Generation of synthetic datasets for transformer's dissolved gas analysis using Monte-Carlo Simulation

Eaby Kollonoor Babu
Dept. of Computing and Information Sciences
Northumbria University
Newcastle, UK
eaby.k.babu@northumbria.ac.uk

Imran Bashir
School of Computing, Engineering and Information Sciences
Teesside University
Middlesbrough, UK
I.Bashir@tees.ac.uk

Gobind Pillai
School of Computing, Engineering and Information Sciences
Teesside University
Middlesbrough, UK
g.g.pillai@tees.ac.uk

Kiran Chandrakumar Jyothi
Kent Business School
University of Kent
Gillingham, UK
kc603@kent.ac.uk

***Abstract*—The fault diagnosis in power transformers is carried out using Dissolved Gas Analysis (DGA). Although DGA does provide key information for fault detection, the method is inherently complex. Several methods have been developed for DGA, but still possess challenges in accurately detecting the fault. A method has been developed to generate synthetic data using Monte-Carlo simulation. The generated synthetic data is feed into DGA excel tool to investigate the accuracy of fault detection. The synthetic data can be used to further enhance the DGA tool, improve its accuracy and investigate the inclusive faults. A model has been proposed for the integration of synthetic data generator with DGA tool for machine learning and to obtain an automated and improved DGA tool for fault diagnoses in power transformers.**

Keywords—Power transformers, fault diagnosis, dissolved gas analysis, Monte-Carlo simulation

I. INTRODUCTION

Power transformers are an integral part of any electrical power network and it's essential to have the means for early fault detection and preventive maintenance in them. Dissolved Gas Analysis (DGA) is a common method that is in use currently for incipient fault diagnosis. Certain gases are generated because of the thermal and electrical stress produced in the transformer's insulating oil (mineral oil) during its prolonged operation. Faults in the transformer can drastically increase the generation rates of these gases in certain characteristic ways. The relevant gases generated are hydrogen (H2), methane (CH4), acetylene (C2H2), ethylene (C2H4), and ethane (C2H6). The decomposition of the insulating paper results in the occurrence of carbon monoxide (CO) and carbon dioxide (CO2) in addition [1, 2]. DGA involves analyzing the concentration of these gases in the transformer oil and comparing it with historical data to find the gas generation rates. Certain ratios of the gas concentrations are calculated which can help us with identifying the fault type. The many approaches developed for analyzing these gases and interpreting their significance include Key Gas, Dornenburg Ratio, Rogers Ratio, Nomograph, IEC Ratio, Duval Triangle, and CIGRE [3].

DGA has traditionally been done manually but many ways to automate the process can be found in the literature. A diagnostic tool based on MS Excel spreadsheet has been produced in [3]. An expert model based on the fuzzy logic approach to finding the transformer incipient fault in [5]. A Support Vector Machine (SVM) with six classifiers and minor training data has been used to do four DGA methods and provide the appropriate result in [6]. A similar method in which the wavelet technique is combined with least-squares SVM has been studied in [7]. [8] proposes the use of a DGA diagnostic tool based on Artificial Neural networks (ANN) for more accurate fault detection in transformers with multiple faults. A Neural Pattern Recognition (NPR) has been suggested in [9] which uses different data transformation techniques to improve train the Artificial Intelligence (AI) the diagnostic accuracy of the model. The main drawbacks with these AI-based methods are the huge costs involved and the high amounts of time and computing power required to train the model to start producing accurate results. The results produced in these works are from limited DGA data sets that are available in the public domain, which is not enough to sufficiently train the models. Hence, it is proposed in this paper the use of synthetic data sets to train the proposed DGA tool, which can be created using synthetic data generation techniques.

Synthetic data generation has been used before in recent literature to create datasets for model training. A machine-learning-based synthetic data generation method has been used in [10] for healthcare applications, which generates synthetic time-series data that composed of nested sequences using hidden Markov models and regression models. A complex nested model has been utilized to create terabytes of structurally similar data for Internet of Things (IoT) research by the authors in [11]. The application of synthetic data generation to create training datasets can be observed in many other fields such as in plasma current quench studies [12], to analyze and predict seismic activities [13], to correctly identify and detect people using omnidirectional cameras [14], wastewater treatment modeling studies [15], and meteorological studies [16, 17]. The synthetic data has been used to test the DGA toolbox. The accuracy of DGA toolbox has been investigated using synthetic data along with its applications for machine learning.

DGA is important for the security, reliability, maintenance, and economic operation of power system infrastructure. It is evident from the literature review that there is a knowledge gap when it comes to online / machine learning-based DGA methods and one of the obstacles in the way of addressing this is the lack of synthetic data sets. Therefore, the objective of this paper is to address the gap by

978-1-6654-4389-0/21/$31.00 ©2021 IEEE

developing a method for generating synthetic dissolved gas data for DGA and validating it.

The rest of the paper is organized as follows: Section II explains the methodology used for synthetic data generation and its segregation, Section III discusses the use of machine learning model implementation, Section IV is the discussion of the results and Section V is the Conclusion.

II. METHODOLOGY

For DGA, it is important to identify and interpret the fault correctly. Multiple DGA methods have been developed, but still discrepancies exist in detecting the fault. Duval triangle and Pentagon methods were developed to reach a conclusive decision. Nevertheless, there still exists some unidentified fault. Therefore, the developed DGA toolbox is being trained using synthetic data. The synthetic data is generated to identify all possible faults with a focus on fault regimes, where a conclusive decision cannot be taken.

The procedure for data synthesis mainly consists of 2 stages. In stage one we find the data pattern from a preexisting dataset of DGA's. This data is passed to a data segregator block which understands the data structure and finds the number of classes, upper limit, lower limit, mean, median, variant, mode, standard deviation, degree of freedom, and range from any given numerical dataset. These values are stored in a data collector block which acts as an input to the Monte-Carlo data generator in stage 2 of the process. The block diagram of the proposed process is shown below.

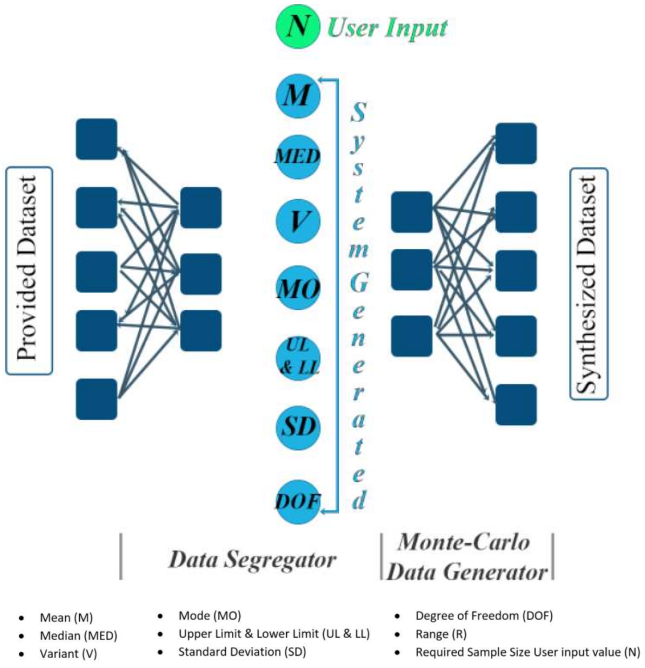

Fig. 1. Block diagram for data synthesis

A. Data Segregator

This functional block takes any numerical dataset as input and understands the structure of the dataset provided. This is carried out with 3 functions stated below.

Class Finder: From the provided dataset this function finds all the column header and checks for numerical value content in each column, if multiple numerical values are found it makes a class with the column header name. This process is carried out till all columns are scanned. Once this scanning is completed the Class finder passes the class names to the data collector class.

Limit Finder: For each of the found classes this function finds the upper limit and lower limit after analyzing each of the class data. The formula used in the limit finder is:

$$f_{Max} = MAX(A2: A_n)$$
$$f_{min} = MIN(A2: A_n)$$

Stat Generator: Once all the classes are passed to the Data Collector, we then use Stat Generator to get mean, median, variant, mode, range, standard deviation, degree of freedom, and number repeatability.

For finding the median of the series and Degree of Freedom, we use a different approach which is as below;

$$f_{MED} = \begin{cases} A\left[\frac{n}{2}\right] & if\ n\ is\ even \\ \frac{A\left[\frac{n-1}{2}\right]+A\left[\frac{n+1}{2}\right]}{2} & if\ n\ is\ odd \end{cases}$$

And for Degree of Freedom the formula is;

$$DF = n - 1 \quad for\ each\ of\ the\ class\ series$$

Once all the above paraments are calculated the results are passed to the Data Collector. The Data Collector as the input to the Monte-Carlo data generator.

B. Monte-Carlo Data Generator

The algorithm for the proposed Mote-Carlo generator is given as,

Algorithm 1: Monte-Carlo Data Synthesizer

Step 1: Get the required sample size from the user **(N)**

Step 2: Get the total number of classes required from the Data Collector **(NC)**

Step 3: **for** i from 1 to NC

do

for j from 1 to N

do

$$\lim_{Value \to between\ UL\ \&\ LL} \left(1 + \frac{1}{MED} + \frac{DOF * SD}{V} + M\right) = Value$$

if (N = MO)

Integer of *Value* $= \left(1 + \frac{1}{MED} + MO\right)$ and MO = MO+MO

else

Value = Integer of ***Value***

Outcome ***Value*** *is Stored in the corresponding class header cell*

end

end

Our proposed computation algorithm used in the Monte-Carlo method consist of a combination of parameters as listed:

- Mean (M)
- Median (MED)
- Variant (V)
- Mode (MO)
- Upper limit & Lower limit (UL & LL)
- Standard deviation (SD)
- Degree of freedom (DOF)
- Range (R)
- Required sample size (user input value) (N)

The DGA excel tool as proposed by A. Mekkayil, G. Pillai, and M. Malcolm [4] has been used to validate the generated synthetic dataset accuracy. This tool is capable of predicting all the 7 basic faults in DGA by using the existing methods like Key Gas Method, Ratio Method, IEC Ratio Method, Duval Triangle Method, and Duval Pentagon Method. The faults are classified as Partial Discharge (PD), Discharges of low energy (D1), Discharges of high energy (D2), Combination of thermal faults and discharges (DT), Thermal faults not exceeding 3000C (T1), Thermal faults between the range of 3000C and 7000C (T2) and Thermal faults exceeding 700º C (T3). A summary of faults is given in Table I.

TABLE I. IEC Ratio Method Diagnosis [4]

Fault Code	Fault Type	C2H2 / C2H4	C2H2/ C2H4	C2 H4/ C2 H6
PD	Partial discharges	NSa	<0.1	<0.
D1	Discharges of low	>1	0.1 - 0.5	2
D2	energy	0.6-	0.1 - 1	>1
T1	Discharges of high	2.5	>1 but	>2
T2	energy	NSa	NSa	>1
	Thermal fault, T<	<0.1		
T3	300 °C	<0.2b	>1	1– 4
	Thermal fault		>1	>4
	300 °C < T < 700 °C			
	Thermal fault, T >			
	700 °C			

*NSa = Non-significant whatever the value.

*An increasing trend in the concentration of C2H2 may indicate that the hotspot temperature is above 1000 °C.

III. DGA AND MACHINE LEARNING MODEL

As the world is moving towards Industry 4.0 revolution, digital twin and machine learning are coming into play. Therefore, large data sets are required to train the machine learning and feed it into digital twin models. However, in some cases i.e. DGA, large data sets are not available or difficult to obtain. To resolve this, the paper presents an innovative solution for integrating the DGA toolbox with synthetic data generation as shown in Figure 3.

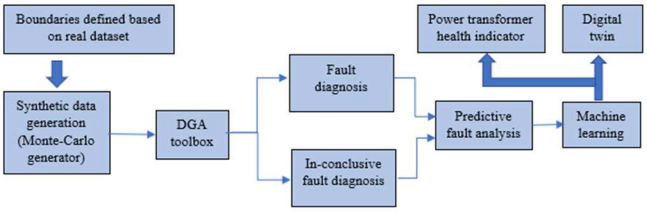

Fig. 2. Integration of synthetic data with DGA toolbox and machine learning

The proposed model in Figure 3 needs to validate and continuously feed into available small data sets. This will not only validate the model but also built a database. The synthetic data generator can be used for the optimization of DGA toolbox. The Duval triangle and Pentagon methods do provide good indicators for fault diagnosis, but the limitations can be optimized using synthetic data with the possibility of new shape development. Historic DGA real data can also be feed to train and optimized the model.

Although the proposed model in Figure 3 does provide a promising enhancement of DGA toolbox. However, it does also bring significant challenges. The generation of synthetic data is entirely dependent on the sample of real data provided as a seed value. Similarly, the larger sample of real data means a better generation of synthetic data. However, the data is obtained from different transformers might provide different seed values irrespective of similar faults. The difference in seed value for similar faults can lead to the generation of different synthetic datasets and thus a different final analysis. Although, the synthetic data generator and excel tool provide an accuracy of more than 80 %. Nevertheless, this error can be minimized by using the modified feedback control loop model as shown in Figure 4. The feedback system can help to validate and optimize fault detection and improve accuracy. The accuracy and reliability of the synthetic data generation can be further improved by using data from multiple sources, which are well-defined. As the number of available real datasets increases, the model can be modified to make it more robust and uncertainty can be reduced.

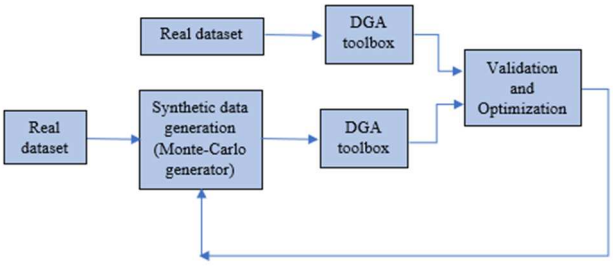

Fig. 3. Optimization and validation of DGA toolbox

Once the proposed models have been optimized and validated, it is expected that the performance of DGA toolbox

will improve significantly. Moreover, the proposed models in Figure 3 and Figure 4 are a step forward to implement the DGA toolbox in Industry 4.0.

IV. Results and Discussion

The main reason for using synthetic data for DGA study is because it's not easy to obtain large sets of real data. Therefore, the expected boundaries of data have been used to generate synthetic data using the Monte-Carlo Data generator. DGA excel tool has been feed-in into synthetic data to study the accuracy and inclusive faults [18].

Synthetic data is used to study the accuracy of DGA faults as given in Table I. Figure 2 plots the accuracy of the predicted faults by DGA tool for synthetic data input as generated by Monte-Carlo simulations. For 70 data points generated for Thermal faults exceeding 700º C (T3) by the synthetic data generator, the excel tool predicted 65 T3 faults accurately, with an accuracy of 92.8%. Similarly, the error was calculated for all other faults. It has been calculated that the Discharge of low energy (D1) has the highest accuracy of 93.33%, followed by T1 and T3 at 92%, then PD at 90%, T2, D2, and S stands at 86%. The synthetic data generator was able to generate a dataset for all the desired faults as the user's requirement with an overall accuracy of above 90%.

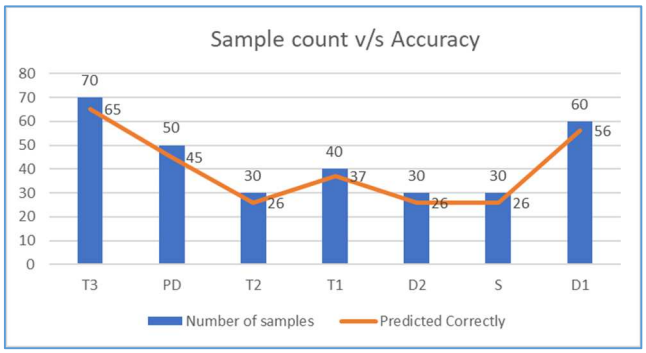

Fig. 4. Accuracy of the predicted faults by DGA tool for synthetic data input as generated by Monte-Carlo simulations.

Several simulations have been carried out to test the accuracy of DGA excel tool for different sets of synthetic data generation. The accuracy of the fault diagnosis remained above 86 %. It has been concluded that the DGA tool and synthetic data generator is shown the results as expected. The DGA excel tool be integrated with a synthetic data generator for machine learning and artificial intelligence applications.

V. Conclusions

DGA is important for the security, reliability, maintenance, and economic operation of power system infrastructure. In this paper, we developed a method for generating synthetic dissolved gas data for DGA using Monte-Carlo Simulation. In the previous work, an excel tool was developed for DGA which was used in this work to validate the synthetic data. The synthetic data generator and DGA tool are integrated and the accuracy of the synthetic data has been investigated. It has been concluded that for all fault detection cases using synthetic data, accuracy of more than 85 % has been achieved. The synthetic data can be used to further enhance the DGA tool, improve its accuracy and investigate the inclusive faults. A model has also been proposed to integrate the synthetic data generator with DGA tool for machine learning applications. This model can help to train machine learning by generating several different synthetic datasets. Further work will be a focus on the implementation of the proposed model for machine learning applications.

References

[1] N.A. Bakar, A. Abu-Siada, and S. Islam, "A review of dissolved gas analysis measurement and interpretation techniques", IEEE Electri. Insul. Mag., pp. 39-49, May 2014.

[2] H. de Faria Jr., J. G. S. Costa and J. L. M. Olivas, "A review of monitoring methods for predictive maintenance of electric power transformers based on dissolved gas analysis", Ren. Sust. Energy Rev., vol. 46, pp.201-209, June 2015.

[3] H. Sun, Y. Huang and C. Huang, "A review of dissolved gas analysis in power transformers", Energy Procedia, vol. 12, pp. 1220-1225, March 2012.

[4] A. Mekkayil, G. Pillai and M. Malcolm, "A spreadsheet based dissolved gas analysis diagnostic tool for oil-filled transformers", 55th Int. Univ. Power Eng. Conf., September 2020.

[5] A. Abu-Siada and S. Hmood, "A new fuzzy logic approach to identify power transformer criticality using dissolved gas-in-oil analysis", Int. J. Elec. Power & Ener. Sys., vol. 67, pp. 401-408, May 2015.

[6] K. Bacha, S. Souahlia and M. Gossa, "Power transformer fault diagnosis based on dissolved gas analysis by support vector machine", Elec. Power Sys. Res., vol. 83, pp. 73-79, February 2012.

[7] H. Zheng et al., "A novel model based on wavelet LS-SVM integrated improved PSO algorithm for forecasting of dissolved gas contents in power transformers", Elec. Power Sys. Res., vol. 155, pp. 196-205, February 2018.

[8] G. S. Naganathan et al., "Internal fault diagnosis of power transformer using artificial neural network", Materials Today: Proceedings, March 2021.

[9] I. B. M. Taha, S. S. Dessouky and S. S. M. Ghoneim, "Transformer fault types and severity class prediction based on neural pattern-recognition techniques", Elec. Power Sys. Res., vol. 191, February 2021.

[10] J. Dahmen and D. Cook, "SynSys: A synthetic data generation system for healthcare applications", Sensors, vol. 19, March 2019.

[11] J. W. Anderson, K. E. Kennedy, L. B. Ngo, A. Luckow and A. W. Apon, "Synthetic data generation for the internet of things", 2014 IEEE Int. Conf. Big Data, pp. 171-176, October 2014.

[12] N. Dalsania, Z. Patel, S. Purohit and B. Choudhary, "An application of machine learning for plasma current quench studies via synthetic data generation", Fusion Eng. Des., vol. 171, April 2021.

[13] G. Roncoroni, C. Fortini, L.Bortolussi, N.Bienati and M. Pipan, "Synthetic seismic data generation with deep learning", J. App. Geophysics, vol. 190, April 2021.

[14] N. Aranjuelo, S. Garcia, E. Loyo, L. Unzueta and O. Otaegui, "Key strategies for synthetic data generation for training intelligent systems based on people detection from omnidirectional cameras", Comp. Elec. Engg., vol. 192, March 2021.

[15] L.J.P.Snip, X. Flores-Alsina, I. Aymerich, S. Rodríguez-Mozaz, D. Barceló, B. G. Plósz, Ll. Corominas, I. Rodriguez-Roda, U. Jeppsson and K.V. Gernaeya, "Generation of synthetic influent data to perform (micro)pollutant wastewater treatment modelling studies", Sci. Tot. Env., vol. 569-570, pp. 278-290, November 2016.

[16] M. A. Hassan, M. Abubakr and A. Khalil, "A profile-free non-parametric approach towards generation of synthetic hourly global solar irradiation data from daily totals", Renewable Energy, vol. 167, pp. 613-628, April 2021.

[17] M. Bervida, L. Patruno, S. Stanic and S. de Miranda, "Synthetic generation of the atmospheric boundary layer for wind loading assessment using spectral methods", J. Wind Engg. Ind. Aero., vol. 196, January 2020.

[18] A. M. Berg, S. T. Mol, G. Kismihok, and N. Sclater, "The Role of a Reference Synthetic Data Generator within the Field of Learning Analytics," J. Learn. Anal., vol. 3, no. 1, pp. 107–128, 2016.