Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020 Doi Number

Accuracy Improvement of Power Transformer Faults Diagnostic using KNN Classifier with Decision Tree Principle

O. Kherif[1], Member, IEEE, Y. Benmahamed[2], M. Teguar[2], A. Boubakeur[2], Senior Member, IEEE, and Sherif S. M. Ghoneim[3], Senior Member, IEEE

[1] Advanced High Voltage Engineering Research Centre, School of Engineering, Cardiff University, Cardiff, CF24 3AA, UK
[2]Laboratoire de Recherche en Electrotechnique, Ecole Nationale Polytechnique (ENP), B.P 182, El-Harrach, 16200 Algiers, Algeria
[3]Electrical Engineering Department, College of Engineering, Taif University, Taif, 21944, Saudi Arabia

Corresponding author: youcef.benmahamed@g.enp.edu.dz

ABSTRACT Dissolved gas analysis (DGA) is the standard technique to diagnose the fault types of oil-immersed power transformers. Various traditional DGA methods have been employed to detect the transformer faults, but their accuracies were mostly poor. In this light, the current work aims to improve the diagnostic accuracy of power transformer faults using artificial intelligence. A KNN algorithm is combined with the decision tree principle as an improved DGA diagnostic tool. A total of 501 dataset samples are used to train and test the proposed model. Based on the number of correct detections, the neighbor's number and distance type of the KNN algorithm are optimized in order to improve the classifier's accuracy rate. For each fault, indeed, several input vectors are assessed to select the most appropriate one for the classifier's corresponding layer, increasing the overall diagnostic accuracy. On the basis of the accuracy rate obtained by knots and type of defect, two models are proposed where their results are compared and discussed. It is found that the global accuracy rate exceeds 93% for the power transformer diagnosis, demonstrating the effectiveness of the proposed technique. An independent database is employed as a complimentary validation phase of the proposed research.

INDEX TERMS DGA, fault diagnosis, power transformer oil, KNN algorithm, decision tree principle.

I. INTRODUCTION

High voltage power transformers are mainly required for the associated heavy and powerful applications in an industrial environment. These transformers use particular insulation systems that depend primarily on the voltage levels. Indeed, the higher the voltage, the greater the impact on the transformer's lifetime and reliability [1]. Otherwise, the transformer insulation systems may deteriorate when exposed to numerous defects arising from overheating, paper carbonization, arcing, and discharges of low or high energy [2]. Therefore, early-stage detection of faults should be conducted to ensure an efficient service of these transformer [3], [4]. For this purpose, several methods were proposed in the literature. Among them, Dissolved Gas Analysis (DGA) represents one of the fastest, economical, and widely used techniques referring to those related to the dielectric insulation systems [5], [6]. Hydrogen (H_2), Methane (CH_4), Acetylene (C_2H_2), Ethylene (C_2H_4), and Ethane (C_2H_6) might be generated within the oil during a faulty mode [7]. Hence, the power transformer's abnormal state can be identified by the DGA method according to the dissolved gases composition and content. The concentrations of these gases are associated with six basic electrical and thermal faults, which might occur separately or in a combination [1], [8].

Based on DGA, different approaches have been developed to diagnose the multiple transformer faults and quantitatively indicate each fault's likelihood. These approaches are mainly based on (i) graphical (e.g., [2], [9]), (ii) artificial intelligence techniques (e.g., [10], [11]), and (iii) other improved coupled techniques (e.g., [12]). Overall, the analysis is generally based on the concentration of the five principal hydrocarbon gases where the said concentration is used directly in ppm or percentages to the total sum. Likewise, various methods are based on a combination of ratios of some specific gases. For instance, three-gas ratios (C_2H_2/C_2H_4, CH_4/H_2, and C_2H_4/C_2H_6) are used in Roger's method [6] and IEC 60599 [8], three relative percentages in Duval's triangle (%CH_4, %C_2H_2 and %C_2H_4) [2], and four-gas ratios (CH_4/H_2, C_2H_2/CH_4, C_2H_4/C_2H_6, and C_2H_2/CH_4) in Dornenburg method [1]. Such techniques are mostly required for the associated heavy and powerful applications. These techniques and other ones (e.g., [13], [14]) are presented in the literature to identify the different kinds of

faults occurring in operating transformers. However, their diagnostic accuracy requires further improvement. In this light, Duval's pentagon has been developed as a complementary tool for interpreting the DGA in power transformers [9].

For this technique, five relative percentages of the (five) leading hydrocarbon gases (%H_2, %CH_4, %C_2H_2, %C_2H_4, and %C_2H_6) are used. In the same context, probabilistic classifiers based on Parzen Windows, Bayesian and Mexican hat functions (e.g., [15], [16]) have been employed for transformers fault classification using actual DGA data. Moreover, various artificial intelligence techniques have been also applied for the transformer fault diagnosis, such as a fuzzy logic technique [17]. On the other hand, bootstrap and genetic programming have been developed to extract classification features for each fault class. These extracted features have been employed as the inputs to an artificial neural network (ANN), a support vector machine (SVM), or a K-Nearest Neighbor (KNN) classifier to perform multicategory fault classification [18]. Also, combined Duval pentagon with SVM and KNN algorithms have been proposed to improve the fault diagnostic accuracy [7]. It is worth noting that the KNN algorithm was first suggested by Cover and Hart in 1967 [19]. This algorithm has encountered several recent improvements (e.g., [20-23]).

Overall, the originality of this work consists in introducing several input vectors into the KNN classification algorithm based on a decision tree principle (DT) in order to select the best one that achieves high accuracy for the transformer faults diagnostic. Various types and combinations of input vectors have been employed, namely, the concentration of gas in ppm, relative concentration of gas in percentage, IEC ratios, Rogers four-ratios, Dornenburg ratios, Duval triangle coordinates, Duval pentagon coordinates, a combination of Rogers and Dornenburg ratios, and the combination of Duval triangle-pentagon coordinates. The accuracy rate has been analyzed to select the most appropriate input vector for the proposed method.

The current paper is organized as follows: In Section 2, we formulate the faults classification problem in power transformers and describe the database set used in this investigation. A general description of the KNN technique's theory is introduced in Section 3, including the basic theory of the KNN technique. In Section 4, the interpretative methods based on DGA are reported. The selection of an appropriate input feature for each classification layer and the proposed combined KNN classifier's performance with the decision tree rule are accomplished in Section 5 to demonstrate the proposed method effectiveness. Finally, conclusions are summarized, and potential future works are discussed in Section 6.

II. PROBLEM FORMULATION

Highly reliable transformers are mainly made of iron core and windings. Both components are placed in a tank filled with insulating oil. Figure 1 shows a cross-section of a typical oil-immersed power transformer.

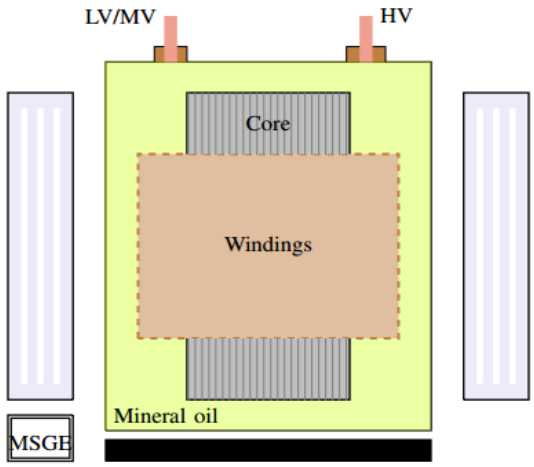

Figure 1. Representation of oil-immersed power transformer cross-section

Dissolved gases, that are liberated under particular electrical or thermal constraints, represent a powerful feature indicating the affection of oil properties. In general, the most important gases are H_2, CH_4, C_2H_2, C_2H_4, and C_2H_6. The concentration of these gases is strongly related to the type of transformer fault, and the rate of gas generation can be used to identify the fault type [24]. For instance, acetylene is associated with arcs where temperatures reach several thousand degrees. Ethylene is related to hotspots between 150 °C and 1000 °C, and hydrogen with cold gas plasma from corona discharges. Although mixtures of all gases, including other saturated hydrocarbons, are generally obtained in most cases of faults where their relative proportions have been correlated with the different fault types [25].

Oil samples in small volumes are periodically taken from the drain valve at the bottom of the transformer tank for the DGA test using chromatography as described in [25], [26]. The most common defects are partial discharge (PD), low energy discharges (D1), high energy discharges (D2), thermal faults < 300 °C (T1), thermal faults of 300 °C to 700 °C (T2), and thermal faults > 700 °C (T3) [7].

A database set of 501 samples is used to train, test, and validate the proposed classifier in the present investigation. This database is collected from the literature [4, 28-31]. The fault type samples distribution (number of samples associated with each defect) is explained in Table I.

TABLE I.
DATABASE DISTRIBUTION

Defect		Interpretation	Samples
Electric	PD	Partial discharge	53
	D1	Low energy discharges	81
	D2	High energy discharges	130
Thermal	T1	Thermal faults of < 300 °C	98
	T2	Thermal faults of 300 °C to 700 °C	51
	T3	Thermal faults of > 700 °C	88
Total			501

III. K NEAREST NEIGHBORS CLASSIFIER (KNN)

KNN algorithms are ranked among the simplest intelligent algorithms that do not require any learning phase. It is based on calculating the distances between the sampling points to the nearest neighbors of the set of assigned points [19]. The decision is based on the majority vote of the KNN. Many types of distances can be used to decide the nearest neighbors, such as Gaussian, triangular, cosine … etc. [10].

Figure 2 illustrates the principle as well as the influence of the choice of neighbors number. The three closest star neighbors selection allows the star classification as a square (objects inside the small circle in a continuous line). However, the star is classified as a triangle if we consider five closest neighbors (items inside the large ring in discontinuous line). Indeed, the choice of the neighbor's number k is a leading factor during the classification process.

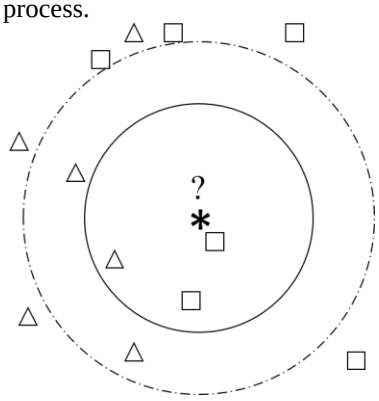

Figure 2. Influence of the number of neighbors on the classification results (reproduced from [7])

A. 1-NN ALGORITHM (KNN WITH K=1)

We consider L the data set of awarded points given by:

$$L : \{(y_i, x_i),\ i = 1, ..., n\}$$

where yi ∈ {1,..., c} denotes the class of sample i and the vector xi = (x_{i1} ,..., x_{ip}) represents the variables that characterize the sample i.

The distance function *d*(.,.) determines the nearest neighbor. P variables characterize the Euclidean distance between a sample xi and an attributed point x_i and is defined by :

$$d((x_{i1}, x_{i2}, ..., x_{ip}),(x_{j1}, x_{j2}, ..., x_{jp})) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2} \quad (1)$$

The observation of the sample (*y*, *x*) by the nearest neighbor (*y*(1), *x*(1)) in the learning sample is determined by :

$$d(x_i, x_j) = min_j(d(x_i, x_j)) \quad (2)$$

We designate by ŷ = y(i) the estimated class of the nearest neighbor. This class is selected for the prediction of y.

Considering the Minkowski formula of equation (3), the Euclidian distance is obtained by replacing p by 2 and is given by equation (4)

$$d(x_i, x_j) = \left(\sum_{s=1}^{p} \left| x_{is} - x_{js} \right|^p \right)^{\frac{1}{p}} \quad (3)$$

$$d(x_i, x_j) = \left(\sum_{s=1}^{p} (x_{is} - x_{js})^2 \right)^{\frac{1}{2}} \quad (4)$$

B. KNN ALGORITHM

In applications, several closer neighbors are usually employed. The decision favors the classes majority represented by the k neighbors obtained from equation 5, k_r being the number of observations from the group of closest neighbors belonging to class r and "*c*" the number of classes.

$$\sum_{r=1}^{c} k_r = k \quad (5)$$

The observation is predicted in class l, only if l = max (k_r) with k ∈ N.

It is recommended to choose the odd k, o avoid equal votes in the binary classification. However, in the case of multi-classes, the best choice of k depends on the nature of the data. The noise effect on classification (risk of overlearning) is reduced when k takes large values. However, this choice makes the boundaries between classes less distinct. The best selection of k is the one that minimizes the classification error [27].

C. KERNEL FUNCTION

The classification is influenced by parameter k and the type of the kernel function K(x). The K function must check the following properties:

- K(d) ≥ 0 for all d ∈ ℜ;
- K(d) reaches its maximum for d=0;
- K(d) decreasing for d→ ±∞.

Several kernel functions [27] are rectangular (uniform law), triangular, Epanechnikov, cosine, Gaussian, and reverse.

D. KNN ALGORITHM ADVANTAGES AND DISADVANTAGES

This technique is easy to implement and apply to any data, including complex ones such as geographic information, text, images, and sound. Also, it is robust to noise. The introduction of new data does not require the reconstruction of a model. The class is assigned to an object with ease and clarity once the closest neighbors are displayed.

As mentioned, the method performance depends on the distance type, and the number of neighbors, and how the neighbors' responses are combined. The results could be of poor quality if the number of relevant attributes is low relative to the total number of characteristics. The distances on the irrelevant attributes will drown out the proximity on the appropriate attributes. The calculations made in the classification phase can be very time-consuming.

IV. DGA–BASED INPUT VECTOR

Many interpretative methods based on DGA were reported in the literature to detect the incipient fault nature within an oil-immersed power transformer [7, 10]. These techniques mainly include the following input vectors:

- **Vector 1:** Since the database contains the concentrations of the five gases in parts per million or ppm, each sample **X** is represented as follows:

$$X=[H_2, CH_4, C_2H_2, C_2H_4, C_2H_6] \quad (6)$$

- **Vector 2:** Since the weight percent of the gases would result in an inopportunely small number, percent concentration to the total sum is also used where each sample $\mathbf{X} = [x_1, x_2, \cdots, x_5]$ is scaled as follows:

$$X \leftarrow \frac{X}{\sum_{i=1}^{5} x_i} \text{ x } 100\%$$

$$X=[\%H_2, \%CH_4, \%C_2H_2, \%C_2H_4, \%C_2H_6] \quad (7)$$

- **Vector 3:** The IEC Ratios method is used to produce the following input vector containing three ratios of the dissolved gases given by:

$$X=[\frac{CH_4}{H_2}, \frac{C_2H_2}{C_2H_4}, \frac{C_2H_4}{C_2H_6}] \quad (8)$$

- **Vector 4:** Roger's four-ratio method has been selected in this case to transform each sample to the following one:

$$X=[\frac{CH_4}{H_2}, \frac{C_2H_2}{C_2H_4}, \frac{C_2H_4}{C_2H_6}, \frac{C_2H_6}{CH_4}] \quad (9)$$

- **Vector 5:** Dornenburg's method is also investigated in this study. In this method, the input consists of four ratios computed as a function of the dissolved gases in ppm as follows:

$$X=[\frac{CH_4}{H_2}, \frac{C_2H_2}{C_2H_4}, \frac{C_2H_4}{C_2H_6}, \frac{C_2H_2}{CH_4}] \quad (10)$$

- **Vector 6:** Duval triangle is a graphical method that use only the concentration of three gases (CH_4, C_2H_2, and C_2H_4) to produce the input vector as follows:

$$X=[C_x, C_y] \quad (11)$$

where the components C_x and C_y are computed by:

$$C_x = \frac{1}{3} \frac{\sum_{i=0}^{n-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i)}{\sum_{i=0}^{n-1}(x_i y_{i+1} - x_{i+1} y_i)} \quad (12)$$

And

$$C_y = \frac{1}{3} \frac{\sum_{i=0}^{n-1}(y_i + y_{i+1})(y_i x_{i+1} - y_{i+1} x_i)}{\sum_{i=0}^{n-1}(y_i x_{i+1} - y_{i+1} x_i)} \quad (13)$$

The parameters x_i are defined as follows:

$$x_0 = \%CH_4 \cos(\frac{\pi}{2})$$

$$x_1 = \%C_2H_4 \cos(\frac{\pi}{2} + \alpha) \quad (14)$$

$$x_2 = \%C_2H_2 \cos(\frac{\pi}{2} + 2\alpha)$$

The parameters y_i can be found by replacing the cosine with the sine in the previous expressions with $\alpha = 2\pi/3$.

- **Vector 7:** The input vector, in this case, has two components given as follows:

$$X=[C_x, C_y] \quad (15)$$

X is computed according to the Duval pentagon that uses the concentration of five gases in percentages. The components C_x and C_y are calculated by:

$$C_x = \frac{1}{6} \frac{\sum_{i=0}^{n-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i)}{\sum_{i=0}^{n-1}(x_i y_{i+1} - x_{i+1} y_i)} \quad (16)$$

And

$$C_y = \frac{1}{6} \frac{\sum_{i=0}^{n-1}(y_i + y_{i+1})(y_i x_{i+1} - y_{i+1} x_i)}{\sum_{i=0}^{n-1}(y_i x_{i+1} - y_{i+1} x_i)} \quad (17)$$

where the parameters *xi* are defined as follows:

$$x_0 = \%H_2 \cos(\frac{\pi}{2})$$

$$x_1 = \%C_2H_6 \cos(\frac{\pi}{2} + \alpha)$$

$$x_2 = \%CH_4 \cos(\frac{\pi}{2} + 2\alpha) \quad (18)$$

$$x_3 = \%C_2H_4 \cos(\frac{\pi}{2} + 3\alpha)$$

$$x_4 = C_2H_4 \cos(\frac{\pi}{2} + 4\alpha)$$

The parameters y_i can be found by replacing the cosine with the sine in the previous expressions with $\alpha = 2\pi/5$.
Other possible combinations of the above technique have also been proposed to give strong credibility to the obtained results. Two combinations are given below.

- **Vector 8:** The first combination consists of an input vector of five ratios given by:

$$X=[\frac{CH_4}{H_2}, \frac{C_2H_2}{C_2H_4}, \frac{C_2H_4}{C_2H_6}, \frac{C_2H_2}{CH_4}, \frac{C_2H_6}{CH_4}] \quad (19)$$

Equation (14) refers to the mixture of Roger's and Dornenburg's methods.

- **Vector 9:** The input vector, in this case, has four components given as follows:

$$X=[C_{x1}, C_{y1}, C_{x2}, C_{y2}] \quad (20)$$

According to Duval's triangle and pentagon, this vector has computed that use the concentration of five gases in percentages. C_{x1} and C_{y1} are calculated according to the triangle method, while C_{x2} and C_{y2} employ the pentagon technique.

V. SIMULATIONS AND RESULTS

The database of 501 samples, already used to evaluate the KNN classifier's accuracy rate, has been also employed in this section. Randomly selected, 321 samples have been utilized for the training phase,160 samples for the testing phase, and 20 samples to examine the validity of the proposed classifier. Each previous vector has been used independently as an input of a KNN classifier. Several types of distance have been used, namely "Euclidean", "Cityblock", "Cosine", and "Correlation". For the training phase, the number of neighbors k has been varied from 1 to 321 where the value corresponding to the better accuracy rate is maintained. This procedure has been repeated for the nine types of the previously defined input vectors. Figure 3 shows an example of the classification results obtained when using the first input vector (gases in ppm).

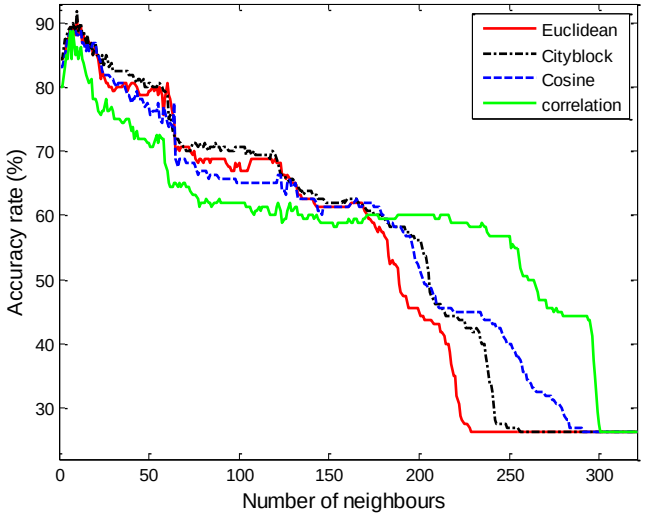

Figure 3. Impact of input vector "type 1" and number of neighbors on the classification performance (accuracy rate)

From the obtained results in Figure 3, an accuracy rate of 88.75% has been developed for both distances "Correlation" and "Cosine" with the same neighbors number k= 6. Whereas, an accuracy rate of 77.50% (respectively 76.25%) is obtained for "Cityblock" (respectively Euclidean) achieved for a distance k=8 (respectively k=4).

For several neighbors and distance types, all input vectors have been separately tested. Based on the obtained results, the best classification accuracy rate is selected for each input vector. Figure 4 regroups the best nine classification results.

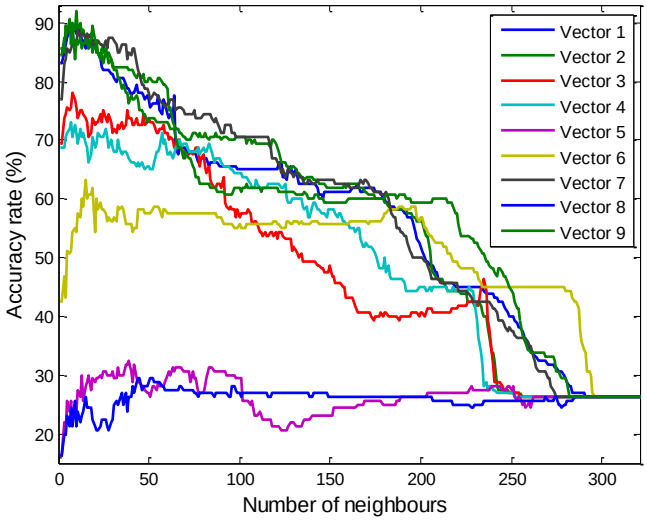

Figure 4. Performance of KNN classifier as function of the number of neighbors for different input vectors

From this figure, it is clear that the accuracy rate is affected by both the type and the number of the neighbors. Higher accuracies are obtained for a relatively low number of neighbors, tending toward a plateau of about 28%. In order to quantify the results of Figure 4, the best accuracy rate over the number of the neighbors is subtracted for each input vector. The obtained results are given by Table II.

TABLE II.
KNN CLASSIFIER FOR ALL INPUT VECTORS

Input vector	Vector 1	Vector 2	Vector 3	Vector 4	Vector 5	Vector 6	Vector 7	Vector 8	Vector 9
Accuracy rate (%)	88.75	**91.88**	78.13	73.13	32.50	63.13	88.75	29.88	**90.63**
Best distance type	Cosine	**Cityblock**	Cityblock	Cityblock	Euclidean	Cosine	Euclidean	Cosine	**Cityblock**
Best K	6	**10**	8	7	39	15	11	44	**6**

TABLE III.
KNN ANALYSIS OF ACCURACY RATE

Electrical and thermal faults separation node									
Input vector	Vector 1	Vector 2	Vector 3	Vector 4	Vector 5	Vector 6	Vector 7	Vector 8	Vector 9
Accuracy rate (%)	**98.75**	**98.75**	**98.75**	95.625	55.625	91.25	**98.75**	41.875	**98.75**
The accuracy rate of the first electrical fault node (PD vs D1&D2)									
Input vector	Vector 1	Vector 2	Vector 3	Vector 4	Vector 5	Vector 6	Vector 7	Vector 8	Vector 9
Accuracy rate (%)	88.10	**91.67**	71.43	66.67	55.95	70.24	88.10	51.19	90.48
The accuracy rate of the second electrical fault node (D1 vs D2)									
Input vector	Vector 1	Vector 2	Vector 3	Vector 4	Vector 5	Vector 6	Vector 7	Vector 8	Vector 9
Accuracy rate (%)	86.76	**91.18**	70.59	66.18	48.51	63.24	86.76	58.82	88.23
The accuracy rate of the first thermal fault node (T1&T2 v T3)									
Input vector	Vector 1	Vector 2	Vector 3	Vector 4	Vector 5	Vector 6	Vector 7	Vector 8	Vector 9
Accuracy rate (%)	89.47	**93.24**	90.54	82.43	6.75	70.27	91.89	5.40	**93.24**
The accuracy rate of the second thermal fault node (T1 vs T3)									
Input vector	Vector 1	Vector 2	Vector 3	Vector 4	Vector 5	Vector 6	Vector 7	Vector 8	Vector 9
Accuracy rate (%)	85.42	87.5	89.58	77.08	4.16	68.75	87.5	0	**91.67**
Accuracy rate by fault type									
Input vector	Vector 1	Vector 2	Vector 3	Vector 4	Vector 5	Vector 6	Vector 7	Vector 8	Vector 9
PD	93.75	93.75	75	68.75	87.5	37.5	93.75	18.75	**100**
D1	**80.77**	**80.77**	46.15	42.31	11.53	26.92	96.23	15.38	80.77
D2	90.47	**97.62**	85.71	80.95	71.43	85.71	97.62	85.71	93.86
T1	93.75	**96.87**	96.87	78.12	3.12	84.37	100	0	**96.87**
T2	68.75	68.75	75	75	6.25	37.5	75	01	**81.25**
T3	**96.43**	**96.43**	85.71	85.71	10.71	67.86	92.86	14.29	89.29

As shown in Table II, it is found that the highest accuracy rate, of 91.88%, is obtained when employing vector 2 (gas in percentage) as an input of the KNN with Cityblock distance and k=10. Furthermore, the combined triangle and pentagon coordinates (vector 9, using Cityblock distance and k= 6) came in second place with an accuracy rate of 90.63%. Finally, the combined Roger's and Dornenburg' gave the poorest results (29.88%) compared to the others.

As stated in the introduction, the novelty of this investigation is to examine several types of input vector, in the proposed classifier, and to compare the obtained diagnostic accuracies. Select the best input vector that effectively separates the electrical and thermal faults, which represents the first node of the classification process. For the electrical fault, there are two stages (nodes); (i) the first electrical node separates the fault PD from D1&D2, and (ii) the second electrical node distinguish between the faults D1 and D2. It is important to note that the same scenario is repeated for the thermal fault. The first thermal node aims to separate T3 and T1&T2, and the last thermal node intends to isolate T1 and T2. The decision tree's strategy is represented in Figure 5. Note that the best number of neighbors k and the selected distance types were employed in the following study.

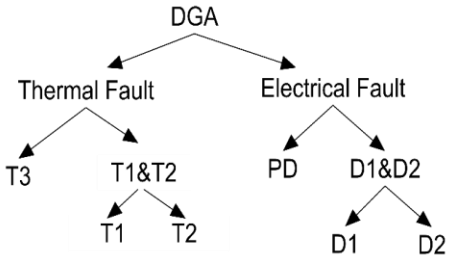

Figure 5. General representation of the decision tree strategy

Elaborating the whole of the previous input vectors, an analysis has been made to compute the accuracy rate at each node. Table III illustrates a detailed analysis of the accuracy rate leading to selecting the more appropriate input vector and neighbor number for each node.

From these results, one can clearly see that the input vectors 1, 2, 3, 7, and 9 separate the electrical and thermal faults with the same accuracy rate of 98.75%. Indeed, both nodes of electrical faults are suitable with the input vector in ppm. Regarding the thermal faults node, the first node (T3 vs. T1&T2) is more compatible with both input vectors, in percentage, and the combined Duval triangle and pentagon. The accuracy rate in this stage reaches 93.24 %. The second thermal node is consistent with the combined Duval triangle and pentagon input vector, with an accuracy rate of 91.67%.

Based on the results of Table III as well as the accuracy rate by nodes and type of fault, it can be concluded that the use of the input vectors 2 and 9 in the same algorithm can improve the overall accuracy of the diagnostic. This reasoning can be recapitulated in the flowchart given by Figure 6 (denoted by model 1).

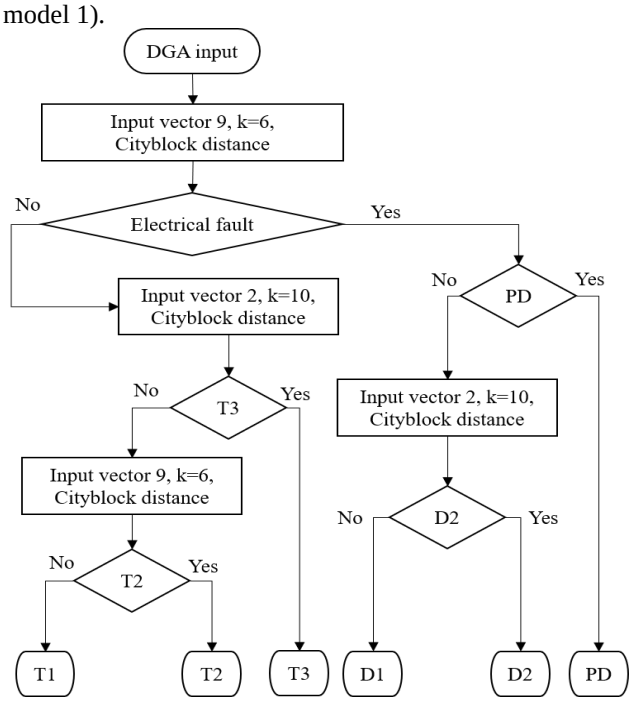

Figure 6. Proposed flowchart of improved diagnostic accuracy (Model 1)

Table IV shows the diagnostic accuracy rate of the KNN algorithm when using the input vector of DGA in ppm and the combined Duval triangle and pentagon input vectors and the proposed enhancement of diagnostic accuracy of the KNN algorithm by the strategy of a decision tree.

TABLE IV.
COMPARISON OF ACCURACY RATE BEFORE AND AFTER IMPROVEMENT WITH MODELS 1

	KNN with input vector 9	KNN with input vector 2	KNN based on decision tree principle (model 1)
Accuracy (%)	90.63	91.88	92.5

One can see that the diagnostic accuracy rate achieved a value of 92.5% with the proposed method. It was 91.88 % with the input vector of DGA in ppm and 90.63 % with the combined Duval triangle and pentagon input vector when used alone (Table III). The first thermal fault node detected many cases by T3, but the actual fault was T2. Therefore, another model has been proposed, as in Figure 7, to overcome this situation.

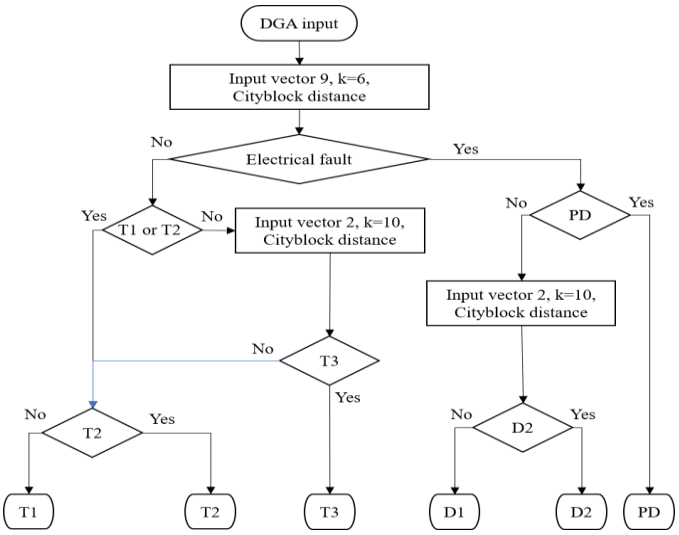

Figure 7. Proposed flowchart of improved diagnostic accuracy (Model 2)

The second proposed model shows a high accuracy rate of 93.75 % compared with the previous application given by Table IV. This imply that the second model is more accurate. For further verification, all proposed models are tested by an independent new dataset of 20 samples. Table V shows the results of both proposed models along with those related to the KNN classifier with input vectors 2 and 9.

TABLE V.
THE ACCURACY RATE OF THE VALIDATION PHASE

	KNN with input vector 9	KNN with input vector 2	KNN based on decision tree principle (model 1)	KNN based on decision tree principle (model 2)
Accuracy (%)	14/20	16/20	15/20	**18/20**

For the validation phase, the results of Table V confirm that the second proposed model (model 2) gave the best accuracy rate compared to the other ones. An accuracy rate of 90% was obtained against 80% when employing the first proposed (model 1). Both models are more accurate than the traditional methods of power transformer diagnosis as shown in Table V. Table VI illustrates the comparison between the KNN-DT (with model 2) diagnosis results and the Duval triangle, Rogers ratios, and IEC 60 599 methods, the common DGA methods in the literature. The results in Table VI explained the high performance of the KNN- DT for correctly diagnose of transformer faults where the KNN-DT accuracy succeeded in detecting 18 samples of 20 with 90% accuracy. On the other hand, the other three methods developed poor diagnostic accuracies, which are 30 % (6/20), 25 % (4/20), and 30 % (6/20) for Duval triangle, Rogers' ratio, and IEC code, respectively. Therefore, the KNN-DT has high reliability to diagnose transformer faults.

TABLE VI.
COMPARISON BETWEEN THE PROPOSED ALGORITHM AND OTHER TRADITIONAL METHODS

Sample	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	ACT	Duval	Rogers' 4	IEC-60599	KNN-DT	Ref.
1	6454	2313	121	6432	2159	D2	D2	UD*	D2	D2	[28]
2	305	100	33	541	161	D1	D2	UD	D2	D2	[28]
3	1230	163	27	692	233	D2	D2	UD	D2	D1	[28]
4	33046	619	58	0	2	PD	PD	PD	UD	PD	[28]
5	796	999	234	31	1599	T3	D1	UD	UD	T3	[28]
6	34	21	4	56	49	D2	D2	D2	D2	D2	[28]
7	960	4000	1290	6	1560	T2	D1	UD	UD	T2	[28]
8	6	2990	29990	67	26076	T1	D1	UD	UD	T1	[28]
9	2500	10500	4790	6	13500	T2	D1	UD	UD	T2	[28]
10	300	700	280	36	1700	T3	D1	UD	UD	T3	[28]
11	37800	1740	249	8	8	PD	PD	PD	PD	PD	[28]
12	1450	940	211	61	322	T1	D1	UD	UD	T1	[28]
13	120	140	30	0	120	T1	D1	UD	UD	T1	[29]
14	3700	6400	2400	10	7690	T2	D1	UD	UD	T2	[29]
15	125	680	290	20	900	T3	D1	UD	UD	T3	[29]
16	120	10	30	25	5	D1	T3	UD	PD	D1	[29]
17	140	95	10	80	60	D2	D2	D2	D2	D2	[29]
18	240	17	0	5	40	PD	D1	UD	UD	PD	[29]
19	650	53	20	0	34	PD	D1	PD	UD	PD	[29]
20	1076	95	71	231	4	PD	T3	UD	UD	PD	[29]
							6/20	4/20	6/20	18/20	
						Accuracy (%)	30	25	30	90	

UD* (Undetermined): means that the DGA method fails to interpret the transformer fault type.

VI. CONCLUSIONS

Oil-immersed power transformer fault diagnosis was investigated using different DGA methods, in which a database set of 501 samples was exploited. A decision tree algorithm was improved the KNN classifier to enhance the accuracy of the transformer faults diagnostic. The neighbor's number and distance type of the KNN algorithm were optimized to improve the classifier's accuracy rate. Several input vectors were assessed for each fault to select the most appropriate vector associated with this type of fault for the combination of the KNN classifier with the decision tree principle. The obtained results were discussed, and two models were proposed in order to improve the global accuracy rate. Both proposed models confirmed their accurateness where the global accuracy rate exceeded 93% for the power transformer diagnosis. A complementary validation phase of the proposed research was also considered using an independent database.

VII. AKNOWLDGEMENT

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.

VIII. REFERENCES

[1] "IEEE Guide for the Interpretation of Gases Generated in Mineral Oil-Immersed Transformers," IEEE Std C57.104-2019 (Revision of IEEE Std C57.104-2008), vol., no., pp.1-98, 1 Nov. 2019.

[2] M. Duval, "Dissolved Gas Analysis: It Can Save Your Transformer," *IEEE Elec. Insul. Mag.*, vol. 5, no. 6, pp. 22–27, Nov./Dec. 1989.

[3] S. Ghoneim, "Intelligent prediction of transformer faults and severities based on dissolved gas analysis integrated with thermodynamics theory," *IET Sci. Meas. Tech.*, vol. 12, no. 3, pp. 388–394, May 2018.

[4] I.B.M. Taha, A. Hoballah and S.S.M. Ghoneim, "Optimal ratio limits of rogers' four-ratios and IEC 60599 code methods using particle swarm optimization fuzzy-logic approach," *IEEE Trans. Dielectr. Elect. Insulation*, vol. 27, no. 1, pp. 222–230, Feb. 2020.

[5] "Recent Developments in DGA Interpretation," CIGRE Technical Brochure 296, Jun. 2006.

[6] R.R. Rogers, "IEEE and IEC codes to interpret incipient faults in transformers using gas in oil analysis," *IEEE Trans. Elec. Insul.*, vol. 13, no. 5, pp. 348–354, Oct. 1978.

[7] Y. Benmahamed, M. Teguar, and A. Boubakeur, "Application of SVM and KNN to Duval Pentagon 1 Transformer Oil Diagnosis," *IEEE Trans. Elec. Insul.*, vol. 24, no. 6, pp. 3443–3451, Dec. 2017.

[8] Mineral Oil-impregnated Electrical Equipment in Service–Guide to the Interpretation of Dissolved and Free Gases Analysis, IEC Publication 60599, 2007.

[9] M. Duval and L. Lamarre, "The Duval Pentagon - A New Complementary Tool for the Interpretation of Dissolved Gas Analysis in Transformers," *IEEE Elec. Insul. Maga.*, vol. 30, no.6, pp.9–12, Nov./Dec. 2014.

[10] Y. Benmahamed, Y. Kemari, M. Teguar and A. Boubakeur, "Diagnosis of Power Transformer Oil Using KNN and Nave Bayes Classifiers," *in Proc. 2018 IEEE 2nd Int. Conf. on Dielectrics (ICD)*, Budapest, Hungary, 1–5 Jul. 2018.
[11] 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*, vol. 9, pp. 30817- 30832, Feb. 2021.
[12] Y. Benmahamed, M. Teguar and A. Boubakeur, "Diagnosis of Power Transformer Oil Using PSO-SVM and KNN Classifiers," *in Proc. 2018 Int. Conf. on Electrical Sciences and Technologies in Maghreb (CISTEM)*, Algiers, Algeria, 28-31 Oct. 2018.
[13] W. Chen, C. Pan, Y. Yun and Y. Liu, "Wavelet Network in Power Transformers Diagnosis Using Dissolved Gas Analysis", *IEEE Trans. Power Del.*, vol. 24, no. 1, pp. 187–194, Jan. 2009.
[14] M. Hasmat, "Application of Gene Expression Programming (GEP) in Power Transformers Fault Diagnosis Using DGA", *IEEE Trans. Indus. Apps.*, vol. 52, no. 6, pp. 4556–4565, Nov./Dec. 2016.
[15] W.H. Tang, J.Y. Goulermas, Q.H. Wu, Z.J. Richardson, and J. Fitch, "A Probabilistic Classifier for Transformer Dissolved Gas Analysis With a Particle Swarm Optimizer", IEEE *Trans. Power Del.*, vol. 23, no.2, pp. 751–759, Apr. 2008.
[16] Z. Yongli, H. Limin and L. Jinling, "Bayesian networks-based approach for power systems fault diagnosis", *IEEE Trans. Power Del.*, vol. 21, no. 2, pp. 634–639, Apr. 2006.
[17] Q. Su, C. Mi, L.L. Lai and P. Austin, "A Fuzzy Dissolved Gas Analysis Method for the Diagnosis of Multiple Incipient Faults in a Transformer", *IEEE Trans. Power Del.*, vol. 15, no. 2, pp. 593–798, May 2000.
[18] A. Shintemirov, W.H. Tang and Q.H. Wu, "Power Transformer Fault Classification Based on Dissolved Gas Analysis by Implementing Bootstrap and Genetic Programming", *IEEE Trans. Power Del.*, vol. 39, no.1, pp. 69–79, Jan. 2009.
[19] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification", *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21--27, Jan. 1967.
[20] Y. Shih and C. Ting, "Evolutionary Optimization on k-Nearest Neighbors Classifier for Imbalanced Datasets," *2019 IEEE Congress on Evolutionary Computation (CEC)*, pp. 3348-3355, Wellington, New Zealand, 2019.
[21] S. Liu, P. Zhu and S. Qin, "An Improved Weighted KNN Algorithm for Imbalanced Data Classification," *2018 IEEE 4th Int. Conf. on Computer and Communications (ICCC)*, pp. 1814-1819, Chengdu, China, 2018.
[22] Y. Cui, H. Ma and T. Saha, "Improvement of power transformer insulation diagnosis using oil characteristics data preprocessed by SMOTEBoost technique," *IEEE Trans Dielect. Elect. Insul.*, vol. 21, no. 5, pp. 2363-2373, Oct. 2014.
[23] K. Li, P. Xie, J. Zhai and W. Liu, "An improved adaboost algorithm for imbalanced data based on weighted KNN," *2017 IEEE 2nd Int. Conf. on Big Data Analysis (ICBDA)*, pp. 30-34, Beijing, China, 2017.
[24] P. Mirowski and Y. LeCun, "Statistical Machine Learning and Dissolved Gas Analysis: A Review", *IEEE Trans. Power Del.*, vol. 27, no. 4, pp.1791–1799, Oct. 2012.
[25] N. Abu Bakar, A. Abu-Siada, and S. Islam, "A Review of Dissolved Gas Analysis Measurement and Interpretation Techniques", *IEEE Electr. Insul. Mag.*, vol. 30, no. 3, pp. 39–49, May-June 2014.
[26] K. Nagapriya, S. Shashank, R. Prashanth et *al.*, "Laser Calorimetry Spectroscopy for ppm-level Dissolved Gas Detection and Analysis", *Nature Scientific Reports*, vol. 7, pp. 42917, Feb. 2017.
[27] Eve Mathieu-Dupas, "Algorithme des K plus proches voisins pondérés et Application en diagnostic", nria-00494814, pp. 1-24, Jun. 2010.
[28] J. Dukarm and F. Jakob, "Thermodynamic estimation of transformer fault severity," *2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*, Dallas, TX, USA, 2016, pp. 1-1.
[29] J. Soni and D. Suthar, "An Experimental Analysis to Check Accuracy of DGA Using Duval Pentagonal Method in Power Transformer", *Selected Papers in Engineering (in Prod ICRISET2017 the Int. Conf. on Research and Innovations in Science, Engineering and Technology)*, vol 1, pp. 394--401, Aug. 2017.

OMAR KHERIF (S'17)(M'19) received the M.Eng. and the Ph.D. degrees in electrical engineering from the Ecole Nationale Polytechnique (ENP) in 2015 and 2019, respectively. He was a mathematical-analysis assistant teacher with ENP from 2016 to 2018. He was also a volunteer researcher with the Laboratoire de Recherche en Electrotechnique of the ENP, where he co-supervised several Master theses. He is currently a KTP Associate with Cardiff University. His main interests are in high voltage techniques, electromagnetic transients, earthing and power system protection.

YOUCEF BENMAHAMED (S'17) (M'19) received the degree of Engineer and Master's degree in power electronics engineering in 2014 and Ph.D in High Voltage techniques in 2019 from Ecole Nationale Polytechnique (ENP) of Algiers. His research interests are in diagnosis, artificial intelligence and optimization techniques.

MADJID TEGUAR has obtained a first degree in Electrical Engineering in 1990, a Master's degree in 1993 and a Ph.D in High Voltage Engineering in 2003 from Ecole Nationale Polytechnique (ENP) of Algiers. He is now a Professor in electrical engineering with ENP. His main research interests are in insulation systems, insulation coordination, earthing of electrical energy systems and polymeric cables insulation.

AHMED BOUBAKEUR (M'02) (SM'03) was born in Biskra, Algeria in 1952. He received in 1975 the degree of Engineer in electrical engineering from Ecole Nationale Polytechnique (ENP) of Algiers, and in 1979 he obtained the Doctorate in Technical Sciences from the Institute of High Voltage Engineering of the Technical University of Warsaw of in Poland. He is currently a professor at ENP of Algiers where he has been giving lectures and supervising research in the field of High Voltage Engineering since 1982. His principal research areas are discharge phenomena, insulators pollution, lightning, polymeric cables insulation, transformer oil ageing, neural network and fuzzy logic application in HV insulation diagnosis, and electric field calculation and measurement. He is an IEEE senior member, member of IEEE/DEIS and a member of the Algerian HV Power Systems Association ARELEC (National Algerian Commity of CIGRE and ENP Elders Association ADEP). He has been member of the Editorial Board and Associate Editor of IET/SMT.

SHERIF S. M. GHONEIM (M'06) (SM'19) received his B.Sc. and M.Sc. degrees from the Faculty of Engineering at Shoubra, Zagazig University, Egypt, in 1994 and 2000, respectively. Since 1996, he has been teaching at the Faculty of Industrial Education, Suez Canal University, Egypt. From the end of 2005 to the end of 2007, he was a guest researcher at the Institute of Energy Transport and Storage (ETS) of the University of Duisburg–Essen in Germany. In 2008, he earned his Ph.D. degree in electrical power and machines from the Faculty of Engineering, Cairo University (2008). He joined Taif University as an associate professor in the Electrical Engineering Department, Faculty of Engineering. His research areas include grounding systems, dissolved gas analysis, breakdown in SF6 gas, and AI technique applications.