# Feature Selection and Classification of Field Leakage Current Waveforms using Genetic Algorithms

**D. Pylarinos[1], K. Theofilatos[2], K. Siderakis[3], E. Pyrgioti[4], T. Papazoglou[5], I. Vitellas[6], E. Thalassinakis[7]**

**[1] Independent Scientific Consultant, [2,4]University of Patras, [3,5]Technological Educational Institute of Crete, [6,7] Public Power Corporation (PPC)**
**[1-7]Greece**

## SUMMARY

Leakage current monitoring is a commonly employed tool for the investigation of HV insulators' performance. Several techniques have been applied on in order to extract activity indicating information from leakage current waveforms. However, a fully representative value is yet to be defined. A recent approach to cope with this problem is to classify waveforms using a feature set comprising from commonly used values from time and frequency domain, and different feature selection and pattern recognition algorithms for the classification. In this paper, Genetic Algorithms are employed to perform both feature selection and classification. Eleven features were selected (out of the original twenty) and the accuracy achieved ranged from 82.2% to 88.48%, which is a considerable increase compared to earlier GA classification with no feature selection. Results are also compared to other feature selection and classification techniques that were previously applied to the same data set. Results emphasize the importance of feature selection and showed that Genetic Algorithms may give comparable results to nonlinear classifiers, provided that they are also employed for feature selection.

## KEYWORDS

leakage current waveform, insulator, classification, feature extraction, wavelet, Fourier, feature selection, pattern recognition, genetic algorithm

dpylarinos@yahoo.com

## Introduction

The performance of insulators is a matter of great concern for system operation, since a single insulator failure can result to an excessive outage of the power system, especially when the insulator is located in a HV station or substation. Outdoor insulators are greatly affected by local operation conditions, with pollution being probably the most significant factor [1-3]. Several standardized tests are employed in order to investigate insulators' performance in the lab, e.g. [4-6]. However, since insulators' performance is strongly correlated to environmental conditions, field testing is also employed with a guide for the establishment of high voltage insulator test stations being recently published [7].

Leakage current measurement is commonly employed as a tool to monitor and investigate the performance of HV insulators both in lab and field monitoring [8]. The basics stages of activity have been correlated with certain waveform shapes during lab tests [9-11] and an excessive investigation of field monitoring recently showed that the same shapes should be expected in the field [12-13], provided that field related noise has been removed [14-15]. Several techniques have been applied on leakage current waveforms in order to extract and record information regarding surface activity. In the field, the most commonly extracted values are the peak value, the charge and the number of pulses exceeding pre-defined thresholds, whereas the harmonic content is commonly investigated in lab measurements [8]. However, it is commonly accepted that it is the shape of the leakage current waveform that corresponds to surface activity, and a fully representative value of the waveforms' shape is yet to be defined.

Recently, a new approach has been proposed for the classification of leakage current waveforms [12, 16]. According to this approach, twenty different features are extracted from the leakage current waveform to create a pattern. The features used, equally represent the time and the frequency domain (ten features from each domain) and are commonly used features [8, 12, 16]. Then, classification techniques were used to classify each waveform in two different classes depending on the duration of discharges. At first a linear classification was attempted employing an Euclidian classifier and a simple Genetic Algorithms (GAs) approach, and results were not that encouraging [16]. This was attributed to the most probably non-linearity of the problem and the absence of an effective feature selection scheme. Then, non-linear classification techniques were employed, including three different classification algorithms (knn, Naïve Bayes, SVMs) and two feature extraction techniques (student's t-test and mRMR) [12]. The classification algorithms were employed on five feature sets (time domain features, frequency domain features, time and frequency domain features, student's t-test selected features and mRMR selected features). Results showed the superior performance of SVMs and of the feature set provided by the mRMR algorithm. Results also indicated that feature selection is rather significant for the success of the classification. However, feature selection was not employed when the GAs were used [16]. Therefore, in this paper, a linear classification methodology based on GAs is developed which incorporates a feature subset optimization procedure. New results are provided for the same set of waveforms for consideration and comparison. Experimental results suggested that even linear classifiers when combined with feature selection schemes can provide effective classification of the examined waveforms.

## Set-up and Measurement Sites

The work presented in this paper is part of a large project of the Greek Public Power Corporation, the Technical Educational Institute of Crete and the University of Patras for the investigation of the performance of high voltage insulators. The final step of the project is the construction and operation of TALOS High Voltage Test Station in Iraklion, Crete [17-19]. The Cretan Transmission Network is exposed to intense marine pollution and several techniques have been employed by the Greek Public Power Corporation to cope with the problem [20-22]

Eighteen different 150 kV post insulators (porcelain, RTV SIR coated and composite) have been monitored than more than six years. The insulators were part of the grid, installed at 150 kV Substations. A collection ring was installed at the bottom side of each monitored insulator and the current was driven through a Hall current sensor in order to acquire the measurement. The acquired data was then transmitted to a commercially available Data Acquisition system (DAQ). Detailed

specifications of the DAQ can be found in [13]. A schematic representation and pictures from the measuring system are shown in Figure 1.

Sampling was performed continuously and simultaneously for all monitored insulators, at a rate of 2 kHz and resolution of 12bit. Each waveform recorded has a duration of 480 ms. The monitoring system incorporated the time-window technique [14-15] to record waveforms. The waveform portraying the highest peak value in the considered time window is recorded. The time-window is user defined (from 6 hours to 24 hours), with a 24 hours window being mostly employed due to hardware restrictions.

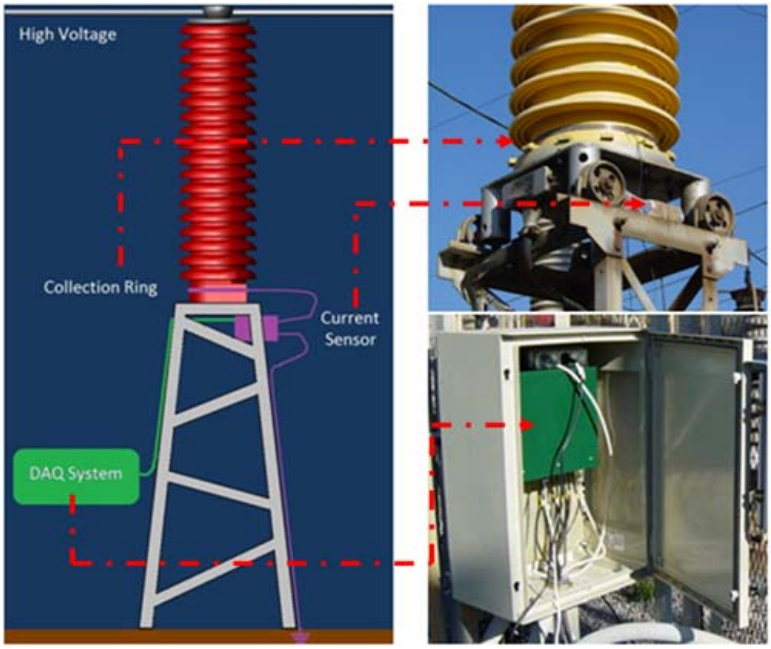


Figure 1. A schematic representation and pictures of the measuring system showing

**Extracted Features and Waveform Data Set**

The noise reduction/removal techniques described in [14-15] have been applied in order to remove noise related waveforms. The $S_R$ ratio [12, 15] has also been used in order to remove isolated spikes and the D3/D5 ratio derived from wavelet analysis has also been used to remove sinusoids waveforms [12]. The finally considered waveforms have been hand-picked and classified in two different classes, depending on the duration of discharges. Class C1 includes waveforms that portray discharges that last four halfcycles or less, whereas class C2 includes waveforms that portray discharges that last five or more halfcycles. A set of 20 features are employed as a pattern. The features can be seen in Table I. Features 1-10 derive from the time domain, and features 11-20 from the frequency domain, and have been chosen in order to evenly represent both domains and also considering the literature [8].

Regarding the time domain features, frequently used values such as the amplitude and charge, along with commonly used statistical values are employed. The charge is calculated from the waveform considering the digitization process and the applied sampling rate. Regarding the feature domain features, it was considered that the content of odd harmonics is commonly correlated to the occurrence of discharges and the distortion of the waveforms' shape and therefore several commonly used ratios of odd harmonics [8] are used. It should be noted that the fundamental frequency is 50Hz and that the HD ratio is similar to the THD ratio, with the numerator being the sum of the odd harmonics' content.

Further, wavelet analysis and especially MRA is employed in order to acquire the STD_MRA VECTOR [12, 15, 24, 25]. The STD_MRA VECTOR contains the standard deviation (STD) of the details of each level of the wavelet multiresolution analysis (MRA) of the original waveform, with $D_1$ referring to the first decomposition level, $D_2$ to the second level etc [12, 15]. The distortion ratio [8]

given by: $D_R = \frac{D_1 + D_2 + D_3 + D_4}{D_5}$, is also considered. The frequency bands of the STD_MRA VECTOR's components are shown in Table II.

TABLE I – Feature Set

| No. | Feature | No. | Feature |
|---|---|---|---|
| **1** | Amplitude | **11** | Third to First Harmonic Ratio |
| **2** | Mean | **12** | Fifth to First Harmonic Ratio |
| **3** | Median | **13** | Fifth to Third Harmonic Ratio |
| **4** | Variance | **14** | Total Harmonic Distortion Ratio (THD) |
| **5** | Standard Deviation | **15** | Harmonic Distortion Ratio (HD) |
| **6** | Median Absolute Deviation | **16** | STD_MRA VECTOR Ratio: D1/D5 |
| **7** | Skewness | **17** | STD_MRA VECTOR Ratio: D2/D5 |
| **8** | Kurtosis | **18** | STD_MRA VECTOR Ratio: D3/D5 |
| **9** | Interquartile Range | **19** | STD_MRA VECTOR Ratio: D4/D5 |
| **10** | Charge | **20** | Distortion Ratio: DR |

TABLE II –Frequency Bands of MRA

| Decomposition Level | (A) Approximation (Hz) | (D) Details (Hz) |
|---|---|---|
| 1 | 0~500 | 500~1000 |
| 2 | 0~250 | 250~500 |
| 3 | 0~125 | 125~250 |
| 4 | 0~62.5 | 62.5~125 |
| 5 | 0~31.25 | 31.25~62.5 |
| 6 | 0~15.625 | 15.625~31.25 |

**Genetic Algorithms**

Genetic Algorithms (GAs) are general search meta-heuristic algorithms based on the initial creation of a population of candidate solutions, called chromosomes, and their iterative differentiation using the operators of evaluation, selection, crossover and mutation until some termination criteria are reached [26]. GAs have been proved useful and efficient in optimization problems where the search space is big and complicated or there is not any available mathematical analysis of the problem.

Their function is based on the initial creation of a population of candidate solutions, called chromosomes, and their iterative differentiation using the operators of evaluation, selection, crossover and mutation until some termination criteria are reached. Evaluation is responsible for evaluating the candidate solution using a problem specific fitness function. The various selection operators which have been proposed in the literature are used to enforce the search process to emphasize in the most promising areas of the search space. Crossover and mutation operators are used to produce new, probably better, solutions using the population of the candidate solutions. The crossover operator is responsible for the recombination of existing solutions in the population whereas the mutation operator is responsible for the random differentiation of existing solutions. The probabilities, under which the mutation operator is applied, control the search behavior of the produced algorithm. For example, high mutation probability may degenerate the algorithm to a random search whereas a very small mutation rate may lead the algorithm in getting trapped in local optima.

The ability to define a different fitness function for each problem, and the flexibility of GAs to function deploying various representations of candidate solutions have enabled their extensive use in a variety of problems. In the present paper, we applied GAs for classifying waveforms in two categories and on parallel for detecting the optimal feature subset which should be used as input for the final

classifiers. Specifically, the chromosome (Figure 2) of the proposed GA consist of 20 genes to represent the first classification center, 20 genes to represent the second classification center and 20 genes to determine which features should be used as inputs for the classifier. The feature selection genes take values in the interval [0,1] and force the classifier to use a specific feature as input if each value is higher than 0.5. Our approach attempts to classify the two classes of waveforms linearly using a set of 20 features as candidate inputs. In order to achieve this linear classification, the two centers of the two classes should be determined. The positions of these two centers are the first 40 genes of each chromosome of the GA. After the estimation of these centers and the optimization of the feature subset are achieved, each waveform is classified in the class for which it has the minimum Euclidean distance from its centre. Our problem thus, is formulated as finding the optimal centers of each class and the optimal feature subset which should be used as input.

| Classification Center 1 Genes (20 genes) | | | | | Classification Center 2 Genes (20 genes) | | | | | Feature Selection Genes (20 genes) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| x1 | x2 | x3 | ... | x20 | y1 | y2 | y3 | ... | y20 | F1 | F2 | F3 | ... | F20 |

Figure 2. GA's chromosome structure

A simple GA is employed to solve this optimization problem [27]. The data set gets normalized and one thousand generations are employed with a crossover probability of 90%. The crossover operator which was used is the two points crossover operator, depicted in Figure 3. This operator is considered superior to the one point crossover operator as it exchanges fewer genes from each parent and thus the probability of being destructive on the performance of the chromosomes is smaller. As for the mutation operator, the Gaussian mutation operator is applied because of the GA's genes being decimal variables. This operator adds in each gene which is selected for mutation a value taken from the Gaussian Distribution with center equal to zero and variance equal to 0.1.


Figure 3. Two points crossover operator

The fitness function which was used to measure the performance of each individual is the one described in equation 1:

$$Fitness = Accuracy + Geometric\ Mean \quad (1)$$

where the *Accuracy* term is the classification accuracy which is derived when the centers and the feature subset of each candidate solutions are used. *Geometric Mean* is equal to the square root of the product of the sensitivity and the specificity of the derived classifiers. This term is used to force our classifier towards balanced classification solutions which perform equally well for both classes.

To optimize the initial population which should be used in the proposed GA we experimented using various values and their results are shown in Table III (mutation probability of 90% was applied). Then, for the initial population showing the best results, different mutation probabilities are employed as shown in Table IV. Fifty present of the data was used as a training set and fifty present as a test set. The mean value of the results of five runs is presented in each table. The results provided in Table III and Table IV show the performances of the proposed GA approach in the test set.

From the tables it is observed that the best mean accuracy which was achieved was for Population Size equal to 100 and Mutation Probability equal to 10%. The best iteration of the algorithm returned a linear classifier with accuracy 88.48% and a feature subset of eleven features. The eleven features which were selected are: {2, 3, 5, 6, 7, 9, 11, 12, 15, 18, 19}.

Table III: Experimental results for various population sizes

| execution no. | population=20 | population=60 | population=100 | population=140 |
|---|---|---|---|---|
| 1 | 86.39% | 86.91% | 88.48% | 84.29% |
| 2 | 82.20% | 85.34% | 86.91% | 86.39% |
| 3 | 87.96% | 85.86% | 85.86% | 82.72% |
| 4 | 82.72% | 85.86% | 84.82% | 85.86% |
| 5 | 86.91% | 84.82% | 86.91% | 88.48% |
| Mean | 85.24% | 85.76% | 86.60% | 85.55% |

Table IV. Experimental results for various mutation rates

| execution no. | mutation_rate=10% | mutation_rate=5% | mutation_rate=20% |
|---|---|---|---|
| 1 | 88.48% | 86.91% | 87.43% |
| 2 | 86.91% | 86.39% | 82.72% |
| 3 | 85.86% | 83.77% | 84.29% |
| 4 | 84.82% | 85.86% | 86.39% |
| 5 | 86.91% | 88.48% | 82.20% |
| Mean | 86.60% | 86.28% | 84.61% |

**Discussion**

The best accuracy achieved using GAs in a single run was 88.48% whereas the best mean was 86.6%. All accuracy percentages in every run were over 82% and usually around 85%. This is a significant increase compared to the previously GA classification, which had a maximum accuracy percentage of 56.12% [16]. Further, this means that the GA approach employed in this paper provided results comparable to the non linear classifiers employed in [12] as shown in Table V. Further, the number of features selected using the GAs is similar to the number of features selected using the mRMR algorithm (11 to 10), and the GA classification using this set provides similar results with the non-linear classifiers (that ranged from 85.74% to 90.21% for the mRMR feature set).

Table V. Previous results from nonlinear classifiers [12]

| ***Features*** | ***TD*** ***{1~10}*** | ***FD*** ***{11~20}*** | ***All*** ***{1~20}*** | ***t-test*** ***{1, 3~11, 13~17, 19~20}*** | ***mRMR*** ***{3, 5, 7, 8, 11, 13, 15, 16, 18, 19}*** |
|---|---|---|---|---|---|
| ***knn*** | 82.13% | 86.77% | 85.22% | 83.85% | 85.74% |
| ***Naïve Bayes*** | 69.41% | 77.66% | 73.02% | 73.88% | 86.43% |
| ***SVMs*** | 82.48% | 88.49% | 87.80% | 87.80% | 90.21% |

**Conclusion**

Leakage current monitoring is commonly employed as a tool for the investigation of the performance of High Voltage insulators. Several techniques have been applied on raw waveform measurements in order to extract and record information regarding surface activity. However, it is the shape of the leakage current waveform that portrays an exact image of the experienced activity. A

recent approach to cope with this problem is to classify waveforms based on the duration of discharges. For the classification, a 20 feature pattern is extracted from each waveform. The features used, equally represent the time and the frequency domain Then, different feature selection and classification algorithms may be used to perform the classification. In this paper, Genetic Algorithms are employed to perform both feature selection and classification. Results are compared to other feature selection and classification techniques that were applied to the same data set and also to an earlier GA classification with no feature selection. Results show that the GA approach, even though a linear one, may give comparable results to previously used nonlinear classifiers, provided that GA feature selection is also employed.

## BIBLIOGRAPHY


[1] CIGRE WG 33-04, TF 01, A review of current knowledge: polluted insulators, CIGRE, 1998

[2] IEC/TS 60815, Selection and dimensioning of high-voltage insulators intended for use in polluted conditions, 2008

[3] CIGRE WG 33-04, "The measurement of site pollution severity and its application to insulator dimensioning for a.c. systems", Electra Vol. 64, p. 101-116, Cigre, 1979

[4] IEC 60507, Artificial pollution tests on high-voltage insulators to be used on a.c. systems, 1991

[5] IEC 60587, Electrical insulating materials used under severe ambient conditions-Test methods for evaluating resistance to tracking and erosion, 2007)

[6] IEC 62217, Polymeric insulators for indoor and outdoor use with a nominal vlotage > 1000 V – General definitions, test methods and acceptance criteria, 2005

[7] CIGRE WG B2.03, Guide for the establishment of naturally polluted insulator testing stations, (CIGRE, 2007)

[8] D. Pylarinos, K. Siderakis, E. Pyrgioti "Measuring and analyzing leakage current for outdoor insulators and specimens", Reviews on Advanced Materials Science, Vol. 29, No. 1, pp. 31-53, 2011

[9] M. A. R. M. Fernando, S. M. Gubanski, "Leakage current patterns on contaminated polymeric surfaces", IEEE Trans. Dielectr. Electr. Insul., Vol. 6, No. 5, pp. 688–694, 1999

[10] T. Suda, "Frequency characteristics of leakage current waveforms of an artificially polluted suspension insulator", IEEE Trans. Dielectr. Electr. Insul., Vol. 8, pp. 705–709, 2001.

[11] J. Li, W. Sima, C. Sun, S. A. Sebo, "Use of Leakage Current of Insulators to Determine the Stage Characteristics of the Flashover Process and contamination Level Predicition", IEEE Trans. Dielectr. Electr. Insul., Vol. 17, No. 2, 2010

[12] D. Pylarinos, K. Theofilatos, K. Siderakis, E. Thalassinakis, I. Vitellas, A. T. Alexandridis, E. Pyrgioti, "Investigation and Classification of Field Leakage Current Waveforms", IEEE Transactions on Dielectrics and Electrical Insulation, Vol. 19, No. 6, pp. 2111-2118, 2012

[13] D. Pylarinos, K. Siderakis, E. Thalassinakis, E. Pyrgioti, I. Vitellas, "Investigation of leakage current waveforms recorded in a coastal high voltage substation", Eng. Technol. Appl. Sci. Res., Vol. 1, No. 3, pp. 63-69, 2011

[14] D. Pylarinos, K. Siderakis, E. Thalassinakis, E. Pyrgioti, I. Vitellas, S. L. David, "Online applicable techniques to evaluate field leakage current waveforms", Electr. Power Syst. Res., Vol. 84, No. 1, pp. 65-71, 2012

[15] D. Pylarinos, K. Siderakis, E. Pyrgioti, E. Thalassinakis, I. Vitellas, "Impact of noise related waveforms on long term field leakage current measurements", IEEE Trans. Dielectr. Electr. Insul., Vol. 18, No. 1, pp. 122-129, 2011

[16] D. Pylarinos, K. Theofilatos, K. Siderakis, E. Pyrgioti, T. Papazoglou, I. Vitellas, E. Thalassinakis, "Classification of Field Leakage Current Waveforms using Genetic Algorithms and an Euclidian Classifier", DEMSEE 7th International Workshop on Deregulated Electricity Market Issues in South-Eastern Europe, 2012

[17] TALOS High Voltage Test Station, www.talos-ts.com

[18] D. Pylarinos, K. Siderakis, E. Thalassinakis, I. Vitellas, E. Pyrgioti, "Recording and managing field leakage current waveforms in Crete. Installation, measurement, software development and

signal processing", ISAP 16th International Conference on Intelligent System Applications to Power Systems, Hersonissos, Crete, Greece, September 25-28, 2011
[19] "Greek utility readies to energize new insulator Test Station", INMR, Issue 82, Vol. 16, No. 4 p. 32, 2008, (available at: http://www.inmr.com/archive/U82032.html)
[20] E. Thalassinakis, K. Siderakis, D. Agoris, "Experience with New Solutions to Combat Marine Pollution in the Power System of the Greek Islands", INMR 2003, World Conference & Exhibition on Insulators, Arresters & Bushings, Malaga, Spain, November 16-19, 2003 (available at: http://www.inmr.com/2003WorldCongress/2003INMRWorldCongress/pap/Thala_paper.html)
[21] "Greek Utility Battles Pollution Affecting Island Transmission System", INMR, Issue 78, Vol. 15, No. 4, p. 24, 2009 (available at http://www.inmr.com/archive/U78024.html)
[22] K. Siderakis, D. Pylarinos, E. Thalassinakis, E. Pyrgioti, I. Vitellas, "Pollution maintenance techniques in coastal high voltage installations", Engineering, Technology & Applied Science Research, Vol. 1, No. 1, pp. 1-7, 2011
[23] K. Siderakis, D. Pylarinos, E. Thalassinakis, I. Vitellas, "High voltage substation pollution maintenance: the use of RTV silicone rubber coatings", Journal of Electrical Engineering, Vol. 11, No. 2, Article 11.2.22, pp. 1-6, 2011
[24] S. G. Mallat, A Wavelet Tour Of Signal Processing, Academic Press, 1999.
[25] S. G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 11, pp. 674-693, 1989.
[26] J. Holland, Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence, Cambridge: Mass ;MIT Press, 1995.
[27] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag, 3rd edition, 1996.