Effect of Data Transformation on the Diagnostic Accuracy of Transformer Faults and the Performance of the Supervised Classifiers

Sherif S. M. Ghoneim*, Ibrahim B. M. Taha*,***, Rizk Fahim**, Saad A. Mohamed Abdelwahab**,****‡

*Department of Electrical Engineering, College of Engineering, Taif University, Taif 29144, Saudi Arabia

** Electrical Department, Faculty of Technology and Education, Suez University, Suez 43527, Egypt

***Department of Electrical Power and Machines Engineering, Faculty of Engineering, Tanta University, Tanta 31521, Egypt

****Department of Computers & Systems Engineering, High Institute of Electronic Engineering, Ministry of Higher Education, Bilbis-Sharqiya 44621, Egypt

(S.ghoneim@tu.edu.sa, i.taha@tu.edu.sa, Rizkfahim16@gmail.com, Saad.Abdelwahab@suezuniv.edu.eg)

‡ Corresponding Author; Saad A. Mohamed Abdelwahab, 43527, Saad.Abdelwahab@suezuniv.edu.eg

Received: 05.04.2022 Accepted:25.05.2022

Abstract- Dissolved gas analysis (DGA) is a common method used to diagnose transformer faults. The DGA methods such as IEC Code, Rogers' ratios, Duval triangle, and key gas methods failed to interpret the transformer faults in some cases and have poor diagnostic accuracy. Therefore, the researchers try to enhance the diagnostic accuracy by combining the traditional DGA techniques with artificial intelligence and optimization techniques. Still, they also have a complex way of interpreting the transformer faults. In the current work, a classification learner toolbox in MATLAB presented several Classifiers to classify the transformer faults and construct a classifier model used to diagnose some other test samples. The classification learner in MATLAB is so easy to understand and implement in classification application. Several data transformations were carried out to investigate their effect on diagnostic accuracy to identify which transformation method can achieve the highest diagnostic accuracy. The results indicated that the ensemble bagged classifier with raw data (data without any transformation) had the highest diagnostic accuracy of the transformer faults, reaching 83.4 %.

Keywords: Dissolved gas analysis, IEC Code, Rogers' ratios, Duval triangle, artificial intelligence and optimization techniques

1. Introduction

Diagnosing the transformer faults in the early stage avoids the loss of continuous operation of the power systems and loss of revenue of the electrical utility [1-4]. Dissolved gas analysis (DGA) is a standard method used to diagnose transformer faults, and it considers the preliminary test carried out to inspect the transformer state [5-7]. Several traditional DGA techniques use to interpret the transformer faults, some of them are based on the fixed rules, such as IEC 60599 Code [8] and the IEEE C57-104 [9], and the others are the graphical DGA methods [10-13]. However, these DGA traditional techniques have poor diagnostic accuracy, and, in many cases, it fails to diagnose the type of transformer faults correctly [14-15].

Recent work focused on enhancing the diagnostic accuracy of traditional and graphical DGA methods by combining them with artificial intelligence. Several artificial intelligence techniques are used with the conventional DGA techniques, such as artificial neural network [16-17], fuzzy logic [18-19], SVM [20-21], KNN [1, 6], and other methods [22-27].

Several researchers addressed the optimization techniques to enhance the diagnostic accuracy of the traditional DGA techniques. Taha et al. [28] proposed a new particle swarm optimization-fuzzy system (PSO-FS) platform to control the ratio limits of Rogers' Four ratios and IEC code DGA methods. The work was based on the ability of PSO to specify the ratio limits and corresponding rules to diagnose the transformer fault types correctly. Hoballah et al. [29] presented an efficient code matrix to diagnose the transformer faults. The FS was used to adjust the rules that

mapped the gas ratio limits for each fault type. Hybrid Grey Wolf Optimization (HGWO) is used to produce the code matrix, limiting the impact of uncertainties on the fault type. Sherif et al. [7] utilized the teaching learning-based optimization (TLBO) to adjust the gases concentration limits and ratio limits based on the work in [30]. An adaptive dynamic polar rose guided Whale optimization algorithm is used to improve the classification techniques' parameters to enhance the diagnostic accuracy of the transformer faults [2]. All the above DGA methods combined with the artificial intelligence or optimization methods were very complex to understand. Therefore, a classification learner tool in MATLAB is used to classify the transformer faults and the trained model to test any other samples. Ben Mahamed et al. [20] enhanced the diagnostic accuracy of the transformer faults using the support vector machine (SVM)-bat (BA) algorithm. The BA algorithm adjusted the model conditioning parameter "l" and penalty parameter "c" of the SVM to develop a maximum diagnostic accuracy rate.

In this research, an Investigation of data normalized effect on the diagnostic accuracy of the transformer faults was developed. Several data manipulation was addressed using several data transformation methods. The main dissolved gases that were used in this study are Hydrogen (H_2), Methane (CH_4), Ethan (C_2H_6), ethylene (C_2H_4), and acetylene (C_2H_2). Six classes of the transformer fault types are proposed, which were labeled and categorized as 1, 2, 3, 4, 5, and 6, referring to partial discharge (PD), low energy discharge (D1), high energy discharge (D2), low thermal (T1), Medium thermal (T2), and high thermal (T3) faults, respectively. Several classifiers from the classification learner tool in MATLAB were considered to select the best one that develops the highest diagnostic accuracy of the transformer fault. A total of 475 samples were used, 386 data samples were used to train the classifiers, and the other 89 samples were used to test the trained classifier. The 386 raw data were used as training to check which classifier will develop the highest accuracy. Therefore, it was used with the data transformation to investigate its effect on the diagnostic accuracy of the transformer faults. The study results indicated that the ensemble bagged tree classifier develops the highest diagnostic accuracy of the transformer faults.

2. Data Collection

The trained and testing data are collected from the chemical laboratory of the holding electricity company in Egypt and literature. The collected samples' distribution was illustrated in Table 1 based on the type of fault and their sources. The data are classified based on the fault types as a total number of 43, 69, 115, 81, 24, and 54 for samples for partial discharge (PD), low energy discharge (D1), High energy discharge (D2), low thermal (T1), Medium thermal (T2), and high thermal (T3), respectively. In addition, there are 240 samples from the chemical laboratory of the holding electricity company in Egypt [23], which are considered field data. In addition, most of the literature data is from [24]. Therefore, there are 89 data samples for testing purposes. Table 2 illustrates their source and fault types. According to fault types, the distribution of the data samples is 8, 13, 19, 13, 7, and 29 for PD, D1, D2, T1, T2, and T3, respectively. The 89 data samples are new samples, which did not consider in the trained data.

Table 1. Distribution of the training samples according to the fault types and the references

Ref.	PD	D1	D2	T1	T2	T3	Total
[31]	27	42	55	70	18	28	240
[32]	9	24	48	0	0	18	99
[33]	3	0	4	4	3	5	19
[34]	1	0	5	2	0	1	9
[35]	0	2	1	1	3	1	8
[36]	1	1	2	1	0	1	6
[37]	2	0	0	3	0	0	5
Total	43	69	115	81	24	54	386

Table 2. Distribution of the testing samples according to the fault types and the references

Ref.	PD	D1	D2	T1	T2	T3	Total
[31]	0	0	2	0	0	0	2
[32]	4	3	4	1	0	0	12
[37]	1	0	0	0	1	0	2
[38]	1	0	0	4	2	14	21
[39]	1	4	2	0	2	8	17
[40]	1	0	4	4	0	2	11
[41]	0	1	4	2	1	0	8
[42]	0	3	1	2	0	1	7
[43]	0	2	2	0	0	0	4
[44]	0	0	0	0	1	2	3
[45]	0	0	0	0	0	2	2
Total	8	13	19	13	7	29	89

3. Data Preparation

The data preparation can be presented as follows,

3.1. Logarithmic transformation

The new dissolved gas concentration can be obtained in the first normalized case considering the logarithmic transformation used to reduce the margin between the samples' datasets. New dissolved gases can be developed as follows:

$$X_n = Log(X_c) \quad (1)$$

X_n refers to the new gas concentration value of the dissolved gases based on the logarithmic transformation, and X_c the row dissolved gas magnitude for each of the main five gases (H_2, CH_4, C_2H_6, C_2H_4, and C_2H_2) for each sample.

2.2. Total dissolved combustion gas (TDCG) transformation

The total dissolved combustion gases are the sum of the main five gases (H_2, CH_4, C_2H_6, C_2H_4, and C_2H_2), and the new normalized gas concentration can be developed by dividing each gas concentration by the TDCG. The total dissolved combustion gases can be presented as (2)

$$TDCG = H_2 + CH_4 + C_2H_6 + C_2H_4 + C_2H_2 \quad (2)$$

Equation 3 can compute the new normalized gas concentration as

$$X_n = \frac{X_c}{TDCG} \quad (3)$$

3.3. Data transformation based on the minimum and maximum of all sample data

The normalized gas concentration data can be introduced by subtracting the gas concentration of each sample from the minimum value of all data concentrations and dividing the results by subtracting the minimum value of all gas concentrations from the maximum value of all gas concentrations in the sample space [46].

$$X_n = \frac{X_c - Min(H_2, CH_4, C_2H_6, C_2H_4, \text{and } C_2H_2)}{Max(H_2, CH_4, C_2H_6, C_2H_4, \text{and } C_2H_2) - Min(H_2, CH_4, C_2H_6, C_2H_4, \text{and } C_2H_2)} \quad (4)$$

where Min((H_2, CH_4, C_2H_6, C_2H_4, and C_2H_2)) and Max((H_2, CH_4, C_2H_6, C_2H_4, and C_2H_2)) refer to the lowest and largest values of all gases' concentration in the sample space, respectively.

3.4. Data transformation based on the mean and standard deviation of each dissolved gas

In this case of data transformation, each gas concentration is subtracted from the mean of the corresponding gas concentration data (i.e., each of H_2 gas concentration and the mean of all H_2 gas concentrations). The new gas concentration can be determined as in (5),

$$X_n = \frac{X_c - \mu}{\sigma} \quad (5)$$

where μ and s are the mean and standard deviation of each gas's main five combustible gases in each dataset sample, respectively.

3.5. Data Transformation based on the maximum value of each gas of the five main gases

Dividing each gas concentration by the corresponding maximum value of the gas data gave new gas' concentrations, i.e., all H_2 concentration is divided by the maximum value of the H_2 gas of all data sample. As a result, the new transformation ratios can be developed as follows:

$$X_n = \frac{X_c(H_2, CH_4, C_2H_6, C_2H_4, \text{and } C_2H_2)}{Max(H_2, CH_4, C_2H_6, C_2H_4, \text{and } C_2H_2)} \quad (6)$$

4. Classification Techniques

This section addressed several classification techniques, which gave transformer faults the highest diagnostic accuracy. These classification techniques included ensemble bagged tree; Ensemble boosted tree, Ensemble RUSBoosted tree, Weighted KNN, quadratic SVM, and linear SVM.

4.1 Ensemble bagged tree

Ensemble methods utilized multiple learning algorithms to enhance the predictive performance rather than the constituent learning algorithm alone [47]. A set of finite models constitutes machine learning ensembles, allowing more flexible structures between these models. For example, an ensemble bagging tree is a parallel learning method used for a two-class processing model, multiple classifiers, and regression [48-50]. As in Fig. 1., the ensemble bagging tree classifier can be constructed via preparing the data samples as training data in a random form. The decision tree is trained to get the basic model with or without the cross-validation, and the voting method is used to combine every basic model.

The k-fold cross validation depends on dividing the trained data to groups, for examples (5 groups), then the first four groups were taken to train the model and the last group is for validation. In the second round the last one was added to the four groups and extract another group of data to use it as a validation group and so on. At the end of this process, the average of the classification accuracy of the 5 rounds was computed and considered the accuracy of the model. This process is carried out by the classification learner toolbox without any action of the user.

A voting ensemble method combines the predictions from different models to get the final decision, where the models should be different because it utilizes all training data for train purpose. A main concept of voting is generalizing better by compensating for the errors of individual model separately.

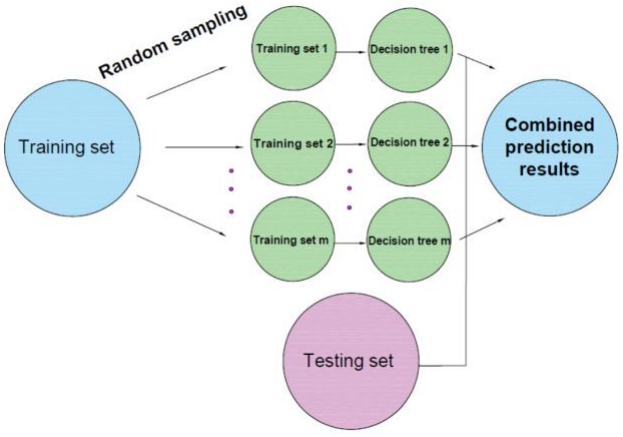

Fig. 1 flowchart of ensemble bagging tree

4.2 Ensemble boosted tree

Ensemble boost-tree algorithm is derived from deterministic gradient boosting's fundamental numerical tools, which can use for classification and prediction purposes [51]. The boosting model is used to boost the performance of a weaker classifier rather than random guessing [52]. Figure 2 illustrates the boosting tree that generates only one model improved every time to reduce classification errors. The boosting models developed with few errors can be overfitting because they modified themselves [53].

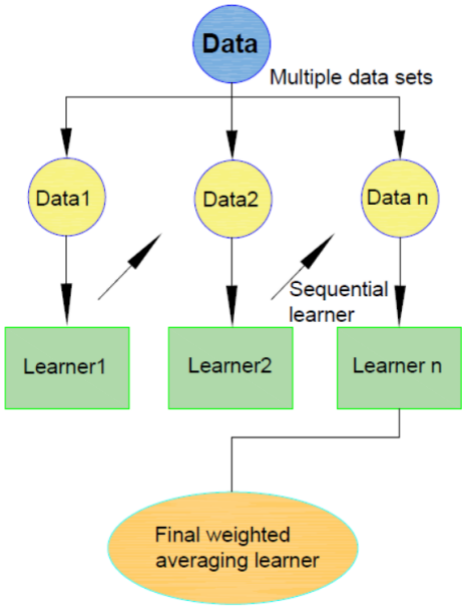

Fig. 2. Ensemble boosting tree flowchart

4.3 Ensemble RUSBoosted tree Algorithm

RUSBoost tree algorithm is a combination of Boosting algorithm and under-sampling method. It divides the dataset into K parts. The initial weight of the sub-data set is the random weight of the total number divided by K. Then, the sub-data set is trained, and regularization is used to update the weights. Finally, a sub-data set repeats the training classifier was meeting the required conditions to select the best model [54-55].

4.4 Weighted K-Nearest Neighbors (K.N.N.)

K-Nearest Neighbours (K.N.N.) considers one of the common classifiers of supervised machine learning. It uses to classify data input into pre-defined classes (k). First, the Euclidean distance function between pre-defined classes and each varying sample is computed, then KNN selects the minimum nearest neighbors according to each class. Finally, each sample assigns their class based on the nearest K neighbors [56-57]. The Weighted kNN is an updated version of KNN Several issues influence the performance of the kNN algorithm relating to the choice of the hyperparameter k. For example, smaller k results in more sensitivity to outliers, and larger k leads to many points from other classes, including neighbors. Figure 3 depicts the operation of KNN for separating among different classes of the data.

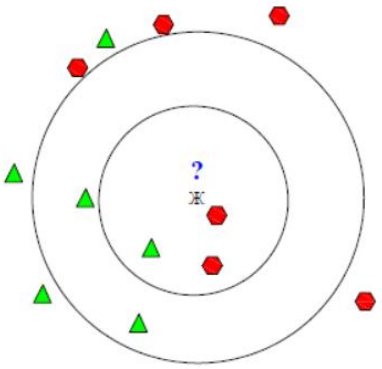

Fig.3. KNN classifier based on k-parts

4.5 Support Vector Machine (SVM)

It is a machine learning tool using to separate the data into two-class of data via a hyperplane. This hyperplane must achieve the maximum distance between the points of each class; then, accurate classifying can occur. If any point lay outside the hyperplane margin, it belongs to a different class. Greater features lead to more difficult to separate among different classes. Figure 4 illustrates the margin condition of SVM. A good classification can take place when a large margin exists [58-59].

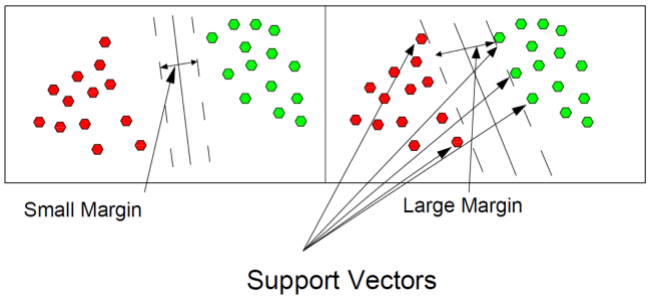

Fig. 4. SVM algorithm indicating the margin separating two classes

5. Results and Discussions

In this section, the results of the diagnostic accuracy of the classification learners were reported. The diagnostic accuracy of each classifier was explained with and without transformation of the raw data. The main contribution of the current work is, which one of the data transformation will give the highest accuracy with the classifiers.

In the MATLAB learner toolbox, the input and output data is identified in the workspace. The classification-learner command must be written in command page, then the classification toolbox is appeared. A new session was developed, then the input and output variables were identified. The learner assumed that the last column of the data is the output. The k-fold cross-validation was used as 10, which leads to construct stable classification model. The all learners were selected and then the learner was run to get the classification accuracy for each classifier as in Fig. 5.

When all of the classifications learners were used with the data whether raw or normalized data, the results revealed that the Ensemble tree classifier developed the highest diagnostic accuracy of the transformer faults. Therefore, to reduce the explanation of all classifications' results, only Ensemble bagged tree was shown. The results in Table 3 illustrates that the ensemble bagged tree develops the highest diagnostic accuracy (81.6%) of the trained data (386 Samples). Hence, the ensemble bagged tree will use as the best classifier in this case study.

Table 3. Comparison between all of classifiers based on the raw data

Classifier	Diagnostic accuracy % based on raw data
fine tree	72.5
medium tree	86.4
coarse tree	59.3
linear discriminant	33.4
quadratic discriminant	47.7
linear SVM	66.6
quadratic SVM	56.5
cubic SVM	39.1
fine Gaussian SVM	54.4
medium Gaussian SVM	46.9
coarse Gaussian SVM	35.8
fine KNN	74.6
medium KNN	71.2
coarse KNN	51
cosine KNN	68.9
cubic KNN	70.2
weighted KNN	77.2
ensemble boosted trees	80.1
ensemble bagged trees	83.4
ensemble subspace Discriminant	33.4
ensemble subspace KNN	73.3
ensemble RUSBoosted trees	70.8

In this manuscript the results of the classifier, which develops the highest diagnostic accuracy of the transformer faults were presented. To present the all results of all classifiers that developed highest accuracy, it need more and more pages, the idea of identifying the results were explained by the classifier that developed highest diagnostic accuracy (Ensemble bagged tree).

5.1 Ensemble bagged tree results for data without normalization

The raw data of the dissolved gases was taken as the first case in this study to compare its results with the normalized dissolved gases data.

The ensemble bagged tree developed the highest accuracy classification when the raw data was used based on the results of Table 3. The classifying accuracy was 83.4% when considering 5/5 features. The Five features refer to the number of input parameters, which are the five main dissolved gases (H_2, CH_4, C_2H_6, C_2H_4, and C_2H_2) in the current study. Figure 5 illustrates the distribution of the data samples as a scatter plot of the ensemble bagged tree. In the scatter plot, the name of the trained file appears as input_output_0, the number of observations (386 data samples). The number of predictors is 5 related to the five main gases. The six column refers to the output (transformer fault type), the fault types are categorized to 6 classes, and the cross-validation is 10.

The correct diagnosis point is shown in Fig. 5 as a color circle, and the incorrect diagnosis point is color x. The x-axis (column 1) and y-axis (column 2) refer to the predictors of dissolved gases. It also showed the number of the correct and incorrect observations in the confusion matrix in Fig. 6. Figure 6a illustrates that the number of observations expressing PD is 43 samples. The ensemble bagged tree classifier correctly diagnoses 38 samples and incorrectly diagnoses five samples (two of them diagnose as D2 and the other three are T1). The correct diagnosis is in green color, and the red color refers to the incorrect diagnosis. The correct diagnose of each fault as in confusion matrix of Fig. 6a is 38 samples from a total number of 43 samples for PD, 49 samples out of 69 samples for D1, 105 samples out of 115 samples of D2, 74 samples out of 83 samples of T1, 12 samples out of 23 samples of T2, and 44 samples out of 54 samples of T3. Fig. 6b indicated the classifier accuracy percentage of each fault type where 88 % is the accuracy percentage of correct diagnose of PD (38/43). The highest diagnostic accuracies are 91% for D2 and T1, and the lowest diagnostic accuracy is 50% for T2. The positive predictive values and negative predictive values were indicated in Fig. 7. Figure 7 illustrates that the predicted class 1 referring to PD appeared 40 times, 38 times is correct with the accuracy of 95% for PD, and incorrect predict two times one with actual fault T1 and another one for T2 (3% for each T1 and T2).

Similarly, class 6 referring to a high thermal fault (T3) appeared 51 times in prediction, 44 times for correct diagnosis with the diagnostic accuracy of 86%. There are seven times incorrect diagnoses with incorrect interpretations (1 time for D2, 3 times for T1, 3 times for T2) with the false diagnoses 2, 6, and 6%, respectively. Figure 8 can indicate the receiver operating characteristic (ROC). The marker on ROC depicts the current classifier performance where the false positive rate (FPR) is on the x-axis, and the true positive rate (TPR) is on the y-axis. Figure 8 illustrates that the FPR is 0.01, which indicates that 1% of the observations were assigned incorrectly to the positive class. The TPR is 0.88, referring that the classifier assigns 88% of the

observations correctly to the positive class. When the ROC curve gives the right angle, it means perfect classifying results were obtained, and when it makes 45°, it refers to a poor classification result. The area can measure the overall accuracy of the classifier under the curve (AUC). The greater the AUC, the higher the classifier accuracy. Figure 8 explained that the AUC is 98%, referring to better classifier performance. Table 4 compares between the classifiers for diagnosing the transformer faults that were developed highest accuracies based on the raw data. The results in Table 4 explains that the ensemble bagged tree develops the highest diagnosing accuracy of the transformer faults with 83.4 % (322/386). The diagnostic accuracies of the other three classifiers (Ensemble boosted tree, ensemble RUSBoosted tree, and weighted KNN) are 80.1, 78, and 77.2%, respectively, which are considered the highest diagnostic accuracy beyond the ensemble bagged tree classifier. The diagnostic accuracy of individual transformer fault types can be illustrated in detail in Table 4, computed by dividing the number of correct diagnoses per the total number of observations. The diagnostic accuracy of PD was 88 (38/43), 81(35/43), 93(40/43), and 67% (29/43) for ensemble bagged tree, ensemble boosted tree, ensemble RUSBoosted tree, and weighted KNN, respectively. Similarly, the diagnostic accuracy for high T3 was 81 (44/54), 78(42/54), 70(38/54), and 89% (48/54) for ensemble bagged tree, ensemble boosted tree, ensemble RUSBoosted tree, and weighted KNN, respectively.

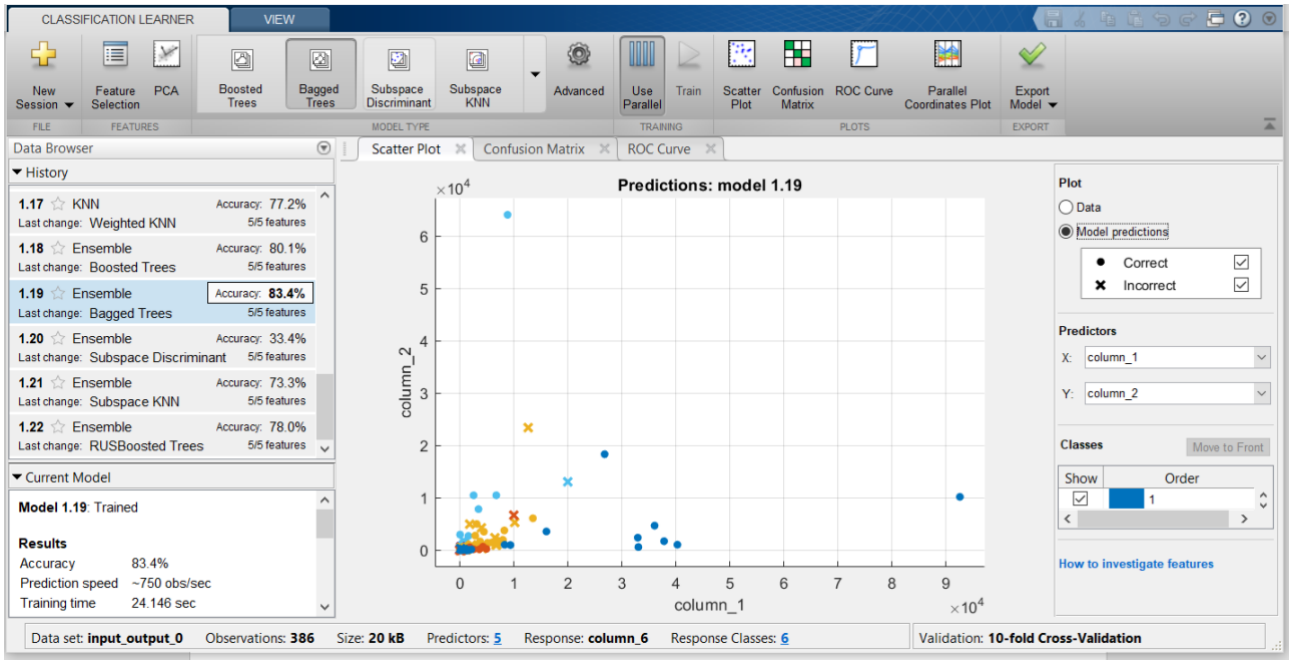

Fig. 5. Scatter plot of the ensemble bagged tree

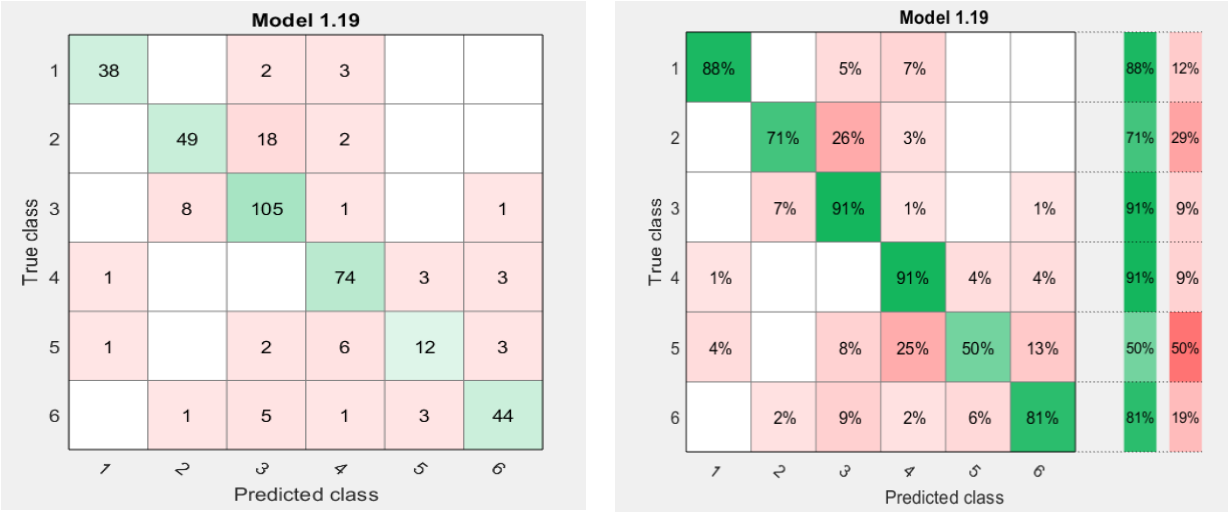

Fig. 6. Confusion matrix a) the number of correct and incorrect observations, b) true positive rates-false negative rates for ensemble bagged tree

Table 3. Comparison of different classifiers' accuracies for classifying the trained data without normalization

Fault type/Classifier accuracy	Ensemble bagged tree	Ensemble boosted tree	Ensemble RUSBoosted tree	Weighted KNN
PD	38/43 (88%)	35/43 (81%)	40/43 (93%)	29/43 (67%)
D1	49/69 (71%)	43/69 (62%)	51/69 (74%)	40/69 (58%)
D2	105/115 (91%)	104/115 (90%)	91/115 (79%)	91/115 (79%)
T1	74/81 (91%)	72/81 (89%)	65/81 (80%)	75/81 (93%)
T2	12/24 (50%)	13/24 (54%)	16/24 (67%)	15/24 (63%)
T3	44/54 (81%)	42/54 (78%)	38/54 (70%)	48/54 (89%)
Overall	322/386 (83.4%)	311/386 (80.1%)	301/386 (78%)	298/386 (77.2%)

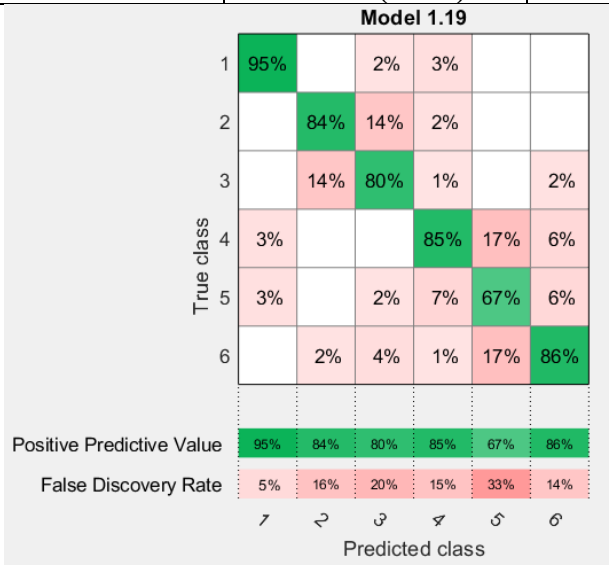

Fig. 7. Positive predictive values-negative predictive values confusion matrix of the trained sample using ensemble bagged tree

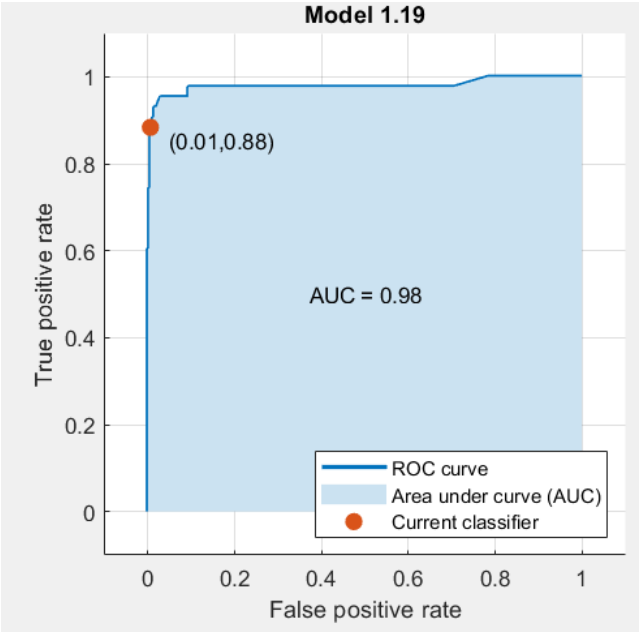

Fig. 8. The ROC result of ensemble bagged tree

Table 5 depicts the accuracy result of different fault types of the constructed model of ensemble bagged tree classifier based on new 89 data samples, which did not include in the trained data (386 samples). The diagnostic accuracy of each fault was as 87.5 % (7 corrected diagnose sample/8 observations of PD), 30.77% for D1, 52.63 for D2, 84.61% for T1, 42.86% for T2, and 72.41% for T3. The overall diagnostic accuracy was 62.91% (corrected diagnosis sample/total tested samples =56/89).

Table 5. The accuracy of the prediction of 89 samples is 62.92% (56/89) based on ensemble bagged trees

Fault type	ACCURACY %
PD	(7/8) 87.5
D1	(4/13) 30.77
D2	(10/19) 52.63
T1	(11/13) 84.61
T2	(3/7) 42.86
T3	(21/29) 72.41
Overall accuracy	(56/89) 62.92

5.2 Data with normalization

Several transformations of the trained data were carried out to investigate their effect on the diagnostic accuracy of the classifiers. The normalization of the data was taken place using five forms. The first form is taken the log of each gas concentration as in (1), the second normalization form was obtained, dividing every gas concentration in each sample by the total dissolved combustion gas of this sample as in (3), the third form was as in (4). The third form of data normalization was developed using each gas concentration column's mean and standard deviation as in (5). The final normalized form of the data can be obtained by dividing each gas concentration by the maximum value of the corresponding gas in (6). All the normalized data developed overall diagnostic accuracy lower than that developed with the raw data where the maximum diagnostic accuracy is from Eqn. (1) (Log (x)) as 82.9 %, but the raw data developed a diagnostic accuracy of 83.4%. It is seen from Table 6 that the diagnostic accuracy of fault type T2 was very poor, which did not exceed 55% and has an adverse effect on the overall diagnostic accuracy of the ensemble bagged tree classifier with the transformation data.

Table 7 illustrates the results of the diagnostic accuracies of the ensemble bagged tree classifier of the constructed model based on a total of 89 data samples (testing data), which were not trained by the constructed model.

Table 6 explained the diagnostic accuracy among different classifiers that developed the highest diagnostic accuracy with the trained data.

	Ensemble bagged tree						
Data state	PD.	D1	D2	T1	T2	T3	Overall accuracy
Eqn. (1)	95% (41/43)	68% (47/69)	94% (108/115)	(88%) 71/81	50% (12/24)	76% (41/54)	82.9%
Eqn. (3)	95% (41/43)	70% (48/69)	91% (105/115)	90% (73/81)	46% (11/24)	67% (36/54)	81.3%
Eqn. (4)	86% (37/43)	70% (48/69)	90% (104/115)	89% (72/81)	54% (13/24)	71% (39/54)	81.1%
Eqn. (5)	95% (41/43)	65% (45/69)	88% (101/115)	91% (74/81)	46% (11/24)	74% (40/54)	80.8%
Eqn. (6)	91% (39/43)	71% (49/69)	90% (104/115)	88% (71/81)	50% (12/24)	74% (40/54)	81.6%

Table 7. An Effect of data normalization method on the diagnostic accuracy of ensemble bagged tree classifier based on new 89 data samples using as testing data

	Ensemble bagged tree						
Data state	PD.	D1	D2	T1	T2	T3	Overall accuracy
Eqn. (1)	87.5	38.46	68.42	84.61	14.28	86.2	69.66% (62/89)
Eqn. (3)	25	0	15.79	100	0	6.9	22.47% (20/89)
Eqn. (4)	87.5	38.46	68.42	84.62	14.29	89.66	70.87% (63/89)
Eqn. (5)	87.5	38.46	63.16	84.62	42.86	72.41	66.29% (59/89)
Eqn. (6)	87.5	30.77	68.42	84.62	57.14	62.07	64.04% (57/89)

In this current work, different data transformation can be used as in Equations (1), (2), (4), (5), and (6) to enhance the diagnostic accuracy of the transformer faults. The data develop by equations (1) to (6) was used with ensemble bagged tree and compute the diagnostic accuracy of the classifier to investigate if the diagnostic accuracy was enhanced or not. Based on the diagnostic accuracy of the testing samples, it is obvious that the diagnostic accuracy of the testing samples (89) samples increased from 62.92% to 70.87% using Eq. (4), therefore, the data transformation based on Eq. (4) enhanced the diagnostic accuracy rather than that in case of raw data.

The results of Table 7 indicated that Eqn. (4) developed the best diagnostic accuracy of the testing samples (70.87%). On the other hand, equation (3) developed the worst diagnostic accuracy as 22.47%, then the data normalization based on Eqn. (3) can be ignored. It was evident from Table 7 that the diagnostic accuracy of D1 and T2 cause low overall accuracies of all data normalization methods. It is attributed to the interfaces between D1 and D2 and T2 and T3.

For ensemble bagged tree, the data transformation for the current study was more efficient with the testing samples than that with the raw data, although the diagnostic accuracy of raw data is higher than that with the data transformation for the training samples. The overall accuracy of ensemble bagged tree based on raw data for testing samples was 62.92%, but in case of data transformation the diagnostic accuracy was enhanced to 70.87 with Eq. (4). Therefore, the performance of the constructed ensemble bagged tree classifier based on Eq. (4) was better than using the ensemble bagged tree with raw data. The strength of the constructed model depends on its ability for correct diagnose with the test samples.

6. Conclusions

The traditional DGA techniques develop poor diagnostic accuracy of transformer faults. The classification learner toolbox in MATLAB presented several data classifiers to classification applications. In the current work it used the MATLAB classifier to identify the transformer fault types (PD, D1, D2, T1, T2, and T3) based on the concentration of the dissolved gases such as (H_2, CH_4, C_2H_6, C_2H_4, and C_2H_2). Based on the classifiers' results with the raw data the ensemble bagged tree developed the highest diagnostic accuracy of the transformer faults, which is divided to six classes (PD, D1, D2, T1, T2, and T3). Data transformation was utilized to study its effect on the ensemble bagged tree classifiers' diagnostic accuracy and performance. This study summarized that the ensemble bagged tree classifier with the raw trained data developed the highest diagnostic accuracy of the transformer faults (83.4%), but the diagnostic accuracy of the test data is poor 62.92%. Therefore, the data transformation was used to enhance the diagnostic accuracy of the transformer faults based on several data transformation as in eqns. (1) to (6). The tested samples explained the data transformation based on Eqn. (4) presented the highest diagnostic accuracy with the ensemble bagged tree classifier (70.87%). Therefore, data transformation can be used to enhance the performance of the classifiers.

Acknowledgements

The authors would like to acknowledge the financial support received from Taif University Researchers Supporting Project Number (TURSP-2020/34), Taif University, Taif, Saudi Arabia.

References

[1] O. Kharif, Y. Benmahamed, M. Teguar, A. Boubakeur, and S. S. M. Ghoneim, "Accuracy Improvement of Power Transformer Faults Diagnostic Using KNN Classifier With Decision Tree Principle", IEEE Access, VOLUME 9, 2021, pp. 81693- 81701.

[2] S. S. M. Ghoneim, T. A. Farrag, A. A. Rashed, E. M. El-Kenawy, And A. Ibrahim, "Adaptive Dynamic Meta-Heuristics for Feature Selection and Classification in Diagnostic Accuracy of Transformer Faults", IEEE Access, VOLUME 9, 2021, pp. 78324- 78340.

[3] M. Rashidi, A. Bani-Ahmed, R. Nasiri, A. Mazaheri and A. Nasiri, "Design and implementation of a multi winding high frequency transformer for MPSST application," 2017 IEEE 6th International Conference on Renewable Energy Research and Applications (ICRERA), 2017, pp. 491-494.

[4] F. B. Gurbuz, R. Bayindir and S. Vadi, "Comprehensive Non-Intrusive Load Monitoring Process: Device Event Detection, Device Feature Extraction and Device Identification Using KNN, Random Forest and Decision Tree," 2021 10th International Conference on Renewable Energy Research and Application (ICRERA), pp. 447-452, 2021.

[5] I. B. M. Taha, D.-E.-A. Mansour, S. S. M. Ghoneim, and N. I. Elkalashy, "Conditional probability-based interpretation of dissolved gas analysis for incipient transformer faults," IET Gener., Transmiss. Distrib., vol. 11, no. 4, pp. 943_951, Mar. 2017, DOI: 10.1049/iet-gtd.2016.0886.

[6] Y. Benmahamed, M. Teguar, and A. Boubakeur, "Application of SVM and KNN to Duval pentagon for transformer oil diagnosis," IEEE Trans. Dielectr. Electr. Insul., vol. 24, no. 6, pp. 3443_3451, Dec. 2017, doi:10.1109/tdei.2017.006841.

[7] S. S. M. Ghoneim, K. Mahmoud, M. Lehtonen, and M. M. F. Darwish, "Enhancing Diagnostic Accuracy of Transformer Faults Using Teaching-learning-Based Optimization", IEEE Access, VOLUME 9, 2021, pp. 30817- 30832.

[8] Mineral oil-filled electrical equipment in service — Guidance on the, I. 60599 E. Analysis, interpretation of dissolved and free gases, s, and 2.1, "Edition 2.1," 2007.

[9] IEEE Guide for the Interpretation of Gases Generated in Mineral Oil-Immersed Transformers, IEEE Standard C57.104-2019 (Revision IEEE Std C57.104-2008), Nov. 2019, pp. 1_98.

[10] D.-E.-A. Mansour, "Development of a new graphical technique for dissolved gas analysis in power transformers based on the five combustible gases," IEEE Trans. Dielectr. Electr. Insul., vol. 22, no. 5, pp. 2507_2512, Oct. 2015.

[11] M. Duval and L. Lamarre, "The Duval pentagon_A new complementary tool for the interpretation of dissolved gas analysis in transformers," IEEE Elect. Insul. Mag., vol. 30, no. 6, pp. 9_12, Nov. 2014.

[12] O. E. Gouda, S. H. El-Hoshy, and H. H. El-Tamaly, "Proposed heptagon graph for DGA interpretation of oil transformers," IET Gener., Transmiss. Distrib., vol. 12, no. 2, pp. 490_498, Jan. 2018.

[13] M. M. Emara, G. D. Peppas, I. F. Gonos, "Two Graphical Shapes Based on DGA for Power Transformer Fault Types Discrimination", IEEE Transactions on Dielectrics and Electrical Insulation, Volume: 28, Issue: 5, pp. 1703 – 1712, 2021.

[14] N. Haque, A. Jamshed, K. Chatterjee, S. Chatterjee, "Accurate Sensing of Power Transformer Faults From Dissolved Gas Data Using Random Forest Classifier Aided by Data Clustering Method", IEEE Sensors Journal, Volume: 22, Issue: 6, pp. 5902 – 5910, 2022.

[15] S. A M. Abdelwahab, A. M. Yousef, M. Ebeed, Farag K. Abo-Elyousr1, A. Elnozohy, M. Mohamed "Optimization of PID Controller for Hybrid Renewable Energy System Using Adaptive Sine Cosine Algorithm," International Journal of Renewable Energy Research-IJRER, pp 670-677, vol 10, no 2, 2020.

[16] S. S. M. Ghoneim, I. B. M. Taha, and N. I. Elkalashy, "Integrated ANN-based proactive fault diagnostic scheme for power transformers using dissolved gas analysis," IEEE Trans. Dielectr. Electr. Insul., vol. 23, no. 3, pp. 1838_1845, Jun. 2016.

[17] M. D. Equbal, S. A. Khan, and T. Islam, "Transformer incipient fault diagnosis on the basis of energy-weighted DGA using an artificial neural network," Turkish J. Electr. Eng. Comput. Sci., vol. 26, pp. 77_88, Jan. 2018.

[18] A. Harrouz, K. Nourdine, K. Kayisli, H. I. Bulbul and I. Colak, "A Fuzzy Controller for Stabilization of Asynchronous Machine," 2018 7th International Conference on Renewable Energy Research and Applications (ICRERA), 2018, pp. 1369-1373.

[19] M. Noori, R. Effatnejad, and P. Hajihosseini, ``Using dissolved gas analysis results to detect and isolate the internal faults of power transformers by applying a fuzzy logic method," IET Gener., Transmiss. Distrib., vol. 11, no. 10, pp. 2721_2729, Jul. 2017.

[20] Y. Benmahamed, O. Kherif, M. Teguar, A. Boubakeur and S. S. M. Ghoneim, "Accuracy Improvement of Transformer Faults Diagnostic Based on DGA Data Using SVM-BA Classifier", Energies, 14, 2021.

[21] M. Beken, B. Hangun and O. Eyecioglu, "Classification of Turkey among European Countries by Years in Terms of Energy Efficiency, Total Renewable Energy, Energy Consumption, Greenhouse Gas Emission and Energy Import Dependency by Using Machine

Learning," 2019 8th International Conference on Renewable Energy Research and Applications (ICRERA), pp. 951-956, 2019.

[22] Md M. Islam, G. Lee and S. N. Hettiwatte, "Application of Parzen Window estimation for incipient fault diagnosis in power Transformers," IET High Voltage, vol. 3, no. 4, pp. 303–309, 2018.

[23] J. T. Hu, L. X. Zhou and ML Song, "Transformer Fault Diagnosis Method of Gas Hromatographic Analysis Using Computer image analysis," Int. Conf. Intelligent Sys. Design and Eng. App., Jan. 2012, pp. 1169–1172.

[24] I.B.M. Taha, S.S Dessouky, S.S.M. Ghoneim, "Transformer fault types and severity class prediction based on neural pattern-recognition techniques" Electric Power Systems Research 191 (2021) 106899.

[25] J. Li, Genxu Li, Chen Hai, M. Guo, "Transformer Fault Diagnosis Based on Multi-Class AdaBoost Algorithm", IEEE Access (Volume: 10), pp. 1522 – 1532, 2022.

[26] I. B. M. Taha, S. Ibrahim, D. A. Mansour, "Power Transformer Fault Diagnosis Based on DGA Using a Convolutional Neural Network With Noise in Measurements", IEEE Access (Volume: 9), pp. 111162 – 111170, 2021.

[27] A. G. C. Menezes, M. M. Araujo,O. M. Almeida, F. R. Barbosa, A. P. S. Braga, "Induction of Decision Trees to Diagnose Incipient Faults in Power Transformers", IEEE Transactions on Dielectrics and Electrical Insulation, Volume: 29, Issue: 1, pp. 279 – 286, 2022.

[28] I. Taha, A. Hoballah, S. Ghoneim "Optimal Ratio Limits of Rogers' Four-Ratios and IEC 60599 Code Methods Using Particle Swarm Optimization Fuzzy-Logic Approach", IEEE Trans. Dielectr. Electr. Insul., vol. 27, no. 1, Feb. 2020, pp. 222-230.

[29] A. Hoballah, D. A. Mansour, I. B. M. Taha, "Hybrid Grey Wolf Optimizer for Transformer Fault Diagnosis Using Dissolved Gases Considering Uncertainty in Measurements", IEEE Access (Volume: 8), pp. 139176 – 139187, 2020.

[30] S. S. M. Ghoneim, I. B. M. Taha, "A New Approach of DGA Interpretation Technique for Transformer Fault Diagnosis", International Journal of Electrical Power and Energy Systems, 81, pp. 265–274, Oct. 2016.

[31] Egyptian Electricity Holding Company (EEHC.) Reports, 1991-2016.

[32] M. Duval and A. DePablo, "Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases," IEEE Electr. Insul. Mag., vol. 17, no. 2, pp. 31–41, Mar. 2001.

[33] M.-H. Wang, "A novel extension method for transformer fault diagnosis," IEEE Trans. on Power Del., 18(1), pp. 164–169, 2003.

[34] A Sanjay, A. K. Chandel "Transformer incipient fault diagnosis based on probabilistic neural network. In: 2012 students conference on engineering and systems (SCES); p. 15, March 2012.

[35] K. Bacha, S. Souahlia, and M. Gossa, "Power transformer fault diagnosis based on dissolved gas analysis by support vector machine," Electr. Power Syst. Res., vol. 83, no. 1, pp. 73–79, Feb. 2012.

[36] M. Wang, Y. Zhu, F. Wang, L. Geng, "Transformer fault diagnosis based on naive Bayesian classifier and SVR", IEEE Region 10 conference (TENCON 2006); November 2006. p. 1–4.

[37] M. Duval, "A review of faults detectable by gas-in-oil analysis in transformers," IEEE Electrical Insulation Magazine, vol. 18, no. 3. pp. 8–17, May-2002.

[38] J. Khelil, K. Khelil, M. Ramdani and N. Boutasseta, "Bearing Faults Diagnosis Using Discrete Wavelets and Artificial Intelligence Approaches," 2019 1st International Conference on Sustainable Renewable Energy Systems and Applications (ICSRESA), 2019, pp. 1-7

[39] D. Sarma G.N.S. Kalyani, "ANN approach for condition monitoring of power transformers using DGA", IEEE Reg. 10 Conf. (TENCON), 2004, pp.444–447.

[40] S. Seifeddine, B. Khamis and C. Abdelkader, "Power transformer fault diagnosis based on dissolved gas analysis by artificial neural network," 1st Int. Conf. Renewable Energies and Vehicular Technology (REVET), pp. 230–236, 2012.

[41] OE Gouda, S. Salem and S. H. El-Hoshy, "Power transformer incipient faults diagnosis based on dissolved gas analysis," TELKOMNIKA Indonesian J. Electr. Eng., vol. 17, no. 1, pp. 10–16, Jan 2016.

[42] Z. Qiaogen, K. Wang and Y. Zhang, "Optimal dissolved gas ratios selected by genetic algorithm for power transformer fault diagnosis based on Support vector machine," IEEE Trans. Dielectr. Electr. Insul, vol. 23, no. pp. 1198–1206, Apr 2016.

[43] J. T. Hu, L. X. Zhou and M. Song, "Transformer Fault Diagnosis Method of Gas Hromatographic Analysis Using Computer image analysis," Int. Conf. Intelligent Sys. Design and Eng. App., pp. 1169–1172J, an. 2012.

[44] M. Rajabimendi and E.P. Dadios, "A hybrid algorithm based on neural fuzzy system for interpretation of dissolved gas analysis in power transformers," IEEE Reg. 10 Conf. (TENCON), 2012, Nov. 2012.

[45] R. Soni and K. R. Chaudhari, "A Novel Proposed Model to Diagnose Incipient Faults of Power Transformer Using Dissolved Gas Analysis by Ratio methods," Int. Conf. on Comp. of Power, Energy, Information and Communication, April 2015.

[46] I. B.M. Taha and D. A. Mansour, "Novel Power Transformer Fault Diagnosis Using Optimized Machine

Learning Methods", Intelligent Automation & Soft Computing, Vol.28, No.3, pp. 739-752, 2021.

[47] T.G. Dietterich, Ensemble learning, The Handbook of Brain Theory and Neural Networks, 2 2002, pp. 110–125.

[48] T.G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization", Mach. Learn. 40 (2000) 139–157.

[49] S. Zhanga , Q. Yub, H. Heb, F. Zhua, P. Wua, L. Gua, S. Jiangc "iDHS-DSAMS: Identifying DNase I hypersensitive sites based on the dinucleotide property matrix and ensemble bagged tree", Genomics, pp. 1-8, 2019.

[50] S. S. M. Ghoneim, "Determination of Transformers' Insulating Paper State Based on Classification Techniques", Processes 9, 427, 2021.

[51] W. Chen, X. Lei, R Chakrabortty, S. C. Pal, M. Sahana, S. Janizadeh, "Evaluation of different boosting ensemble machine learning models and novel deep learning and boosting framework for head-cut gully erosion susceptibility", Journal of Environmental Management, pp.1-15, 2021.

[52] Hassan, A.N.; El-Hag, A. Two-Layer Ensemble-Based Soft Voting Classifier for Transformer Oil Interfacial Tension Prediction. Energies, 13, 1735, 2020.

[53] J. Lee, W. Wang, F. Harrou, Y. Sun, "Reliable solar irradiance prediction using ensemble learning-based models: A comparative study", Energy Conversion and Management, 112582, pp. 1-13, 2020.

[54] M. Adil, N. Javaid, U. Qasim, et al. LSTM and Bat-Based RUSBoost Approach for Electricity Theft Detection", 10 (12), 2020.

[55] S. Rao, K. Li, J. Wu and Z. Mu, "Application of Ensemble Learning in EEG Signal Analysis of Fatigue Driving", Journal of Physics: Conference Series 1744 (2021) 042193, MACE 2020.

[56] S. Zhang, X. Li, M. Zong, X. Zhu, R. Wang, " Efficient kNN Classification With Different Numbers of Nearest Neighbors", IEEE,29:1-12, 2017.

[57] A. Ali, M. Alrubei, L F. M. Hassan, M. Al-Ja'afari, and Saif Abdulwahed, "Diabetes Classification Based on KNN", IIUM Engineering Journal, Vol. 21, No. 1, 2020, pp. 175-181.

[58] Y. Benmahamed, O. Kherif, M. Teguar, Ahmed Boubakeur and S. S. M. Ghoneim, "Accuracy Improvement of Transformer Faults Diagnostic Based on DGA Data Using SVM-BA Classifier", Energies 2021, 14, 2970.

[59] Y. Zhang, X. Fan, J. Liu, H. Zhang, "Moisture Prediction of Transformer Oil-Immersed Polymer Insulation by Applying a Support Vector Machine Combined with a Genetic Algorithm", Polymers 2020, 12, 1579.