Simulated Partial Discharge Harmonic Data generation for Neural Network Training in the Absence of real measurements: A first Approach

Dimitrios A. Barkas
Department of Electrical and Electronics Engineering
University of West Attica
Egaleo
d.barkas@uniwa.gr

Konstantinos Kalkanis
Department of Electrical and Electronics Engineering
University of West Attica
Egaleo
k.kalkanis@uniwa.gr

George Ch. Ioannidis
Department of Electrical and Electronics Engineering
University of West Attica
Egaleo
gioan@uniwa.gr

Stavros D. Kaminaris
Department of Electrical and Electronics Engineering
University of West Attica
Egaleo
skamin@uniwa.gr

Constantinos S. Psomopoulos
Department of Electrical and Electronics Engineering
University of West Attica
Egaleo
cpsomop@uniwa.gr

***Abstract*— Partial Discharges measurements on the High Voltage equipment are assumed as one of the most critical condition assessment measurements which can propose a future lifetime of power equipment and a suitable maintenance schedule. Neural Networks are assumed as one of the most accepted Artificial Intelligence techniques for the condition assessment of High Voltage power equipment. However, the correct use of this technique demands the existence of a large number of datasets to provide valuable results. Many times the required large datasets in specific investigation areas such as the partial discharges on the high voltage electricity network do not exist for several reasons. These reasons are described in this manuscript. In the absence of these datasets, which are typically real measurements, there is a specific need to construct them by using small datasets on the field under interest. This manuscript describes such a data construction algorithm based on the authors' knowledge and experience in the field of partial discharge measurements.**

Keywords—Partial Discharges, Frequency Components, Neural Network, Artificial Intelligence

I. INTRODUCTION

The electricity network reliability is of great importance for the provision of reliable and stable electric power. The electricity network is composed of three main areas according to the voltage level, the distribution network that operates at 150kV and 400kV for the Greek electricity network, the medium voltage network at 22kV and the low voltage network with a nominal voltage of 400V. The integrity of the electricity network is finally translated to the integrity of the insulation. Good insulation results in an electricity network with high integrity and a low probability of failure. Because the insulation is of great importance, its condition assessment is considered critical. An important notice is the dependence of the insulation characteristics on the operating voltage level. The equipment which operates on the high voltage level is more prone to malfunction due to the higher electric strength of the insulation. Insulation includes several operating parts of the network with the most usual and at the same time critical being the solid insulators (for example the bushings of power transformers), oil insulation and gas insulation. The condition assessment of the insulation is based on several measurements that can be applied to High Voltage (HV) equipment. Many proposed measurements have been used such as the delta tangent of an insulator (well known as dissipation factor or loss angle), the dielectric strength of the insulator in kV, and the Sweep Frequency Response Analysis (SFRA), the Partial Discharges and the chemical analysis of the insulating oils [1, 2, 3]. The chemical analysis of insulating oils is a very useful method for their condition assessment . Chemical analysis measures the existence of seven specific chemical components in the oil (hydrogen, carbon monoxide, carbon dioxide, methane, ethane, ethylene and acetylene) [4]. The existence of these specific chemical components is then analyzed by representation models, which are well known as Dissolved Gas Analysis (DGA) models [5, 6]. Many operators including the Independent Power Transmission Operator (IPTO) in Greece analyze insulating oil via their equipment such as oil-immersed power transformers and switches. However, the DGA must not be used as a specific problem decision technique but only as an indicator and complementary measurement. The combination of the DGA with other measurements can overcome meaningless decision situations [7, 8].

A supplementary measurement could be the measurement and the analysis of the electric voltage waveform with the main aim to detect the spectrum of the measured signal. In an ideal world, the alternating electric voltage waveform would be a sinusoidal signal with a frequency of 50Hz in Greece and Europe (or 60Hz in the USA, Canada etc). However, the existence of non-linear electric loads such as capacitive and inductive loads and the use of power electronics have added more frequency components to the waveform. Moreover, the existence of Partial Discharges (PDs) creates frequency components higher than the fundamental frequency of 50Hz. PDs are a special phenomenon which is presented in the insulations of high voltage equipment. These discharges originate from many reasons but the most significant is the local increased value of the electric field. The environmental conditions also affect the severity of this type of discharge.

979-8-3503-9958-5/22/$31.00 ©2022 European Union

The probability of insulation failure is increased when these discharges become severe.

The knowledge of the frequency components which are generated due to the PDs can be used for the extraction of valuable condition assessment of HV equipment. The classification of different types of PDs by harmonic orders has already been done and the results are presented in Table I [9]. There are three basic types of PDs according to the location of the insulation where the discharge is carried out, corona discharges, internal discharges and surface discharges. The corona discharges are usually presented near curved regions such as the cable connections to the chain insulators on transfer and distribution pylons, at the terminals on the power transformers, and on the transmission lines. The internal discharges are created inside the insulation, for example in cavities of solid insulators and bubbles in the insulating oils. The surface discharges are usually presented on the surface of the solid insulators. With knowledge of the type of the discharge, engineers can compute the expected lifetime of equipment (as the lifetime of the equipment is affected by the insulation's lifetime), as well as they can plan the maintenance procedures with relevance to processes and spare parts.

TABLE I. HARMONIC ORDERS ON THE VOLTAGE WAVEFORM CLASSIFYING THE DIFFERENT TYPE OF PDS [9]

Discharge Type	Harmonic Orders				
Corona Discharge	5^{th}: 1.1%	7^{th}: 1.5%	9^{th}: 0.9%	13^{th}: 1.05%	
Internal Discharge	5^{th}: 1.05%	7^{th}: 1.25%	9^{th}:0.95%	13^{th}: 0.55%	
Surface Discharge	2^{nd}: 0.85%	3^{rd}: 0.4%	5^{th}: 1%	7^{th}: 0.95%	9^{th}: 0.90%

The absence of frequency analysis of the high voltage waveform, connected to a specific type of discharge, and therefore the absence of frequency datasets constitutes a significant problem for the development of modern maintenance schedule processes. To solve this problem, an algorithm for the creation of large datasets of frequency components is presented, which will be fully connected to specific PD types. Some of these data will be used to train a Neural Network (NN) while the rest will be used for evaluation. According to Table I, there are six inputs (2nd , 3rd, 5th, 7th, 9th and 13th order harmonics corresponding to fundamental) and three outputs (corona discharge, internal discharge and surface discharge). The selected NN is constructed with two hidden layers. The first hidden layer has six neurons (the number of inputs), while the second hidden layer has three neurons (the number of outputs of the NN). Fig. 1 presents the NN architecture. The authors, owing to their experience and knowledge regarding the behaviour of PDs under realistic situations, real laboratory measurements and observations, concluded that the under investigation frequency components can change up to ±20% due to several reasons, with the most critical being:

- Measuring Error
- Unstable electric and voltage rms value
- Ambient conditions (humidity, temperature, surface dust etc.)

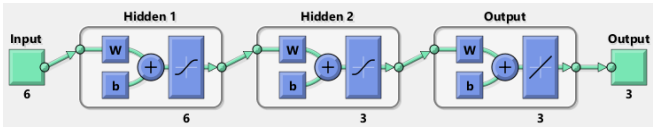

Fig. 1. The selected NN with two hidden layers of six and three neurons respectively

II. CREATION OF DATASET

A Neural Network is an Artificial Intelligence technique with the ability to compute the probability of a situation to be achieved, based on previous situations. The past known situations, which are described in the form of datasets "Inputs - Outputs", are used to train the NN to acquire its knowledge. For this reason, the existence of large datasets in the form of Inputs – Outputs is of great importance [10]. However, many times these large datasets are not available for several reasons, such as in the case of PDs. The difficulty of the existence of the dataset in the case of PDs sourcing from the past lack of exploration of PDs and their connection to specific frequency patterns. The electricity network operators did not and even now do not measure and analyze the high voltage waveforms merely because they ignore that these frequency patterns exist. The need for high voltage measurement and analysis and the connection of the three types of aforementioned PDs with specific frequency components may take a significant time interval. For this reason, until real large PDs frequency patterns are available, it is important to start by theoretically creating these datasets. These will be used to train the NN architecture which is described in Fig. 1. The dataset creation is comprised of three stages:

- Stage 1: Initial dataset creation and NN training and evaluation
- Stage 2: Second dataset creation and testing the NN
- Stage 3: Third dataset creation and testing the NN

A. Stage 1 - Creation of the 1st dataset packet

For the creation of datasets, their structure must be explained. The datasets as already mentioned should be a pair of Inputs and Outputs. For each dataset, there is an input vector of six components and an output vector of three percentages. The six components of the input vector are the six values of the frequency components in percentage. The output vector includes three probability values. The first value refers to the "Corona Discharge" case, the second value to the "Internal Discharge", and the third value to the "Surface Discharge". The higher value from the three probabilities characterizes the resulting discharge type. The initial stage embeds the initial datasets creation used for the training and testing of the selected NN architecture, as well as the NN architecture which must be constructed, tested and evaluated. The NN structure uses as transfer function the "Hyperbolic Tangent Sigmoid" function, as training algorithm the "Levenberg - Marquardt" backpropagation algorithm. Fig. 2 explains the algorithm part for this initial stage. The implementation has been developed through MatLab code. In the first step of this stage, the code reads the initial data measurements (Table I) and creates two matrices. The "Total Data Matrix" containing the constructed input datasets, and the "Target Data Matrix" containing the target value for each input dataset. The way that datasets are created is based on the binomial coefficient (usually it is referred to as "n choose k").

The next step is the initialization of the NN, which contains two hidden layers. The first layer is constructed by six neurons, while the second hidden layer includes three neurons. The initialization process is followed by the training and testing processes of the network. The code finally exports the NN weights through a suitable MatLab matrix and creates the confusion matrix. The confusion matrix describes the efficiency of the NN after the testing process.

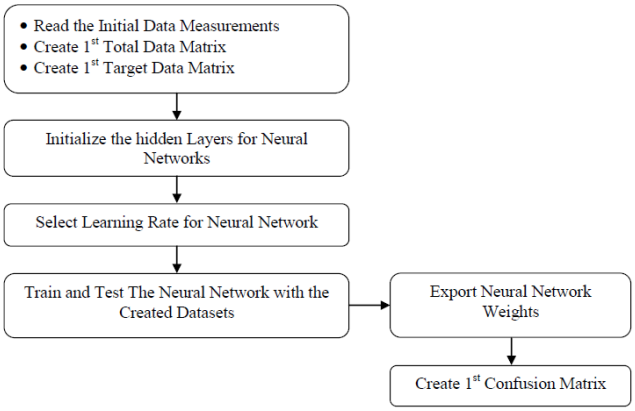

Fig. 2. The basic steps for the creation of the initial datasets and the construction of the NN

B. *Algorithm for the creation of datasets*

As already mentioned, the idea behind the algorithm is the production of new PD cases which can be recognized through the frequency analysis over the voltage signal. This dataset production is based on the binomial coefficient by taking into consideration the authors' experience in the variation of the real measurements.

In the first case of evaluation, 126 new datasets for each case of PD were created, which means that a total number of 378 new datasets were created. These numbers can be exported following the next way of thinking. The input vector includes six input values for the NN. The question is how many combinations can be achieved for each PD case by changing the value of only one, or only two, or only three, and so on, of the six values of the initial input vector. The change of the value will always be by 20% of the initial values (Table I) for this first stage. The binomial coefficient gives this answer through the following formula:

$$n'_{s1} = \sum_{k=1}^{6} \binom{k}{6} = 63 \qquad (1)$$

However, the code takes into consideration that the change in value can be +20% or -20%. By this, the above number must be doubled. Additionally, as there are three different PD cases, the number must also be tripled. So the final number of new datasets for the first stage of the code will be:

$$n_{s1}=378$$

The algorithm that calculates the combinations creates a matrix with a corresponding number of unity elements, and then for each selected group of values that are to be changed in the matrix, it increases or decreases the selected group by 20%. By this logic the final matrix is a matrix that contains two tables, D1 and D2:

$$D=[D1\ D2]^T$$

The D1 submatrix is the matrix with the groups of those values that will be increased by 20%, while the D2 submatrix is the matrix with the groups of values that will be decreased by 20%. The D table contains numbers which are either "1" or "1.2" or "0.8". Element wise multiplying the D matrix with the corona row vector measurement, internal row vector measurement or surface row vector measurement (Table I), the new values with variations from the initial values can be created. The total created data are contained in a table named totalData:

totalData = [corona new_Corona internal new_Internal surface new_Surface]^T

Where corona, internal, and surface row vectors are the initial measurement vectors, while the new_Corona, new_Internal and new_Surface matrices are the new created elements.

From the total data matrix, 65% of the elements are used for the training process of the neural network, while 35% are selected for the testing process. The confusion matrix for the first stage of code, which embeds the construction of NN, the training and the testing processes, is presented below in Fig. 3. The confusion matrix shows that the results agree 100% with the NN decision.

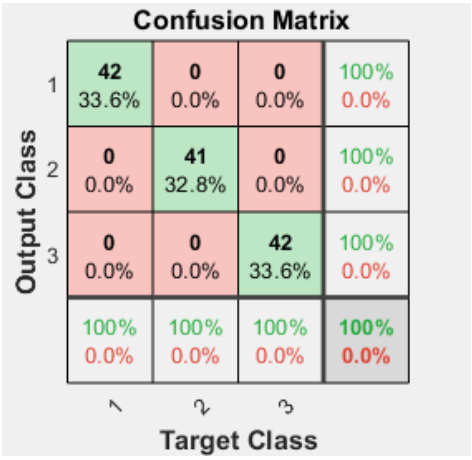

Fig. 3. Testing of NN for the 1st Stage of code

III. CREATION OF TESTING DATASETS

A. *Stage 2 - Creation of the second testing dataset*

In the second stage, the new datasets are created in a manner like in the initial training and testing data, using a binomial coefficient. Specifically, the code computes matrices like the matrix D that was already mentioned but with a significant difference. When a group of components are selected for changing their values, each component alters its value in a different percentage. For example, assume the case of initial measurement row vector "corona" with elements [0.001% 0.001% 1.1% 1.5% 0.9% 1.05%], and the components that are to be changed are the third and fourth one

(1.1% and 1.5%). Assume also that the alternation factor is m=5%. The third element will take a value of 1.05 times 1.1% (0.01155 or 1.55%), while the fourth element will take a value of 1.10 times 1.5% (0.0165 or 1.65%), leading to a different value variation for the same dataset. This example can easily be explained in matrix format. The example assumes that from the six values of the dataset only a group of two can be changed and also that these are the third and fourth. The combination row vector for this case will be:

B = [1 1 **1.05 1.10** 1 1]

The element wise multiplication of B row vector of combination matrix with the "corona" vector will be:

C[i] = B[i] x corona[i]

Where i = 1 to 6

The difference between the third and fourth elements of the combination matrix is m. Generally, if the selected group of values to be changed contains "g" elements, then the first selected element will change by m, the second by 2m and so on.

It must be noted that in this stage of dataset creation, all the components in the group of change will take an increment or a decrement. That means that if one component of the selected group is to be increased, then all the other components in this group must also be increased. The maximum variation will be 30% which means that in the worst case of a selected group of 6 values the B matrix will be:

B = [1.05 1.10 1.15 1.20 1.25 1.30] (for increment in the values) or

B = [0.95 0.9 0.85 0.8 0.75 0.7] (for decrement in the values)

The new datasets are calculated by multiplying the B vector with the initial measurements (Table 1). The corresponding target data were also calculated, and the data was entered in the trained NN. The confusion matrix is presented below in Fig. 4 and shows that the previously trained NN can identify 100% percent of the different PD cases.

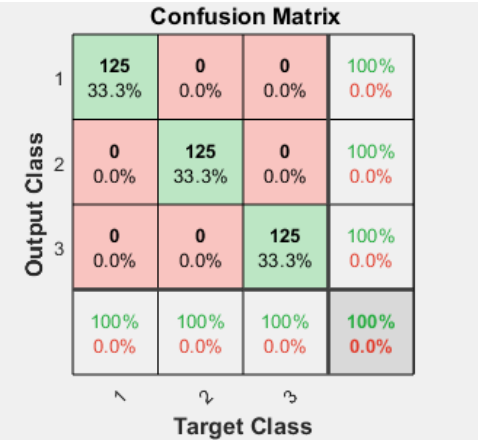

Fig. 4. Confusion matrix for the 2nd stage of testing of NN

B. Stage 3 - Creation of the third testing dataset

In the second case, the values are all increased or decreased for a selected group of a dataset. In the third and last case, for the selected group of a dataset some values are increased while the rest are decreased. The algorithm initially selects the group of components that are to be changed. The algorithm increases the components of the odd position by m and decreases the values of the components that are located in the even position by m. For example, assume the case that a group of four components is selected. Moreover, assume that the components of the group that must change are the first, the second, the third and the fourth of the initial surface PD vector. The initial surface PD vector, which is included in Table I is:

surface = [0.85/100 0.4/100 1/100 0.95/100 0.90/100 0.001/100]

For this case example, the combination row vector will be:

B = [**1.05 0.95 1.10 0.9** 1 1]

Remember that the code at the beginning initializes the combination matrix with ones. As the B vector above illustrates, the first element of the selected group has increased by 5%, the second one has decreased by 5%, the third has increased by 10%, and the fourth has decreased by 10%. The selected alternation factor is m=5% which means that in the worst case for this "m" value, the combination row vectors will be:

B = [1.05 0.90 1.15 0.80 1.25 0.70] or

B = [0.95 1.10 0.85 1.20 0.75 1.30]

The new datasets are calculated by multiplying the B vector with the initial measurements (Table I). The corresponding target data were also calculated, and the data was entered in the trained NN. The confusion matrix is presented below in Fig. 5. Fig. 5 shows that the previously trained NN can identify with 97.9% percentage of the different PD cases, in the case where the alternation factor is 5%, and the worst value has been altered by 30%.

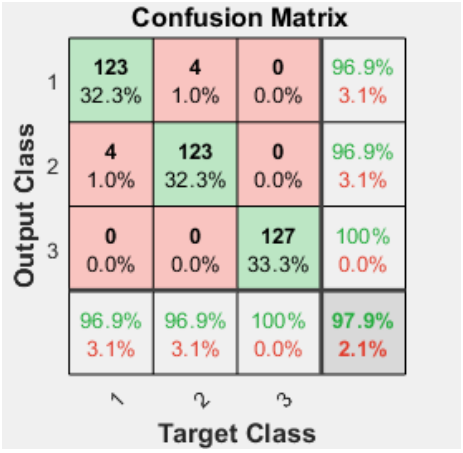

Fig. 5. Confusion matrix for the 3rd stage of testing of NN

This result describes the situation that as the changes in the harmonic distortion of the signal into the interested frequency band become more complicated, the difficulty in the correct recognition of the PD source is increased.

IV. CONCLUSION

The absence of real measurements in specific applications when decision-making applications are required is a great thorn in the science of engineering. There are critical infrastructures that must be monitored, such as the electricity transmission and distribution network, and their health is of great importance for human survival. This paper describes a simple but yet important algorithm, for the calculation of new datasets of specific harmonic components on the electric voltage waveform, according to which the partial discharge categorization can arise. The algorithm begins with the calculation of a starting dataset which is split into training and testing sub-datasets. Then two new datasets are created for the testing of the pretrained NN. These new datasets were created by utilizing the measurement theory of partial discharges and their variation into a band of ±20% relating to their initial measurement values. The testing process showed that the NN achieves high accuracy results. The algorithm offers significant help to the scheduled maintenance of the high voltage equipment which is exposed on PDs and can stand as the first approach in this field. This algorithm could be improved in the future by making more measurements in the field or the lab and describing the statistic that the PD frequency components on the voltage waveform follow. However, the reinforcement of the NN with real measurements from the field will lead to better neural networks and better-scheduled maintenance processes.

REFERENCES

[1] J.S. N'cho, I. Fofana, Y. Hadjadj, A. Beroual, "Review of physicochemical-based diagnostic techniques for assessing insulation condition in aged transformers," Energies 2016, 9, 367, https://doi.org/10.3390/en9050367

[2] T. Toudja, A. Nacer, H. Moulai, I. Khelfane, and A. Debche, "physico-chemical properties of transformer mineral oils submitted to moisture and electrical discharges," International Conference on Renewable Energies and Power Quality (ICREPQ'12), Santiago de Compostela (Spain)

[3] M. Bagheri, and B.T. Phung, "Frequency response and vibration analysis in transformer winding turn-to-turn fault recognition," International Conference on Smart Green Technology in Electrical and Information Systems (ICSGTEIS), pp. 10-15, doi: 10.1109/ICSGTEIS.2016.7885758

[4] W. Chen, X. Chen, S. Peng, and J. Li, "Canonical correlation between partial discharges and gas formation in transformer oil paper insulation," Energies 5(4):1081-1097. https://doi.org/10.3390/en5041081

[5] U.S. Department of the Interior Bureau of Reclamation, "Transformers: basics, maintenance, and diagnostics, " U.S. Department of the Interior Bureau of Reclamation, Denver, Colorado

[6] D.V.S.S Siva Sarma, and G.N.S Kalyani, "ANN approach for condition monitoring of power transformers using DGA," IEEE Region 10 Conference TENCON, pp. 444-447 Vol. 3, doi: 10.1109/TENCON.2004.1414803.

[7] M. Aslam, M.N. Arbab, A. Basit, T. Ahmad, and M. Aamir, "A review on fault detection and condition monitoring of power transformer," International Journal of Advanced and Applied Sciences 6(8): 100-110

[8] Cigre Working Group, "Guide for transformer maintenance," Cigre Working Group A2.34, ISBN:978-2-85873-134-3

[9] M.S. Hapeez, A.F. Abidin, H. Hashim, N.R. Hamzah, and M.K. Hamzah, "Analysis and classification of different types of partial discharges by harmonic orders" Elektronika Ir Elektrotechnika 19, 35-41.

[10] J.S. Russell, and P. Norvig, "Artificial intelligence A modern approach", Prentice Hall, Englewood Cliffs, New Jersey