained with the set *S'* of the *n-nr* remaining patterns constitutes a useful tool to assess the effectiveness of the proposed me-

thodology. Since our objective is to develop an empirical classifier able to correctly classify the patterns that do not belong to contradictory zones of the training space, the performance of the methodology is evaluated with respect to the ratio between the number r of non-contradictory patterns correctly classified and the total number of non-contradictory patterns (in the considered case study n-nc=43-5=38):

$$Perf = \frac{r}{n - nc}$$

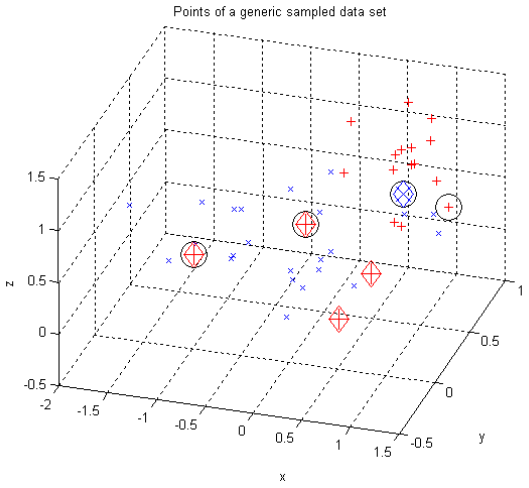

Figure 2: Dataset S: '+' indicates patterns of class 'Bad'; 'x' indicates patterns of class 'Good'; ◊ indicates contradictory patterns: when it contains a '+' marker, then it refers to patterns of class 'Good' in A but of class 'Bad' in S; on the contrary, when ◊ contains a 'x' marker, then it refers to patterns of class 'Bad' in A but of class 'Good' in S. 'O' indicates the patterns selected by the considered criterion.

Notice that the classification of the contradictory patterns is not taken into account since the empirical model has not been trained with patterns in these zones of the input space.
In order to cross-validate the results, this classification performance is computed in a LOO scheme. Basically, an a priori known non-contradictory pattern x_k is omitted from the set S' of the patterns considered by the criterion as non-contradictory; the remaining patterns of S' are used to train a classifier which predicts (correctly or incorrectly) the class of the omitted instances x_k. The process is repeated for all the a priori known non-contradictory instances in S'.

For example, with reference to the dataset considered in Figure 2, the selection criterion removes from the original dataset the nr=4 patterns with the circle marker so that the remaining set S' contains the n-nr=39 points not marked by circles. Figure 3 shows the points of S' in the input space.

In the reference case of Figure 2 and Figure 3 ni=3, i.e., 3 contradictory patterns of S_c are in the removed group $\hat{S}_c$, thus nc-ni=2 contradictory patterns of S_c are still among the 39 patterns of S' (diamond markers in Figure 3) and nr-ni=1 non-contradictory pattern does not belong to S'. In order to compute r in Equation 2, the following procedure is adopted:

1) The (*n*-*nr*)-(*nc*-*ni*)=37 non-contradictory patterns of *S'* (i.e., the number of patterns in *S'* minus the number of patterns of S_c not removed from *S*) are tested in a LOO approach: one of these patterns is omitted from the training set, the classifier is built and the omitted pattern is classified. The process is repeated for all the 37 non-contradictory patterns and the total number of correctly classified patterns gives r_1. Figure 3 shows that there are 3 points misclassified (indicated by star markers). In this case r_1 is equal to 34.

A classifier is trained using all the patterns in *S*' and tested using the *nr*-*ni* =1 non-contradictory pattern not belonging to *S'*; in general, r_2 is given by the number of the *nr*-*ni* patterns correctly classified. In the reference dataset, there is only one point which is correctly classified so that r_2=1.

Finally, *r* is given by r_1+r_2; in the reference case study, *r*=35.

The performance obtained in the classification of non-contradictory patterns by applying the considered selection criterion is compared to the performances obtained in the following two reference cases:

all the contradictory patterns are a priori known and thus the classifier is trained by using all the non-contradictory patterns. In general, this corresponds to the best possible situation since all the contradictory information is not used to train the classifier.

a) all the available patterns of *S* are used to train the classifier. In general, this corresponds to the worst possible case since all the contradictory information is used to train the classifier.

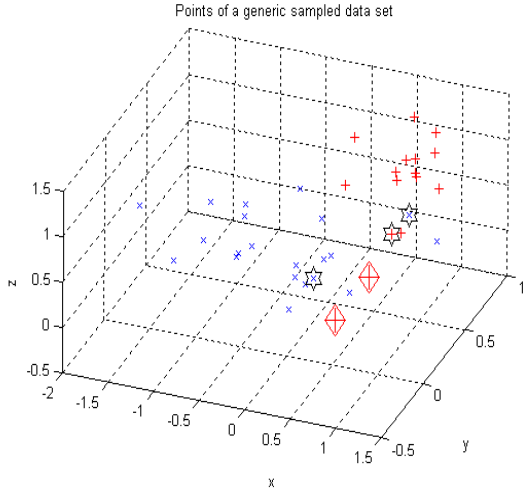

Figure 3: Representation of the Points of S' in the Input Space.

Both in a) and b) the results are cross-validated in a LOO scheme. In order to verify that the results do not depend on the particular Gaussian distribution of the data nor on the particular choice of the 5 contradictory patterns, 50 different datasets have been considered. Table 3 reports the mean values and standard deviations of the obtained performances. It can be noted that the chosen performance indicator of the criterion is between those of the best and the worst case.

Table 3: Mean and Standard Deviation of *Perf* Indicator.

	Mean	Std
Selection Criterion	0.919	0.053
Best Case	0.928	0.053
Worst Case	0.897	0.070

3.5 Sensitivity Analysis

The performances of the proposed methodology depend on some parameters such as the number of classifiers, the distance between the centers of the two clusters and the number of contradictory patterns contained in the dataset. A sensitivity analysis is performed in this Section in order to evaluate their importance. In particular, a local approach in which these parameters are varied one at a time [5] is applied. This investigation also resorts to the creation of artificial datasets in order to increase the confidence on the estimation of the parameters of interest (i.e., *ni* and *nr*).

Figure 4 shows that both the number *ni* of patterns correctly identified and the number *nr* of patterns removed by the considered criterion are not sensitive to variations of the number *B* of classifiers used by the Adaboost algorithm. In particular, Figure 4 left, shows that the ratio between the number *ni* and the number *nc* of contradictory patterns contained in the original dataset (which is an indicator of the effectiveness of the considered selection criterion) remains approximately constant when *B* varies. Figure 4 right, reports an indicator of the efficiency of the selection criterion: the larger the ordinates the smaller the number of patterns incorrectly removed; this efficiency indicator is also nearly constant when *B* varies. This is due to the fact that in the considered case the probability mass functions D_b evolve so that the degrees of contradictoriness w_k of few patterns are outside the interval $W+V$.

Figure 5 describes the behavior of the parameters of interest (*nr* and *ni*) when the number *nc* of contradictory patterns contained in the original dataset varies. In particular, the parameter *ni* is very sensitive to the variation of *nc*: the larger *nc* the lower the performances in identifying incoherent patterns. This can be explained by considering that if the number of contradictory patterns increases, then the classification performances tend to decrease with a generalized enhancement of the degrees of contradictoriness w_k associated to the patterns. Thus, also recognizing the contradictory patterns among the correct ones becomes very difficult.

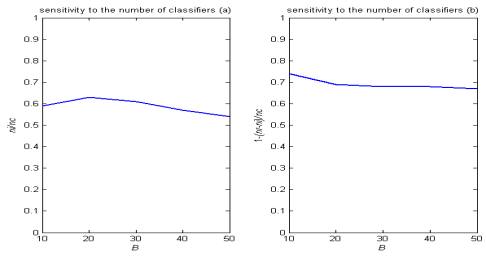

Figure 4: Sensitivity of the Selection Criterion to the Number of Classifiers *B*.

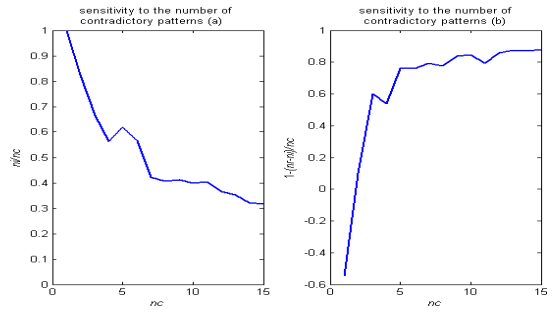

Figure 5: Sensitivity of the Selection Criterion to the Number *nc* of Contradictory Patterns contained in the Dataset.

Figure 6 confirms this aspect; it shows that the standard deviation of the degrees of contradictoriness w_k decreases when the number of contradictory patterns contained in the dataset increases. This also leads to the fact that the number of patterns incorrectly removed decreases when *nc* increases (Figure 5). For example, if *nc*=1, then there are on average *nr*=2.55 patterns with very large degrees of contradictoriness which are removed; among these patterns there is always the contradictory one (*ni*=1); thus, there are 1.55 patterns incorrectly removed from the dataset. On the contrary, when *nc*=15, then *nr*=6.55<<*nc*=15 and *ni*=4.75: there is a large probability (0.72) that in the few removed patterns there are the contradictory ones. In conclusion, the selection criterion distinguishes the correct and the contradictory patterns only when the corresponding mean weights w_k are really far from those of the other patterns.

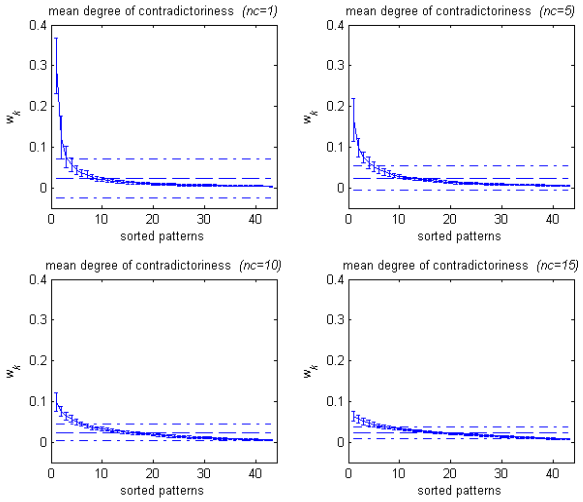

Figure 6: Mean Values and the Corresponding 68% Confidence Intervals of the Degrees of Contradictoriness of the Patterns for Different Values of *nc*.

Figure 7 shows that the number *nr* is quite sensitive to the distance between the two clusters representative of the health state classes. The abscissa of Figure 7 is an index of the distance between the cluster centers with 1 representing the largest considered distance, 4 the distance relative to the clusters considered in Table 1 and 7 the smallest considered distance. In practice, the more the two classes are distinct, the larger are the performances of the selection criterion. This is also an expected result: when the two clusters are more overlapped there is a generalized increment of the degrees of contradictoriness w_k associated to the patterns which makes more difficult the classification. This affects also the capability of the selection criterion to distinguish the contradictory patterns. Figure 7 also shows that when the clusters are very distant, then the criterion is very efficient because it identifies more than 60% of anomalous patterns with a small number of incorrect removals of patterns. This effectiveness decreases when the clusters are more overlapped because of the generalized enhancement of the weights w_k associated to the patterns which increases the probability that a non-contradictory pattern has a large value of w_k.

Finally, a further analysis has been performed in order to investigate the behavior of the proposed methodology in case of more populated training datasets that contain the same percentage (almost 12%) of contradictory patterns. It can be noted (Figure 8) that the larger the number *n* of patterns the larger the fraction of contradictory patterns correctly identified. This is due to the fact that when the cardinality of the training dataset is larger, the shapes of the two clusters are better defined and thus it is easier to recognize patterns that are contradictory. The counterpart is that outlier patterns have more chances to be selected among the *nr* patterns that are candidates to be contradictory, and this results in a loss of information.

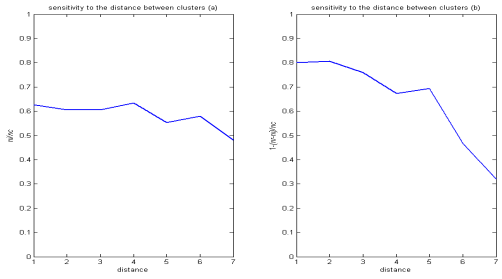

Figure 7: sensitivity of the selection criterion to the distance between the two clusters.

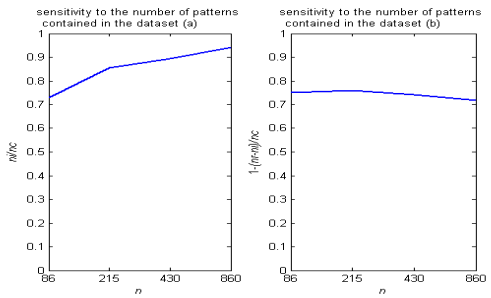

Figure 8: Performances of the considered Selection Criterion when varying the Number of Patterns in *S*.

4 Application to the Erse Pd Measurement Dataset

The present Section reports the results obtained by applying the methodology proposed above for the classification of contradictory patterns to the ERSE PD measurement dataset described in Section 2.

The proposed methodology identifies as contradictory 7 patterns out of the 43 of the dataset (Table 4). For confirmation, these patterns have been further analyzed by resorting to expert opinions supported by the available paper reports describing the results of the visual inspection of the cable.

Table 4: Patterns Identified as Contradictory by the Proposed Methodology

U_i/U_0	PD value	disp ID	PD value	disp ID	U_{max}/U_0	PD value	disp ID	Visual insp. classification
1	905	3	905	3	1.5	4202	3.5	Good
0.88	2000	2	8750	2.5	1.38	10500	4.5	Good
1	349	2.5	349	2.5	1.56	990	2	Good
0.63	560	3	894	4	1	894	4	Bad
0.63	6631	7.5	12291	7.5	1	12291	7.5	Good
0.81	5051	3.5	8453	2	1.19	11546	3	Good
0.88	3385	3	7500	3.5	1.19	11157	4.5	Good

It turns out that (Table 5):

a) Three of the seven patterns (rows 2, 4, 5 of Table 4) can be really considered contradictory since the information provided by the PD measurements is not sufficient for their correct classification. In particular, one of these patterns (row 4 of Table 4) is associated to a joint which belongs to a part of an electrical line made up of XLPE cables. Notice that this information on the cable insulation material is omitted in the digital PD measurement dataset whose other available patterns refer to oil-paper cables for which different relationships between the PD measurements and the health state arise. Thus, it seems correct to consider this pattern as contradictory. Two other contradictory patterns (rows 2 and 5) refer to PD measurements acquired in a test characterized by high dispersion of the discharges along the cables. This fact indicates that the discharges were not due to a localized defect, as in the other considered patterns, but to other causes. Also in this case, due to the missing information on the discharge dispersion not provided by the PD measurements, the patterns can be correctly considered contradictory.

b) The classification reported in the ERSE dataset for two other patterns (rows 6 and 7 of Table 4) identified as contradictory seems not justified considering the information available in the paper reports. Thus, the experts indicate also these two patterns as contradictory.

Finally, no apparent anomalies are found in the remaining two patterns (rows 1 and 3 of Table 4) identified as contradictory by the algorithm. Thus, these patterns seem incorrectly identified as contradictory by the proposed methodology.

Table 5: Expert Judgment on the Patterns identified as Contradictory by the proposed Methodology.

Number of selected patterns *nr*	7

Number of contradictory patterns for the experts	3
Number of patterns whose classification in the dataset seems not justified.	2
Number of non-contradictory patterns for the experts	2

Notice that, expert opinions on the contradictoriness of the patterns selected by the proposed methodology are not expected to be available in future applications, and one can only remove the identified patterns from the training set of the classifier and verify its classification performance. If the classifier trained without considering the patterns identified as contradictory outperforms the classifier trained with all the patterns, the analyst can be confident on the correctness of the identification of the contradictory patterns.

In this case study, once the $nr=7$ selected patterns have been removed from the given dataset, a classifier is trained on the set S' of the remaining 36 patterns in order to provide ERSE with the final diagnostic system for oil-paper electrical cables. In this respect, the Adaboost algorithm has been adopted as classification algorithm, as its classification performance is generally higher than that of other classification algorithms based on a single classifier (e.g., EFCM; see Section 3.1 and related references). The results obtained have been compared with those obtained by an Adaboost algorithm trained with the set S of all 43 available patterns (worst case). In particular, the comparison is made considering S' as test set since the Adaboost classifiers are not trained on the removed patterns.

In order to get a reliable estimation of the classification performance, the LOO cross validation scheme introduced in Section 3.3 has been applied to provide an unbiased estimation of the true error of the classifier. The LOO scheme also allows detecting over-fitting; indeed, over-fitted classifiers are characterized by high training performance and generally low test performance, i.e., they learn by heart the classes associated to the points of the training set, but are not robust enough to correctly classify points different from those of the training set. In the present case study, the very high classification performance in testing by all the 36 different classifiers (trained on 36 different sets), makes us confident that there is no over-fitting. Furthermore, empirical results have shown that Adaboost is resistant to the phenomenon of the over-fitting, as explained on the basis of margin theory [1].

Table 6 summarizes the performance obtained in test: the removal of the patterns identified as contradictory by the proposed methodology results in a remarkable increase of the performance of the diagnostic system, which in the present case becomes infallible.

Table 6: Performance of the Diagnostic System.

Training Set	Performance on Test Set S' (36 patterns)
S (43 patterns)	0.611
S' (36 patterns)	1

5 Conclusions

Errors in data collection can result in databases containing contradictory patterns which could bias the mapping function created by training a classification algorithm; this can significantly affect the classification performance.

An original methodology which allows recognizing contradictory patterns has been proposed here and its performance evaluated on some artificial case studies. The metho-

dology has then been applied with satisfactory results to the PD measurements dataset collected by ERSE for diagnosing the health state of electrical cables.

References

[1]. Polikar, R. *Bootstrap Inspired Techniques in Computational Intelligence.* IEEE Signal Processing Magazine. 2007. Vol. 24 No. 4, pp. 59-73

[2]. Efron, B. *Bootstrap Methods: Another look at the Jackknife.* Annals of Statistics. 1979. Vol.7, no. 1, pp. 1-26.

[3]. Shapire, R.E. *The Strength of weak learnability*. Machine Learning. 1990. Vol. 5 no 2, pp. 197-227.

[4]. Freund, Y. and Shapire, R.E. *A Decision-Theoretic generalization of On line Learning and Application to Boosting*. Journal of computing and system Science. 1997. Vol. 55, pp. 119-139.

[5]. Zio, E. *Computational methods for reliability and risk analysis*. 2009. World Scientifics.

[6]. Boland, P.J. *Majority Systems and the Condorcet jury problem*. Statistician. 1989. Vol. 30, no. 3, pp. 181-189.

[7]. Yuan, B. and Klir, G. *Data driven identification of key variables*. in D. Ruan (Ed.), Intelligent Hybrid Systems Fuzzy Logic, Neural Network, and Genetic Algorithms, pp. 161–187 (Chapter 7). 1997. Kluver Academic Publishers.

[8]. Zio, E. and Baraldi, P. *Identification of Nuclear Transients via Optimized Fuzzy Clustering*. Annals of Nuclear Energy. 2005. Vol. 32, No. 10, pp. 1068-1080.

[9]. Ho, T.K. *Random Subspace method for constructing decision forests.* IEEE Trans. Pattern Anal. Machine Intell. 1998. Vol. 20, No. 8, pp. 832-844.

[10]. DePasquale, J. and Polikar, R. *Random Feature Subset Selection for Ensamble Based Classification of Data with Missing Features*. Lecture Notes in Computer Science. 2007. Vol 4472. pp. 251-260. M. Haindl and F. Roli, Eds Berlin. Springer-Verlag.

[11]. Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget. *Statistical Policy-Working Paper 31: Measuring and Reporting Sources of Error in Surveys.* 2001.

[12]. Kuncheva, L. *Combining Pattern Classifiers: Methods and Algorithms*. 2004. John Wiley & Sons, Chichester.

[13]. Tang, F. Grinding Wheel Condition Monitoring With Boosted Classifiers. PhD Dissertation Thesis. 2006.

Appendix 1

In supervised classification, we are given a training dataset S={x_1,x_2,…,x_n}. where $x_i \in X$ is the i-th instance in the f-dimensional feature space X. A class $c_i \in \Gamma$ and $\Gamma = \{c_1, c_2, \ldots c_\Omega\}$ is assigned to every xi. An unknown mapping function m: X→Γ assigns to each point of S the corresponding true class. A classifier C trained on S produces a mapping function (also referred to as hypothesis) h: X→Γ which is an estimate of *m(x)*. When building a classifier, it is very important to estimate its performances in classifying previously unseen field data. In this respect, the probability that an instance drawn from the input space is misclassified by C, also referred to as the true error, is an unknown value that needs to be estimated. An unbiased estimator of the true error is provided by the Leave-One-Out (LOO) approach.

In the LOO approach, an instance is omitted from the training sample; when the classifier is built, the prediction (correct or incorrect) for the omitted instances is obtained; the process is repeated for all the instances in the training sample; the estimation of the true error is given by the proportion of instances incorrectly classified. This estimator has low bias but its variance tends to be large.

Appendix 2

In this Appendix, the pseudo-code of Adaboost.M1 [1], [4], the most popular Adaboost's variation, is described and commented (see Figure 9).

Inputs for Algorithm:

- Training dataset $S=\{x_1,x_2,\ldots,x_n\}$, with correct classes c_i ($c_i \in \Gamma$ and $\Gamma=\{c_1,c_2,\ldots.c_\Omega\}$) assigned to every x_i.
- Weak learning algorithm **EFCM**.
- Number of classifiers B

Initialization: $D_1(i)=1/n$; $i=1,\ldots n$

Do for $b=1,2,\ldots.B$

1. Draw bootstrap training data subset S_b^* according to current distribution D_b.
2. Train **EFCM** algorithm with S_b^*; the hypothesis h_b is provided in output.
3. Calculate the error of h_b by:

$$\varepsilon_b = \sum_{i=1}^{n} I\left\|h_b(x_i) \neq c_i\right\| \cdot p_b(x_i) \text{ where } I\left\|h_b(x_i) \neq c_i\right\| = \begin{cases} 1 & if\ h_b(x_i) \neq c_i \\ 0 & otherwise \end{cases}.$$

Adaboost algorithm requires that the error ε_b be smaller than one half [3]; this requirement has its root in the Jury Condorcet Theorem [6].

4. **If** ε_b >0.5 **Then**
   Repeat point 3
   End if
5. Calculate normalized error $\beta_b = \dfrac{\varepsilon_b}{1-\varepsilon_b}$ (Notice that $0 \leq \beta_b \leq 1$).
6. Update distribution D_b $p_{b+1}(x_i) = \dfrac{p_{b+1}(x_i)}{Z_b} \cdot \begin{cases} \beta_b & if\ h_b(x_i) = c_i \\ 1 & otherwise \end{cases}$

where Z_b is the normalization constant chosen so that D_{b+1} is a proper probability mass function.

End Do loop

Test

1. Obtain total vote received by each class: $V_j = \sum_{\{b|h_b=c_j\}} \log\left(\frac{1}{\beta_b}\right)$ $j=1,2,\ldots,\Omega$
2. Choose the class that receives the highest total vote as the final classification

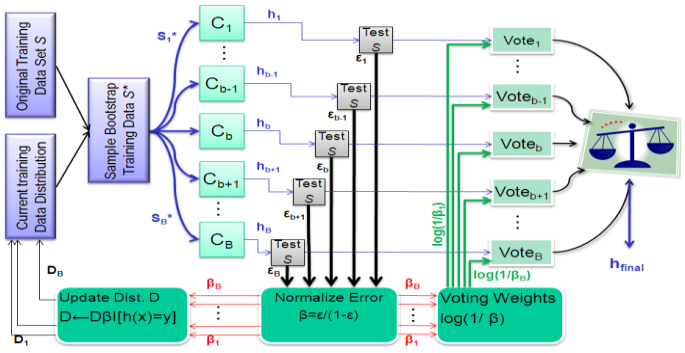

Figure 9: Block Diagram of Adaboost.

Legend

A=Gaussian dataset with no contradictory pattern.
B=number of classifiers created by the Adaboost algorithm.
C_b=b-th classifier (b=1,…,B).
D_b=probability mass function used to create the b-th bootstrap sample S_b*, b=1,…,B.

EFCM=Evolutionary Fuzzy C-Means
f=number of features of the patterns.
h_b=b-th mapping function b=1,…,B.
LOO=Leave One Out
n=number of patterns in the original training dataset S.
nc=number of contradictory patterns contained in S.
ni=number of patterns correctly identified.
nr=number of removed patterns according to the considered selection criterion.
$p_b(k)$=sampling probability associated to the k-th pattern at the b-th iteration of the Adaboost Algorithm (b=1,…,B and k=1, 2,…, n).
Perf=performance indicator.
r=number of non-contradictory patterns correctly classified.
r_1=number of patterns correctly classified among the a priori known non-contradictory patterns which are not selected by the considered selection criterion.
r_2=number of patterns correctly classified among the a priori known non-contradictory patterns which are selected by the considered selection criterion.
S'=dataset obtained by removing the patterns identified by the selection criterion from S
S=working dataset.
S_b*=b-th bootstrap sample created according to the distribution D_b, b=1,…,B.
S_c=set of the contradictory patterns contained in S.
$\hat{S}_c$=estimation of S_c made by a selection criterion (*i.e.*, the set of the patterns identified by a selection criterion as candidates to be contradictory).
T = threshold on w_k.
U_i= the voltage value at which the PDs start.
U_0= the ominal voltage value.
U_{max}= maximum voltage value.
w_k=degree of contradictoriness associated to the pattern x_k, k=1, 2,…, n.
x_k=k-th patterns of S, k=1, 2,…, n.
Ω=number of classes.

Enrico Zio (BS in nuclear engng., Politecnico di Milano, 1991; MSc in mechanical engng., UCLA, 1995; PhD, in nuclear engng., Politecnico di Milano, 1995; PhD, in nuclear engng., MIT, 1998) is Director of the Chair in Complex Systems and the Energetic Challenge of Ecole Centrale Paris and Supelec, Director of the Graduate School of the Politecnico di Milano, full professor of Computational Methods for Safety and Risk Analysis, adjunct professor in Risk Analysis at the University of Stavanger, Norway, and invited lecturer and committee member at various Master and PhD Programs in Italy and abroad.

He has served as Vice-Chairman of the European Safety and Reliability Association, ESRA (2000-2005) and as Editor-in-Chief of the International journal Risk, Decision and Policy (2003-2004). He is currently the Chairman of the Italian Chapter of the IEEE Reliability Society (2001-). He is member of the Korean Nuclear society and China Prognostics and Health Management society.

He is member of the editorial board of the international scientific journals Reliability Engineering and System Safety, Journal of Risk and Reliability, Journal of Science and Technology of Nuclear Installations, plus a number of others in the reliability, safety and nuclear energy fields.

He has functioned as Scientific Chairman of three International Conferences and as Associate General Chairman of two others, all in the field of Safety and Reliability.

His research topics are: analysis of the reliability, safety and security of complex systems under stationary and dynamic conditions, particularly by Monte Carlo simulation methods; development of soft computing techniques (neural networks, fuzzy logic, genet-

ic algorithms) for safety, reliability and maintenance applications, system monitoring, fault diagnosis and prognosis, and optimal design.

He is co-author of three international books and more than 150 papers in international journals, and serves as referee of more than 20 international journals.

Piero Baraldi (BS in nuclear engng., Politecnico di Milano, 2002; Ph.D. in nuclear engng., Politecnico di Milano, 2006) is assistant professor at the department of Energy at the Politecnico di Milano. His main research efforts are currently devoted to the development of methods and techniques for system health monitoring, fault diagnosis, prognosis and maintenance optimization. He is also interested in methodologies for rationally handling the uncertainty and ambiguity in the information. He is co-author of 28 papers on international journals and 32 on proceedings of international conferences.

Michele Compare (BS in mechanical engng., University of Naples Federico II, 2003) is a PhD student in nuclear engng. at the Politecnico di Milano. He has a two years' work experience as RAMS engineer in the aerospace industry and one year of experience in the medical industry as risk manager. His main research efforts are currently devoted to the development of methods and techniques in support of maintenance of complex systems.

Michele de Nigris received the Degree of Electrical engineering from the University of Genoa (Italy) in 1983. From 1984 to 2005 he worked with CESI SpA with growing responsibilities reaching the role of Head of the Electrical Laboratories and Components Business Unit. From 2006 onwards he is Director of T&D Technologies Department in CESI RICERCA, a public owned company carrying out R&D activities in the electroenergetic field. He is author or co-author of more than 70 scientific papers published at national and international level and has a number of assignments in IEEE, IEC, CIGRE, IEA etc.

Giuseppe Rizzi received the Doctor's Degree in Physics from the State University of Milan in 1974. In 1969 he joined CESI, where he started his activity in the field of High Voltage. His main research efforts have been focused on gas insulation and development of measuring techniques, accreditation activities of CESI as Accredited Calibration Laboratory for High Voltage Impulse Measurements according to IEC 60060-2. He is a member of CIGRE WGs 33.03 and 23/21/33-15 as well as senior member of IEEE, TC42 of CEI (Italian Electro-technical Commission), and is author of more than 50 technical papers.