DOI: 10.2478/sbeef-2023-0014

MACHINE LEARNING ALGORITHMS FUSION BASED ON DGA DATA FOR IMPROVING FAULT DIAGNOSIS OF ELECTRICAL POWER TRANSFORMER

Abdelmoumene Hechifa[1], Abdelaziz Lakehal[2,*], Arnaud Nanfak[3], Lotfi Saidi[4], Chouaib Labiod[5]

[1]LGMM Laboratory, Faculty of Technology, University of 20 August 1955-Skikda, Skikda, Algeria.
[2]Laboratory of Research on Electromechanical and Dependability, University of Souk Ahras, Souk-Ahras, Algeria
[3]Laboratory of Energy, Materials, Modelling and Methods, National Higher Polytechnic School of Douala, University of Douala ,Douala , Cameroon.
[4]University of Tunis, ENSIT – Laboratory of Signal Image and Energy Mastery, Tunis, Tunisia
[5]Electrical Engineering Department, Faculty of Technology, University of El Oued, El Oued, Algeria.
* Corresponding author: a.lakehal@univ-soukahras.dz

Abstract*: Dissolved Gas Analysis (DGA) continues to be widely recognized as a valuable method in recent times for the early identification of issues in oil-filled power transformers. It has gained extensive adoption as a primary approach for the early discovery of these issues, relying on the analysis of dissolved gases. This contributes to enhancing the dependability of electrical systems. This paper proposes an efficient fusion method based on DGA data using the two best Machine Learning algorithms , the neural network (MLP), the naïve Bayes (NB) throughdata input vector ppm, a percentage input vector, and an Logarithmic input vector. The fusion method predictively combined the two classifiers and obtained a statistical evaluation: accuracy, recall, precision, and F-measure higher than both classifiers separately. The proposed fusion method was evaluated for performance using a test database and compared with conventional and smart methods. Results showed that the proposed model outperformed both traditional and intelligent methods in terms of diagnostic accuracy when using percentage and logarithmic input vectors. The Prediction Based Fusion (PBF) vector Percentages achieved an accuracy rate of 97.22%, while PBF vector Logarithmic achieved an accuracy rate of 95.83%. These rates were higher than those achieved by traditional methods, such as the Modified RRM/CEGB method 91.67% and Modified RRM/IEC method 90.28%. Additionally, the proposed model surpassed the accuracy rates of intelligent methods, such as CSUS ANN 88.89% and Conditional Probability 93.06%.*

Keywords: Dissolved gas analysis; fusion method, Multilayer Perceptron, Naïve Bayes**,** Percentages, Logarithmic.

1. INTRODUCTION

The importance of power transformers lies in their crucial role in ensuring the smooth operation of power systems. Regular monitoring is necessary to maintain their availability and prevent any potential faults. In the event of a transformer fault, the entire power network can be affected, leading to catastrophic consequences in the transmission of electricity [1]. Power transformers are susceptible to both thermal and electrical stresses, which have the potential to induce decomposition of the insulating oil, subsequently resulting in oxidation. This process leads to the production of various gases, including Methane (CH_4), carbon monoxide (CO), Hydrogen (H_2), Acetylene (C_2H_2), Ethylene (C_2H_4), Ethane (C_2H_6), carbon dioxide (CO_2) [2]. To analyze the gases produced by transformer faults, a commonly used method is Dissolved Gas Analysis (DGA) [3].

DGA is a powerful diagnostic technique and widely used by the majority of power company for detecting thermal and electrical faults in transformers. International committees acknowledge its efficiency in diagnosing issues resulting from oil or paper, including Low/Medium/High thermal faults or electrical faults, including Partial discharge and Low/High energy discharge [4]. To simplify the process of identifying potential faults in transformers, established rules and methods rely on the analysis of dissolved gas concentrations [5].

Accurate methods for diagnosing faults in power transformers are essential for conducting a thorough DGA analysis. At present, there are traditional methods represented in ratio methods: Dornenburg method [6], modified Rogers four ratios, modified IEC ratios methods [7], HYOSUN Corporation gas ratio method [8], three ratios technique [9], In addition, the graphic methods are: Duval triangle method [10], Gouda triangle method [11], pentagon methods [12] [13]. The traditional methods are not accurate enough to know the type of fault, so they are weak and suffer from decision-making, but by turning to artificial intelligence, diagnosing faults for transformers has become available and with high precision. The methods used in the literature are: ANN [14], fuzzy logic [15], support vector machines [16], k-nearest neighbor [17], Bayesian networks [18], ensemble learning [19], and deep learning [20].

In this paper, both the neural network algorithm and the naïve Bayes (NB) theory were presented using different input vectors and a proposed method for fusion them to diagnose the six power transformer faults using the KNIME analytics platform.

This paper is prepared as follows. In Section 2, the methodology utilized in this study is introduced, commencing with an elucidation of the employed dataset, followed by a discussion of the various input vectors considered, and an overview of the faults of transformers targeted for detection. Section 3 provides an overview of the machine learning algorithms used in this study. In Section 4, the model evaluation tools are shown. Section 5 evaluates the proposed fusion technique's performance and effectiveness, comparing it with traditional and intelligent methods in the literature using test data. Finally, some conclusions of the presented study are given in the section 6.

2. METHODOLOGY

2.1 Data Description

Data collection is the first and most important step for any study, as it is an integral part of the work that enables us to enhance the lifespan of machines by diagnosing and predicting their faults. The data for this study [21] was collected from 240 samples, comprising 129 electrical fault samples and 111 thermal fault samples, which are further categorized into six types of faults, as follows: Partial discharge (PD=27 cases), Low energy discharge (D1=42 cases), High energy discharge (D2=55 cases), Thermal fault below 300°C (T1=70 cases), Thermal fault between 300°C and 700°C (T2=18 cases), and Thermal fault above 700°C (T3=28 cases). The training process used 70 % of the data (168 samples), while the remaining 30 % of the testing (72 samples) were used.

2.2 Input vector preparation

In order to ensure the quality of any data mining process, the data must be processed by changing the input vectors and data format to facilitate dealing with them. This procedure represents the pivotal phase for enhancing the model's accuracy. In this study, three different input vectors were employed, namely the original data input vector (ppm), a percentage input vector, and a Logarithmic input vector, as shown in Table 1.

Table 1. Input features

Data Format	
input vectors (ppm)	X= [C_2H_6, C_2H_2, CH_4, C_2H_4, H_2]
input vectors (Percentage)	X= [%C_2H_6, %C_2H_2, %CH_4, %C_2H_4, %H_2]
input vectors (Logarithmic)	X= [LogH_2, logCH_4, LogC_2H_2, LogC_2H_4, LogC_2H_6]

3. THE USED MACHINE LEARNING ALGORITHMS

3.1 Multilayer Perceptron (MLP)

Artificial Neural Networks (ANNs) are symmetric processors that exhibit high interconnectedness and represent an advanced and efficient algorithm for solving problems related to data scarcity and interference. Neurons serve as the basic building blocks of neural networks, drawing inspiration from the complex neurons and neural connections found in the human brain [22].

The Multilayer Perceptron (MLP) is widely considered to be one of the most powerful and effective types of neural networks for handling data. MLPs belong to the category of supervised neural networks, which rely on desired outputs for machine learning [23]. On the other hand, the MLP algorithm comprises three primary layers. The input layer establishes the connection between the data input vector and the network, while the hidden layer varies from one algorithm to another, depending on the sensory cells employed. Finally, the output layer processes the data for the input vector via the sensory cells and directs it through its activation function, as represented by [24].

$$y_i = f\left(\sum_{j=1}^{n} \omega_{ij} x_j + \theta_i\right) \quad (1)$$

x_j: is the jth input of the ith neuron.

θ_i: is called the bias of the ith neuron.

ω_{ij}: is the weight from the jth input to the ith neuron.

y_i: is the output of the ith neuron.

The corresponding Figure 1 represents the classified MLP architecture, which consists of three layers, input layer, hidden layer, and output layer.

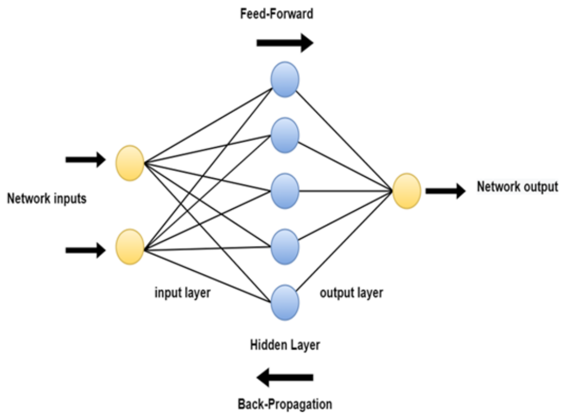

Figure 1. The structure of the (MLP) based classifier

3.2 Naive Bayes (NB)

The Naive Bayes (NB) algorithm is technique inspired by Bayes' theorem. It operates on the principle of probability theory and statistical methods. Bayes theorem establish independent values based on what precedes them [25]. The NB algorithm is considered "naive" because it assumption of independence among all variables concerning class values, which may not hold in real-world scenarios. Nonetheless, this approach is favored for its fast-learning capabilities [26]. The Bayes theorem algorithm offers the advantage of classifying various objects, with the posterior probability equation being described as follows [27]:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

$P(A|B)$ is the posterior probability of *A* under condition *B*.

$P(A)$ i prior probability *A*.

$P(B|A)$ posterior probability under condition *A*.

$P(B)$ is the prior probability *B*.

The posterior probability:

$$Posterior = \frac{likelihood \times prior}{evidence} \quad (3)$$

The principle of the NB algorithm can be illustrated through a simplified diagram and how to classify things; Figure 2 illustrates the basic structure of a NB.

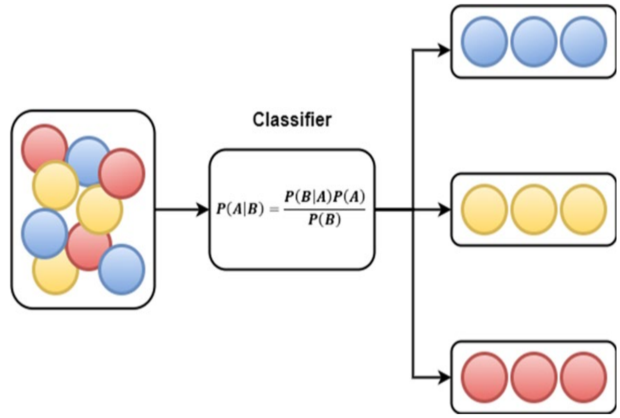

Figure 2. Simplified diagram of the NB algorithm

4. MODEL EVALUATION

In the field of data mining, there exists a dependable method to assess the accuracy of data, which is crucial in supporting documentation systems. The effectiveness of the model is evaluated using statistical measures, including Accuracy, recall, precision, and F-measure [28].

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F_{-measure} = \frac{2(Precision \times Recall)}{Precision + Recall} \quad (7)$$

5. RESULTS AND DISCUSSION

To assess the diagnostic efficacy of the proposed model and to predict transformer faults that could endanger the condition of the electrical system, a dataset consisting of 240 samples was used which is as follows: 168 training samples and 72 testing samples. The used data are in ppm, percentage, and arithmetic. Where this paper represents one of the most powerful and accurate algorithms, namely, the naive rule and neural networks in diagnosing transformer faults, which in turn, their prediction explaining the first stage, which is the preparation and processing of data involves collecting data and extracting features for input vectors. In the second stage, classifiers are trained and tested. The third stage, known as the Prediction Based Fusion (PBF) method, is a machine learning and data science technique used to combine multiple predictions from MLP and NB algorithms to improve the overall accuracy and robustness of predictions. This is achieved by averaging the predictions of different models. Finally, in stage four, the proposed model is evaluated. They were combined through ensemble learning to increase the accuracy of the model, using the KNIME analytics platform. Figure 3 represents the proposed model structure.

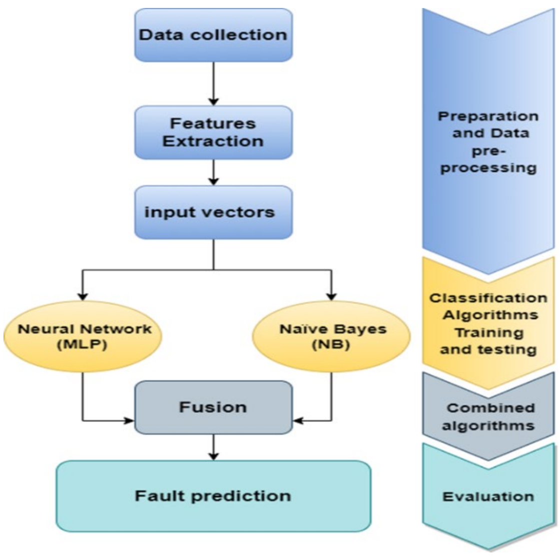

Figure 3. Flowchart of the proposed model

The KNIME analytics platform environment allows engineers to develop and implement algorithms in a short time through a group of interconnected nodes, each of which performs a specific function and enables the expert to enter output and modify data [29]. Figure 4 represents the proposed model and PBF with a simplified explanation of the steps.

Data pre-processing: The input data is processed by the Reader node, which scans the input file to ascertain the quantity and categories of columns present. The Duplicate Rows node then removes all duplicate rows within the input table, followed by processing any instances of missing data.

Model training and testing: First, the data is partitioned into training and testing segments via the partition node. Then, the model is trained and evaluated using a node learner and predictor.

Combined prediction: The predictions are collected first by the joiner node, which keeps the class and prediction columns, and the PBF node is collected using the mean predictions of both NB and MLP algorithms.

Evaluation: The proposed model was evaluated through accuracy, recall, precision, and F-measure.

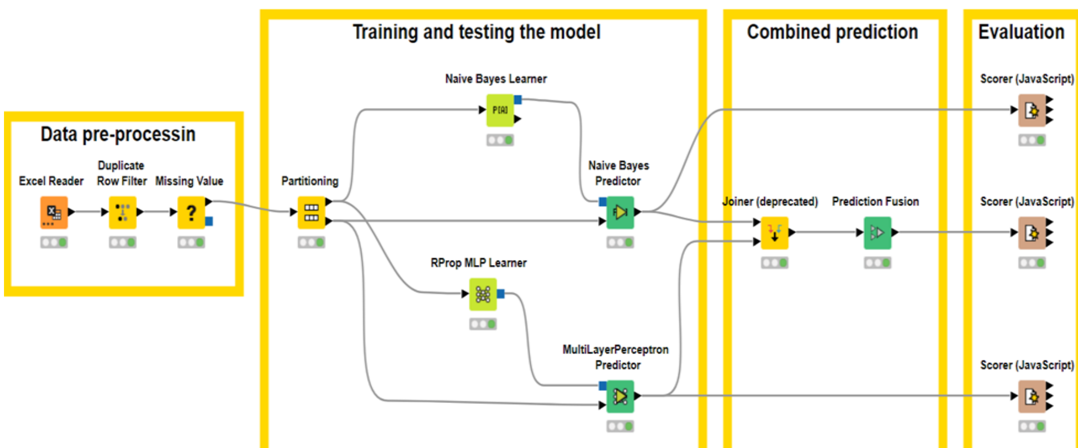

Figure 4. The proposed model using PBF to combine predictions of different classifiers

5.1 Performance evaluation results

When comparing Figure 5 with Table 2, it can be seen that the PBF of assembling the algorithm of NB with the algorithm of MLP over both algorithms separately when the input vectors are ppm, percentages, and logarithmic. The proposed method of PBF is characterized by its reliability in fault detection in terms of overall accuracy, and macro average : precision, recall, and F-measure.

The best results of the proposed method for PBF at input vector ratio of 97.22% accuracy, 98.33% recall, 97.62% precision, and F-measure of 97.83% were superior to both input vector ppm with 91.67% accuracy, 86.08 recall, 92.73% precision, and F-measure. 94.75, as well as a logarithmic input vector with 95.83% precision, 96.14% recall, 94.72% precision and 95.32% F-measure.

While for the algorithms separately, the high efficiency of the MLP algorithm in terms of the input vector was in percentages and ppm, and the algorithm of the NB prevailed when the input vector was logarithmic.

Table 2. Performance results of the proposed model

Proposed classification techniques	Input features											
	ppm				Percentages				Logarithmic			
	Acc%	Rec%	Pre%	F-m%	Acc%	Rec%	Pre%	F-m%	Acc%	Rec%	Pre%	F-m%
NB	77.78	70.94	72.30	81.91	91.67	94.56	90.88	92.45	93.06	94.56	91.81	92.61
MLP	84.72	81.53	86.70	81.6	94.44	96.67	95.83	95.77	88.89	84.97	85.63	84.37
Prediction Based Fusion of NB With MLP	91.67	86.08	92.73	94.75	97.22	98.33	97.62	97.84	95.83	96.14	94.72	95.32

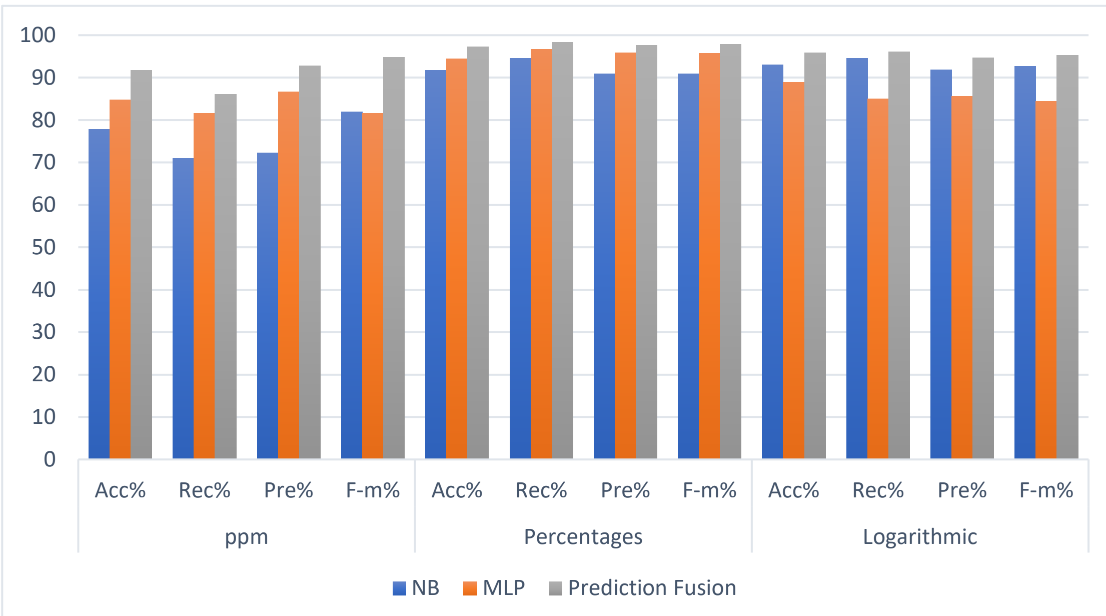

Figure 5. Performance chart algorithms for different evaluation metric

5.2 Comparisons with previous studies

For the effectiveness of the proposed fusion prediction model and to make comparisons with previous studies, the test data set consisting of 72 samples was used to find out the reliability of the proposed method in predicting transformer faults. Table 3 shows a comparison of the results obtained with other techniques, whether traditional or Intelligent.

The comparison results demonstrated that the proposed model outperformed both traditional and intelligent methods in terms of diagnostic accuracy when using percentage and logarithmic input vectors. This improvement in accuracy can be attributed to the preprocessing of the data, which involved converting the original ppm data to either percentage or logarithmic format. The PBF (vector Percentages) achieved an accuracy rate of 97.22%, while PBF (vector Logarithmic) achieved an accuracy rate of 95.83%. These rates were higher than those achieved by traditional methods, such as the Modified RRM/CEGB method 91.67% and Modified RRM/IEC method 90.28%, as well as intelligent methods, such as CSUS ANN 88.89% and Conditional Probability 93.06%.

Table 3. Comparison between the proposed method with traditional and Intelligent methods

Diagnostic methods		Number of correct cases						Total accuracy (%)
		PD	D1	D2	T1	T2	T3	
Traditional methods	Three ratios technique [9]	4	6	1	12	0	8	43.06
	Modified RRM/CEGB [7]	7	10	15	20	6	8	91.67
	Modified RRM/IEC [7]	6	10	15	20	6	8	90.28
	Hyosun Corporation gas ratio method [8]	2	10	1	8	2	8	43.06
	Duval triangle method [10]	4	8	2	10	0	5	43.06
	Gouda triangle method [11]	3	10	3	9	0	8	45.83
Intelligent methods	CSUS ANN [30]	7	10	14	19	6	8	88.89
	Conditional probability [31]	7	11	16	19	6	8	93.06
Proposed method	Prediction Based Fusion (vector ppm)	7	16	16	18	1	8	91.67
	Prediction Based Fusion (vector Percentages)	9	12	18	17	5	9	97.22
	Prediction Based Fusion (vector Logarithmic)	7	16	14	20	7	5	95.83

6. CONCLUSION

This article proposes an efficient fusion method based on DGA data, utilizing the two best Machine Learning algorithms, MLP and NB. By using input vectors of ppm, percentage, and logarithmic data, the proposed fusion method combines the classifiers and achieves higher accuracy rates, recall, precision, and F-measure than the classifiers alone. The performance of the proposed model was evaluated using a test database and compared to traditional and intelligent methods. Results showed that the proposed model outperformed all other methods, achieving an accuracy rate of 97.22% for the PBF vector Percentages and 95.83% for PBF vector Logarithmic. These rates were higher than those achieved by traditional methods such as the Modified RRM/CEGB 91.67% and Modified RRM/IEC 90.28% methods, as well as intelligent methods such as CSUS ANN 88.89% and Conditional Probability 93.06%. Overall, this study demonstrates that the proposed fusion method is a promising approach for DGA analysis and can significantly improve the accuracy of DGA-based diagnosis.

7. REFRENCES

[1] Nasirul, Haque, Jamshed Aadil, Chatterjee Kingshuk, and Chatterjee Soumya. "Accurate Sensing of Power Transformer Faults From Dissolved Gas Data Using Random Forest Classifier Aided by Data Clustering Method." IEEE Sensors Journal 22, no. 6 (2022): 5902 – 5910.

[2] Taha, Ibrahim B.M, and DiaaEldin, A. Mansour. "Novel power transformer fault diagnosis using optimized machine learning methods." Intelligent Automation & Soft Computing 28, no. 3 (2021): 739-752.

[3] Passara, Chanchotisatien, and Vong Chanvichet. "Feature engineering and feature selection for fault type classification from dissolved gas values in transformer oil." International Computer Science and Engineering Conference (ICSEC), 2021: 75-80.

[4] Rao, U. Mohan, I. Fofana, K. N. V. P. S. Rajesh, and P. Picher. "Identification and Application of Machine Learning Algorithms for Transformer Dissolved Gas Analysis." IEEE Trans Dielectr Electr Insul. 28, no. 5 (2021): 1828–1835.

[5] Taha, Ibrahim B. M., Saleh Ibrahim, and Diaa-Eldin A. Mansour. "Power Transformer Fault Diagnosis Based on DGA Using a Convolutional Neural Network with Noise in Measurements." IEEE Access 9 (2021): 111162 –111170.

[6] Wannapring, Ekkarach, Cattareeya Suwanasri, and Thanapong Suwanasri. "Dissolved Gas Analysis methods for distribution transformers." International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI CON), 2016: 1-6.

[7] Taha, Ibrahim B.M, Ayman Hoballah, and Sherif SM Ghoneim. "Optimal ratio limits of rogers' four-ratios and IEC 60599 code methods using particle swarm optimization fuzzy-logic approach." IEEE Transactions on Dielectrics and Electrical Insulation 27, no. 1 (2020): 222-230.

[8] Sung-wook, Kim, Sung-jik, Kim, Hwang-dong, Seo, Jae-ryong, Jung, Hang-jun, Yang, Michel, Duval, et al. "New methods of DGA diagnosis using IEC TC 10 and related databases Part 1: application of gas-ratio combinations." *IEEE Transactions on Dielectrics and Electr Insula* 20, (2013): 685-690.

[9] Gouda, Osama E., Salah H. El Hoshy, and Hassan H. EL Tamaly. "Proposed three ratios technique for the interpretation of mineral oil transformers based dissolved gas analysis." *IET Generation, Transmiss Distribu* 12, (2018): 2650-2661.

[10] Khiar, Mohd Shahril Ahmad, Sharin Ab Ghani, Imran Sutan Chairul, Yasmin Hanum Md Thayoob, and Young Zaidey Yang Ghazali. "On-site OLTC monitoring using Duval triangle and DWRM." In Proceedings of the 2nd International Conference on Technology, Informatics, Management, Engineering and Environment, TIME E 2014, Bandung, Indonesia 1 (2014): 216–221.

[11] Gouda, Osama E., Salah H. El-Hoshy, and Hassan H. EL-Tamaly. "Condition assessment of power transformers based on dissolved gas analysis." *IET Generation, Transmission & Distribution* 13, no.12 (2019): 2299-2310.

[12] Duval, Michel, and Laurent Lamarre. "The duval pentagon-a new complementary tool for the interpretation of dissolved gas analysis in transformers." IEEE Electrical Insulation Magazine 30, no. 6 (2014): 9–12.

[13] Mansour, Diaa-Eldin A. "Development of a new graphical technique for dissolved gas analysis in power transformers based on the five combustible gases." IEEE Transactions on Dielectrics and Electrical Insulation 22, no. 5 (2015): 2507–2512.

[14] Li, Anyi, Xiaohui Yang, Zihao Xie, and Chunsheng Yang. "An optimized GRNN-enabled approach for power transformer fault diagnosis." IEEJ Trans. Electr. Electron. Eng 14, no. 8 (2019): 1181–1188.

[15] Taha, Ibrahim B. M., Sherif S. M. Ghoneim, and Hatim G. Zaini. "A Fuzzy Diagnostic System for Incipient Transformer Faults Based on DGA of the Insulating Transformer Oils." International Review of Electrical Engineering (I.R.E.E.) 11, no. 3 (2016): 305-313.

[16] Zhang, Yiyi, et al. "A fault diagnosis model of power transformers based on dissolved gas analysis features selection and improved Krill Herd algorithm optimized support vector machine." IEEE Access 7 (2019): 102803–102811.

[17] Benmahamed, Y, Y Kemari, M Teguar, and A Boubakeur. "Diagnosis of power transformer oil using KNN and nave bayes classifiers." IEEE 2nd International Conference on Dielectrics (ICD), 2018: 1–4.

[18] Lakehal, Abdelaziz, and Fouad Tachi. "Bayesian duval triangle method for fault prediction and assessment of oil immersed transformers." *Measurement and control* 50, no. 4 (2017): 103-109.

[19] Nanfak, Arnaud, Charles Hubert Kom, and Samuel Eke. "Hybrid Method for Power Transformers Faults Diagnosis Based on Ensemble Bagged Tree Classification and Training Subsets Using Rogers and Gouda Ratios." Int. J. Intell.Eng. Syst 15, no. 5 (2022): 12-24.

[20] Wu, Xiaoxin, Yigang He, and Jiajun Duan. "A deep parallel diagnostic method for transformer dissolved gas analysis." *Applied Sciences* 10, no. 4 (2020): 1329.

[21] Ibrahim, Saleh I., Sherif S.M. Ghoneim, and Ibrahim B.M Taha. "DGALab: an extensible software implementation for DGA." IET Gener. Transm. Distrib 12, no. 18 (2018): 4117-4124.

[22] Patekar, Kalinda D., and Bhoopesh Chaudhry. "DGA analysis of transformer using artificial neutral network to improve reliability in Power Transformers." IEEE 4th International Conference on Condition Assessment Techniques in Electrical Systems (CATCON), 2019: 1–5.

[23] Ramkumar, M., C. Ganesh Babu, K Vinoth Kumar, D Hepsiba4, A. Manjunathan, and R. Sarath Kumar. "ECG Cardiac arrhythmias Classification using DWT, ICA and MLP Neural Networks." Journal of Physics: Conference Series 1831, no. 1 (2021): 012015.

[24] Xiang, Weiming, Hoang-Dung Tran, and Taylor T. Johnson. "Output reachable set estimation and verification for multilayer neural networks." IEEE Transactions on Neural Networks and Learning Systems 29, no. 11 (2018): 5777–5783.

[25] Ketjie, Viny Christanti Mawardi, and Novario Jaya Perdana. "Prediction of credit card using the Naïve Bayes method and C4.5 algorithm. IOP Conference Series." Materials Science and Engineering 1007, no. 1 (2020): 012161.

[26] Mahamdi, Yassine, Ahmed Boubakeur, Abdelouahab Mekhaldi, and Youcef Benmahamed. "Power Transformer Fault Prediction using Naive Bayes and Decision tree based on Dissolved Gas Analysis." ENP Engineering Science Journal 2, no. 1 (2022): 1-5.

[27] Balaji, V.R., S.T. Suganthi, R. Rajadevi, V. Krishna Kumar, B. Saravana Balaji, and Sanjeevi Pandiyan. "Skin disease detection and segmentation using dynamic graph cut algorithm and classification through Naive Bayes classifier." Measurement: Journal of the International Measurement Confederation 163 (2020): 107922.

[28] Rahmad, F, Y Suryanto, and K Ramli. "Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification." IOP Conference Series: Materials Science and Engineering 879, no. 1 (2020): 012076.

[29] Nicola, George, Michael R. Berthold, Michael P. Hedrick, and Michael K. Gilson. "Connecting proteins with drug-like compounds: Open source drug discovery workflows with BindingDB and KNIME." Database-Oxford, 2015: 1–22.

[30] Ghoneim, Sherif SM, Ibrahim BM Taha, and Nagy I. Elkalashy. "Integrated ANN-based proactive fault diagnostic scheme for power transformers using dissolved gas analysis." *IEEE Transactions on Dielectrics and Electrical Insulation* 23, no. 3 (2016): 1838-1845.

[31] Taha, Ibrahim B.M, Diaa-Eldin A. Mansour, S. S. Ghoneim, et Nagy I. Elkalashy, "Conditional probability based interpretation of dissolved gas analysis for transformer incipient faults." *IET Generation, Transmission & Distribution* 11.4 (2017): 943-951.