A Comparative Study of Imputation Methods in Predicting Missing Attribute Values in DGA Datasets

Zahriah Sahri
Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka,
Melaka, Malaysia.
szahriah2511@yahoo.com.my

Rubiyah Yusof
Malaysia Japan Internatioanl Institute of Technology,
Universiti Teknologi Malaysia,
Kuala Lumpur, Malaysia.
rubiyah@ic.utm.my

***Abstract*—Dissolved Gas Analysis (DGA) is one of the most deployable methods for detecting and predicting incipient faults in power transformers. For predicting faults, DGA uses tools such as Doernenburg, Rogers and IEC methods. The presence of missing values in a DGA dataset may affect the diagnostic performances of these three methods. This study applies the mean, regression, expectation-maximization, multiple imputation, and *k*-nearest neighbor methods to replace the missing values with estimated values in a DGA dataset. Using the number of unresolved diagnoses, the number of wrong diagnoses, and the number of correct diagnoses as the criteria to evaluate the effects of the imputation methods on the DGA diagnostic methods, this study shows that *k*-nearest neighbor increases the performances of Doernenburg, Rogers and IEC methods the most on two datasets with actual missing values. Experimental results show that imputing missing values in DGA datasets has increased diagnostic performance of the three ratios methods of DGA**

Keywords-dissolved gas analysis; missing values; imputation methods; gas ratios method; fault diagnosis

I. INTRODUCTION

Power transformers are a must-have item for any power utility company to increase or decrease electrical power for transmission and distribution throughout interconnected power systems. Because of prolonged usage and as transformers age, their internal conditions degrade when faults such as short-circuit, arcing, partial discharges or overheating occur during operations. These faults will release several gases commonly known as the fault gases: hydrogen (H_2), acetylene (C_2H_2), ethylene (C_2H_4), methane (CH_4), ethane (C_2H_6), carbon monoxide (CO) and carbon dioxide (CO_2) that stay dissolved at above threshold values in the insulating oil of a transformer. A faulty transformer must be removed from operation, sent for repair and/or replaced. These processes are costly and time-consuming but are necessary because if left unattended for long, a faulty transformer may trigger worse impacts such as explosions, loss of human lives, or environmental disasters. Recognizing the need for checking the serviceability of transformers, most utility companies perform preventive maintenance either periodically or conditionally to detect these faults.

Among the many existing techniques to detect these faults, Dissolved Gas Analysis (DGA) is recognized as the most effective method and is practiced universally today [1]. It involves a few sequence processes as follows: a) sampling of oil from the transformer, b) extracting the fault gases dissolved in the oil, c) calculating and analyzing the concentration of these gases, their gassing rates and the ratios of certain gases d) finally, the identification of the fault types. Currently, traditional diagnostic tools such as Key Gas [1], IEC ratios [2], Rogers ratios [1], Doernenburg ratios [1] and Duval Triangle [3] are widely used in the fourth process of DGA method. These ratio methods identify fault types using the ratios of certain fault gases and each ratio is assigned to one or more numerical thresholds. These thresholds are coded and mapped to specific faults. However, in some cases, gas concentrations may be *incomplete* which lead to combination of ratios that do not match any predefined threshold. As a result, fault that occurs inside a transformer may be classified unknown or inconclusive - a well-known shortcoming of the DGA ratios methods as documented in [4].

One of the reasons for incomplete gas concentrations is *missing values* for some of the fault gases. Missing values in DGA occur for various reasons, such as acetylene evaporates quickly, or the existence of contamination on the surface of the platinum alloy of a gas meter, and some transformer faults generate only a few fault gases [2]. Such samples containing missing values were usually manually deleted and excluded from subsequent analysis [5-6]. Majority reported only complete-case samples of DGA data [7-9]. Only researchers in [10] estimated the missing values in a DGA dataset using Support Vector Machine regression, which increased the accuracy of their Naive-Bayes classification algorithm. However, there are very few published literature concerning missing value estimation for DGA data especially with the objective of improving the diagnostic performance of the gas ratio methods. On the other hand, certain fields such as microarray analysis [11-12], medical [13-14], and social sciences [15-16] have paid more attention to this issue. Consequently, statistical analyses or machine learning algorithms that were subsequently applied after filling in the missing values have shown better results.

The aforementioned scenario motivates us to fill-in the missing values in DGA datasets with estimated values ("imputing") to minimize the inconclusive diagnoses of the gas ratio methods. Fortunately, there are many imputation methods to choose from [11,17-19], ranging from simple to complex solutions, and statistical to machine learning methods. In this paper, we propose imputing missing values in a DGA dataset using a few established methods Mean/mode (MEAN), linear regression (REG), expectation-maximization (EM), multiple imputation (MI) and *k* nearest neighbour algorithm (*k*NN) are the selected imputation methods, and Rogers ratios, Doernenburg ratios, IEC ratios are the DGA ratios methods to diagnose the transformer faults. A "before and after" experiment is conducted to compare the diagnostic performance of each ratio method applied on the original datasets with missing values and that applied on the complete datasets imputed by each imputation method.

In Section II, we provide related work on missing values and the application of imputation methods. Section III we briefly described the DGA method and its ratio-based diagnostic methods. The compared imputation methods are described in Section IV. Section V presents and analysed experimental results. Section VI concludes our findings.

II. Related Work

According to [17], a missing value indicates a lack of response. “Don’t know”, “Refused”, “Unintelligible”, and “Nil” are some of the possible codes for missing values in a dataset. However, why worry about missing values? Two major negative effects are reported in [20]. First, missing values reduce statistical power. Second, missing values could result in biased statistical estimates in several ways. Before deciding which imputation are suitable for the incomplete datasets, researchers should ask whether the pattern of missing observations is random or not ("patterns of missingness"). Little and Rubin [21] distinguished randomness into the following categories: Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR).

MCAR exists when value(s) are missing because of uncontrolled events, which are totally independent of the potential values of both the observed and the missing variables. In the case of DGA data, acetylene is a very soluble and reactive gas, and disappears fast from absorption; thus, acetylene sometimes does not appear in the collection sample. On the other hand, when data are MAR, the probability of a missing value on some variables is dependent on the value of the observed variables. However, the missing value itself is not the cause for the missing. For DGA data, a fault such as high-energy arcing releases high quantities of hydrogen and acetylene. Such a pattern is considered to be MAR, and most missing data treatment methods are assumed to be MAR Finally, NMAR occurs when the probability of missing a value is a function of the value itself. NMAR is very unlikely to appear in DGA data; because gases are released as a result of faults in power transformer.

Statistical analysis with missing data has been noted in the literature for more than 70 years. Walks [22] initiated a study on the maximum likelihood estimation for multivariate normal models with fragmentary data. Thereafter, extensive discussions on this topic continue. A useful reference for general parametric statistical inferences with missing data can be found in Little and Rubin [23]. Litwise deletion (LD), pairwise deletion (PD), REG, MEAN, EM, and MI are some statistical tools available for imputing missing values. LD and PD are deficient in several aspects. Despite their simplicity, both are inefficient [24-25]. LD, in particular, can discard an enormous amount of potentially useful data. PD may be more efficient than LD in many cases, but for some data structures it is known to be less efficient [26]. MEAN, because of its simplicity, is commonly used in the social sciences as a fast alternative to LD [14]. Also, it is often used as a base for other proposed imputation methods such as in [27-28]. EM and MI, currently represent the state of the art, have been applied to various problem domains. Researchers in [29] estimated the missing values of leaf area index , a biophysical variable, using EM and helped reduce the root mean square error of the Gaussian Bayes Network output. In the medical field [13], MI was found to preserve observed and real data better than complete-case and dropping a particular variable approaches when predicting for deep venous thrombosis in patients.

Machine learning algorithms have also garnered large followings as the choice for data imputation. The Naive Bayes classifier, which is the least sensitive to missing data, learns effectively without the need to treat missing values, especially if the missing rate is less than 30%. It makes full use of the observed data and, thus, is the most adaptive to missing data [30]. In [31], *k*NN was used to impute missing values in software project datasets. They found that the k-NN imputation can improve the prediction accuracy of C4.5, particularly if the missing data percentage exceeds 40%. A comparison of various imputation machine learning methods has also been published [32]. They concluded that self-organizing map and multi-layer perceptron performed slightly better than regression-based imputation and nearest neighbour method. Comparisons were performed between statistical and machine learning imputation methods in [14] for imputing missing values in breast cancer datasets. Their findings show that machine learning based imputation methods outperformed the statistical ones.

III. Dissolved Gas Analysis

During normal operation, oil-insulated power transformers produce gases such as hydrogen and hydrocarbon, albeit in very small quantities. However, when they experience electrical disturbances or thermal decomposition, the chemical reactions of the insulation involves the breaking down of carbon-hydrogen and carbon-carbon bonds. During this process, active hydrogen atoms and hydrocarbon fragments are formed. These fragments can combine with each other to form gases: hydrogen, methane, acetylene, ethylene, ethane, carbon monoxide, and carbon dioxide. These gases are usually referred to as the fault gases, which stay dissolved at above threshold values in the presence of fault(s). The composition of these dissolved gases is dependent on the type of faults that occur as shown in Table 1 of some common power transformer faults.

DGA uses this strong relationship between certain combination of dissolved gases and their associated fault as the underlying principle in identifying incipient faults in power transformer. In general, a DGA dataset is built by taking oil samples over a period of time at regular interval to discern trends and to determine the severity and progression of incipient faults. Generally, daily or weekly sampling is recommended after start-up, followed by monthly or longer intervals. Routine sampling intervals may vary depending on application and individual system requirements [1].

This study briefly describes the ratio-based diagnostic methods mentioned in the Section I that map the dissolved gases found in the oil sample with corresponding fault. Compared to other tools, these ratios methods have become a standard to diagnose fault based on DGA results [1]. These methods employ an array of ratios of certain key combustible gases as the fault type indicators. These five ratios are:

Ratio 1 (R1) = CH_4/H_2
Ratio 2 (R2) = C_2H_2/C_2H_4
Ratio 3 (R3) = C_2H_2/CH_4
Ratio 4 (R4) = C_2H_6/C_2H_2
Ratio 5 (R5) = C_2H_4/C_2H_6

TABLE 1. COMMON TYPES OF TRANSFORMER FAULTS AND THE KNOWN FAULT GASES ASSOCIATED WITH THEM

No	Gases Present Prominently During Operation	Interpretations
1		Normal operation
2	Nitrogen, carbon monoxide, and carbon dioxide	Transformer winding insulation overheated: key gas is carbon monoxide
3	Nitrogen, ethylene, and methane, some hydrogen and ethane	Transformer oil is overheated: key gas is ethylene
4	Nitrogen, hydrogen small quantities of ethylene and ethane	Corona discharges in oil: key gas is hydrogen
5	Same as item No. 4 with the inclusion of carbon dioxide and carbon monoxide	Corona involving paper insulation: key gas is hydrogen
	Nitrogen, high hydrogen and acetylene, minor quantities of ethylene and methane	High-energy arcing: key gas is acetylene
	Same as item No. 6 with the inclusion of carbon dioxide and carbon monoxide	High-energy arcing involves paper insulation of winding: key gas is acetylene

A. Doernenburg Ratios

This method suggests the existence of three general fault types namely, thermal, partial discharge, and arcing. This method requires significant levels of the gases to be present in order for the diagnosis to be valid. The method utilizes Ratios 1, 2, 3, and 4 and the ratios in the order Ratio 1, Ratio 2, Ratio 3, and Ratio 4 are compared to limiting values, providing a suggested fault diagnosis as given in Table 2.

B. Rogers Ratios

This method utilizes Ratios 1, 2, and 5. The method does not depend on specific gas concentrations to exist in the transformer for the diagnosis to be valid. However, it suggests that the method be used only when the normal limits of the individual gases have been exceeded. Table 3 gives the values for the three key gas ratios corresponding to suggested diagnoses.

C. IEC Ratios

This method also utilizes the same three Ratios 1, 2, and 5 as in the revised version of Rogers Ratios. The key differences are the range of code assigned to each ratio and the number of suggested faults. The IEC Ratios fault interpretations can be divided into 9 different types and they are shown in Table 4.

IV. METHODOLOGY OF STUDY

This study imputes the missing values found a DGA datasets using the four established methods mentioned in the Introduction section with the objective of reducing the inconclusive diagnoses of the gas ratio methods. Fig. 1 shows the methodology of this study in meeting this objective.

A. Mean Imputation

One of the most frequently used methods. This method consists of replacing the missing data for a given attribute by the mean (quantitative attribute) or mode (qualitative attribute) of all known values of that attribute. This study chose the mean because DGA datasets contains continuous variables. It is easy to execute as well as suitable for all patterns of missingness. However, this approach also introduces biases. The main reasons are that it shifts the possible extreme values back to the middle of the distribution, and it reduces variances in the variable being imputed. Sample sizes may be overestimated and correlations may be negatively biased, also, when applying this technique [17].

B. Regression Imputation

This can be accomplished by regressing the variable with missing data on other variables in the data set for those cases with complete data. The estimated regression equation is then used to generate predicted values for the cases with missing data. This technique assumes that the data are MAR or MCAR, which makes it suitable for imputing DGA

TABLE 2. DORNERBURG RATIOS [1]

Suggested Fault Diagnosis	R1 CH4/H2		R2 C2H2/C2H4		R3 C2H2/CH4		R4 C2H6/C2H2	
	Oil	Gas Space	Oil	Gas Space	Oil	Gas Space	Oil	Gas Space
Thermal decomposition	>1.0	>0.1	<0.75	<1.0	<0.3	<0.1	>0.4	>0.2
Partial Discharge (low intensity PD)	<0.1	<0.01	Not significant		<0.3	<0.1	>0.4	>0.2
Arcing (high intensity PD)	>0.1 to <1.0	>0.01 to <0.1	>0.75	>1.0	>0.3	>0.1	<0.4	<0.2

TABLE 3. ROGERS RATIOS [1]

R1 CH_4/H_2	R2 C_2H_2/C_2H_4	R5 C_2H_4/C_2H_6	Suggested Fault Diagnosis
<0.1	>0.1 to <1.0	<1.0	Unit normal
<0.1	<0.1	<1.0	Low-energy density arcing-PD
0.1 to 3.0	0.1 to 1.0	>3.0	Arcing-High energy discharge
<0.1	>0.1 to <1.0	1.0 to 3.0	Low temperature thermal
<0.1	>1.0	1.0 to 3.0	Thermal fault < 700 ^{o}C
<0.1	>0.1	>3.0	Thermal fault >700 ^{o}C

datasets. There is, however, a general tendency to produce underestimates of variances and overestimates of correlations [33].

C. *Expectation Maximization*

This algorithm was introduced by [34]. It capitalises on the relationship between missing data and the unknown parameters of a data model. The basic algorithm consists of two steps: expectation (E step) and maximization (M step). First, separate the data into missing and observed, and establish starting values for the parameters (mean, variance, and covariance). In the E step, using these parameters, compute the predicted scores for the missing data (the expectation). In the M step, using the predicted scores for the missing data, maximize the likelihood function to obtain new parameter estimates. Repeat the process until convergence is obtained. EM assumes that the data are MAR. EM requires a fairly large sample for the estimates to be approximately unbiased and normally distributed [17].

D. *Multiple Imputation*

This method was proposed by Rubin [35] in 1987. The whole MI procedure is made of three steps. They are imputation, analysis, and pooling processes. We applied only the first step, imputation process, of the three steps because our interest is to fill in missing values with estimated values. This simulation method replaces each missing value with $m > 1$ plausible values, which are drawn randomly from their predictive distribution. m is the number of repetition. Imputing a missing value with m simulated values produces m apparently completed datasets and then the mean of m imputed values was filled in the missing value. Post-imputation, MI allows analysts to proceed with familiar complete-data techniques and software. Another positive point is that a large number of repetitions is not always necessary to obtain precise estimates. For example, with 50% missing information, m = 10 imputations is efficient. However, like EM, MI does rely on large sample approximations but works better than EM in small to moderate sample sizes. MI also assumes that the missing data are MAR [17].

E. *kNN Imputation*

We use the *k*-NN algorithm to determine the imputed data, where nearest is usually defined in terms of a distance function based on the auxiliary variable(s). In this method a pool of complete instances is found for each incomplete instance, and the imputed values for each missing cell in each recipient is calculated from the mean or median of the respective attribute in complete instances. Mean is used with continuous attributes, whilst median is suitable for discrete attributes. This study chose mean as DGA dataset contains continuous variables. For the *k*NN method, two parameters need to be determined for achieving high estimation accuracy: the number of nearest neighbour (*k)* and the distance metric: the choice of *k*, the number of neighbours used and the appropriate distance metric. Simulation results have demonstrated that for small datasets, k = 10 is the best choice [36], while [37] observed that k is insensitive to values of k in the range of 10-20. Therefore, this study replaced the missing values with estimated values from 1-10 nearest neighbours depending on the size of datasets. The distance metric used was the Euclidean distance as adopted by [37].

The advantages of the *k*NN imputation are [31]:

- It does not require to create a predictive model for each feature with missing data.
- It can treat both continuous and categorical values.
- It can easily deal with cases with multiple missing values.
- It takes into account the correlation structure of the data.

TABLE 4. IEC RATIOS

R1 CH_4/H_2	R2 C_2H_2/C_2H_4	R5 C_2H_4/C_2H_6	Suggested Fault Diagnosis
0.1 to 1.0	<0.1	1.0 to 3.0	Thermal fault < 150 ^{o}C
>1.0	<0.1	<1.0	Thermal fault 150^{o}C - 300 ^{o}C
>1.0	<0.1	1.0 to 3.0	Thermal fault 300^{o}C - 700 ^{o}C
>1.0	<0.1	>3.0	Thermal fault > 700 ^{o}C
0.1 to 1.0	1.0 to 3.0 and >3.0	1.0 to 3.0 and >3.0	Discharge of low energy
0.1 to 1.0	1.0 to 3.0	>3.0	Discharge of high energy
<0.1	<0.1	<1.0	Partial discharge of low energy density
<0.1	1.0 to 3.0	<1.0	Partial discharge of high energy density
0.1 to 1.0	<0.1	<1.0	normal

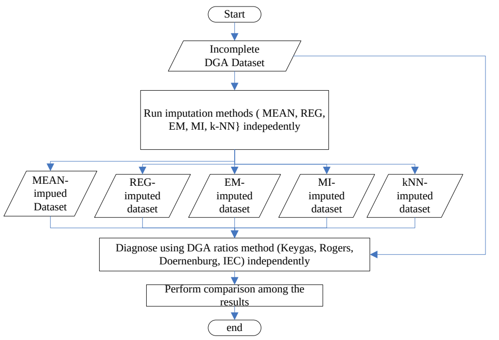
Start
Incomplete DGA Dataset
Run imputation methods (MEAN, REG, EM, MI, k-NN} indepedently
MEAN-impued Dataset
REG-imputed dataset
EM-imputed dataset
MI-imputed dataset
kNN-imputed dataset
Diagnose using DGA ratios method (Keygas, Rogers, Doernenburg, IEC) independently
Perform comparison among the results
end

Figure1. Comparative evaluation of imputation methods

V. EXPERIMENTS AND RESULT

A. DGA Datasets

Two DGA datasets, with different percentages of actual missing values, are imputed and classified in this study. The first DGA dataset (named MAL) is obtained from a local Malaysian utility company which manage various transformers located throughout Malaysia, whilst IECDB10 [38] is the second. The characteristics of the datasets are shown in Table 5. A sample of DGA data consists of a number of dissolved gases in oil and the corresponding fault type as shown in Table 6. Dashes in Table 6 represents missing values (missing gases). Using a gas chromatograph equipped with suitable adsorption columns, the dissolved gas concentrations are measured from the oil samples in parts per million (ppm) by volume of (specific gas) in oil.

B. Experimental setup

This study used SPSS [39] to impute missing values using the MEAN, REG, and EM methods. Meanwhile the MI and *k*NN methods were applied using SOLAS and MATLAB [40], respectively. For the EM method, a normal distribution (the default) of the data was assumed and the default iterations (25) was adopted. This study chose normal variates as the random component to be added for the REG estimation task. For the MI method, this study chose $m = 5$ because according to [16], the MI method does not need a large number of repetitions for precise estimates. Because both datasets in Table 5 were quite small in size, this study chose *k*=1,3,5,7,10 as explained in Section IX.E.

TABLE 5 THE CHARACTERISTICS OF DGA DATASETS USED IN THIS STUDY

	IEC10DB	MAL
Number of samples	167	1228
Number of dissolved gases	7	9
Number of fault type	6	6
Instances with missing values (%)	27.54	76.07
Missing values (%)	7.96	14.21

TABLE 6. DGA DATASET WITH MISSING VALUES (ppm)

H2	O	N2	CH4	CO	CO2	C2H4	C2H6	C2H	Fault
9	3419	29804	5	403	1316	12	1	-	PD
6	5830	50411	-	217	2039	1	-	-	PD
52470	33164	94134	2504	12515	566	640	18	3251	Arcing
-	13877	54589	3	624	2043	56	3	-	Arcing

After independently filling in the missing values in DGA datasets using the compared methods, the three ratios methods were then run independently on each imputed dataset. The effectiveness of each imputation method were evaluated based on the number of unresolved diagnoses reported by each ratio method, the number of wrong diagnoses made by each method and the number of correct diagnoses predicted by each method.

C. Case Study 1:IEC10 Dataset

Table 7 shows the result of diagnoses for each *k* of the kNN method on the IEC10DB dataset for each DGA diagnostic method. It is observed that when *k*=1, all of the three ratios method gave the highest number of correct predictions. Similar observation was seen for the number of wrong diagnoses, all three methods were the least wrong when *k*=1. For the number of unresolved diagnoses, the Rogers and IEC reduced the unresolved cases the most when *k*=1, but Doernenburg did the same when *k*=5. Overall, higher values of *k* do not increase diagnostic performance of the ratio methods on this dataset. It can be explained that the inclusion of complete instances that are significantly different from the instance of interest can decrease accuracy because the neighborhood has become too large and not sufficiently relevant for the estimation task.

Table 8 shows the result of diagnoses for each ratio method on imputed datasets obtained from each imputation method compared in this study. INCOMPLETE is the results of diagnoses from each ratio method before the missing values were estimated by the comparative methods. For the *k*NN method, the result from *k*=1 was used for comparison with the other imputation methods. It can be safely said, for this dataset, replacing missing values with estimated values from each imputation method reduce the unresolved diagnoses compared to INCOMPLETE. The MEAN was the most effective in reducing the number of unresolved diagnosis for Doernenburg. For Rogers and IEC, it was the *k*NN. Consequently, the reduced statistics of unresolved lead to an increase or decrease of wrong and correct predictions for all of the ratio methods. It is seen that, among the compared imputation methods, *k*NN registered the highest number of correct guesses and the least number of wrong diagnoses. Interestingly, four methods (kNN, EM, MI, REG) helped increase the diagnostic performance of all the ratios method in guessing correct faults compared to INCOMPLETE. MEAN however fared worst than INCOMPLETE when applied with Rogers and IEC but did better than INCOMPLETE in Doernenburg. Unfortunately, the number of wrong diagnoses increased for all ratios when combined with the imputation methods in

TABLE 7. DIAGNOSIS OF DGA RATIOS METHOD USING *K*NN WITH DIFFERENT VALUES OF *k*

Diagnose	***k***	**DGA Diagnostic Methods**		
		Doernenburg	Rogers	IEC
Unresolved	1	34.73	**34.73**	**29.94**
	3	35.93	50.90	44.91
	5	**28.74**	51.50	44.91
	7	37.13	51.50	44.91
	10	38.32	52.10	45.50
Wrong	1	**28.14**	**13.17**	**23.95**
	3	29.34	14.37	25.15
	5	28.74	13.77	25.15
	7	29.34	13.77	25.15
	10	28.14	13.77	25.15
Correct	1	**37.13**	**52.10**	**46.11**
	3	34.73	34.73	29.94
	5	34.13	34.73	29.94
	7	33.53	34.73	29.94
	10	33.54	34.13	29.34

comparison with INCOMPLETE. But this is expected with the reduced number of unresolved diagnosis when imputed datasets were used for fault diagnosis.

D. Case Study 2: MAL Dataset

Table 9 shows the result of diagnoses for each *k* of the *k*NN method on the MAL dataset for each DGA diagnostic method. It is observed that when *k*=3, two out of the three ratios method (Rogers and IEC) gave the highest number of correct predictions. The number of unresolved cases were reduced the most when *k* = 3. However, higher *k* (10 for Doernenburg and Rogers, and 5 for IEC) were needed to record lower wrong diagnoses. It can be said that for this dataset, the best *k* needed to improve the performance of each ratio method varies from one ratio method to another. If the number of unresolved guesses is the most important criterion in evaluating the effectiveness of the imputation method, than *k* = 3 was the best. Table 10 shows the result of diagnoses for each ratio method on imputed datasets obtained from each imputation method compared in this study. For the *k*NN method, the result from *k*=3 was used for comparison with the other imputation methods as it reduced unresolved diagnoses the most for all ratio methods.

For this dataset, the effect of each imputation method in reducing the unresolved cases as compared to INCOMPLETE varies. While kNN, EM, MI, and REG met the aforementioned criterion for the three diagnostic method, MEAN, however, only did so for Doernenburg and IEC. For Rogers, MEAN increased the unresolved cases than INCOMPLETE. In increasing correct guesses against INCOMPLETE, only kNN, REG, and EM (albeit very slightly) were successful when paired with all of the three ratios individually. MEAN and MI had mixed results depending on the ratios used. Similar with IEC10DB dataset, *k*NN registered the highest number of correct guesses and the least number of wrong diagnoses for all of the ratio methods on this dataset. This made *k*NN registered the highest number of wrong guesses because the reduction of unresolved were significantly huge compared to other imputation methods.

E. Analysis

As stated above, different experiments are executed on two different datasets and the results show that:

a) imputing missing values in a DGA dataset reduce the number of unresolved diagnoses reported by the three established DGA ratios method when predicting the incipient fault in power transformer. Unresolved diagnoses are a well-known issue with the DGA ratio methods.

b) among the established imputation methods, *k*NN has the best effect to the diagnostic performances of the three DGA ratio methods on both of the datasets. The number of unresolved cases were the least and correct diagnoses were the highest for both datasets. *k*NN also outperformed INCOMPLETE in these criteria.

c) only REG and EM are as consistent as *k*NN in reducing the unresolved guesses produced by the ratios method on both dataset as well as in increasing the number of correct guesses when compared to INCOMPLETE.

d) MI, and MEAN have varied effects to the diagnostic performances of the three ratios methods depending on the datasets used. However, MEAN is the worst performer especially in increasing the number of correct guesses for both datasets for two (Rogers and IEC) ratio methods against INCOMPLETE.

e) *k*NN outperforms the other methods is expected because estimated values are calculated from observed samples having the most similar characteristics with the sample of interest (contains the missing value). *k*NN is a non-parametric method and requires no assumption with regards to pattern of missingness. *k*NN takes account correlation among data structure which coulld be the reason behind its best performance .

f) Interestingly, REG comes second to kNN. As mentioned in Section IX.B, REG is suitable for MCAR and MAR data. Using the chi-square test, both datasets are proved to be MCAR, with *p*-values <=0.001. This could be the possible reason behind the comparable performance of REG with kNN.

g) EM and MI are state-of-the art imputation methods which motivate us to use them with DGA dataset. Experimental results show that they fare behind kNN and REG. According to [16], both assume data are MAR. Because both DGA datasets are MCAR compliance, we assume this could be the reason behind the lesser quality of imputed values produced by them.

h) For both datasets, EM performs better than MI. According to [16], MI estimates better on smaller datasets than EM, but the results using DGA datasets were the opposite.

i) Doernenbug method benefits the most from the imputing procedure where the number of correct guesses increase against INCOMPLETE for most of the experimental settings (only MEAN failed in the MAL dataset).

j) For IEC10DB, the combinations of kNN and Rogers ratio gives the highest correct guesses, whilst for MAL, they are *k*NN and IEC.

VI. Conclusion

The main objective of this research was to investigate the effects of missing values imputation methods on the diagnostic performances of the Doernenburg, Rogers, and IEC for predicting the incipient faults of power transformers. From the experimental results, we can safely say that, in general, imputing missing values can reduce the number of unresolved diagnoses faced by the three diagnostic methods as shown by four out five methods compared in this study. However, the number of correct guesses obtained by the ratios methods vary according to the combination of imputation method and the ratio method. Some combinations increased the correct guesses than INCOMPLETE while other combinations did the opposite. It is to be expected that the number of wrong guesses increase with the reduced number of unresolved cases. It is found that *k*NN brings the best effect to the performances of the three ratio methods compared to the EM, MI, REG, and MEAN methods. The experimental results show that imputing missing values found in a DGA dataset can bring positive effects to the performance of three established DGA diagnostic methods especially in reducing unresolved diagnoses - a common drawback faced by these diagnostic methods. We would like to undertake the study on the impact of imputing missing values to the performance of machine learning algorithms that learn from historic DGA dataset to classify fault in power transformer for future research.

TABLE 8. COMPARATIVE PERFORMANCES OF IMPUTATION METHODS ON THE DGA RATIO METHODS

Diagnose	**Imputation Methods**	**DGA Diagnostic Methods**		
		Doernenburg	Rogers	IEC
Unresolved	INCOMPLETE	45.51	56.89	50.3
	MEAN	33.53	52.69	44.91
	REG	37.13	51.94	45.51
	EM	36.53	50.30	44.31
	MI	36.53	53.29	44.31
	*k*NN	**34.73**	**34.73**	**29.94**
Wrong	INCOMPLETE	25.15	10.78	20.96
	MEAN	35.93	15.57	26.95
	REG	31.14	13.77	24.55
	EM	31.14	15.57	25.15
	MI	31.14	13.17	26.35
	*k*NN	**28.14**	**13.17**	**23.95**
Correct	INCOMPLETE	29.34	32.33	28.74
	MEAN	30.54	31.74	28.14
	REG	31.74	34.73	29.94
	EM	32.33	34.13	30.54
	MI	32.34	33.53	29.34
	*k*NN	**37.13**	**52.10**	**46.11**

TABLE 9. DIAGNOSIS OF DGA RATIOS METHOD USING *K*NN WITH DIFFERENT VALUES OF *k*

Diagnose	***k***	**DGA Diagnostic Methods**		
		Doernenburg	Rogers	IEC
Unresolved	1	58.62	83.33	81.71
	3	**6.67**	**50**	**9.11**
	5	60.73	83.17	83.41
	7	61.14	83	83.09
	10	62.68	83.58	82.85
Wrong	1	22.76	8.78	9.43
	3	76.34	31.87	50.98
	5	21.79	9.10	**9.27**
	7	21.79	9.27	9.51
	10	**20.24**	**8.62**	9.43
Correct	1	**18.62**	7.89	8.86
	3	16.99	**18.13**	**39.92**
	5	17.48	7.72	7.32
	7	17.07	7.72	7.40
	10	17.07	7.80	7.72

TABLE 10. COMPARATIVE PERFORMANCES OF IMPUTATION METHODS ON THE DGA RATIOS METHODS

Diagnose	**Imputation Methods**	**DGA Diagnostic Methods**		
		Doernenburg	Rogers	IEC
Unresolved	INCOMPLETE	84.55	90.65	89.59
	MEAN	68.70	92.60	89.51
	REG	68.78	78.54	76.67
	EM	62.68	88.86	88.04
	MI	60.96	83.58	83
	kNN	**6.67**	**50**	**9.11**
Wrong	INCOMPLETE	3.58	3.09	4.47
	MEAN	**20.98**	**2.36**	**5.69**
	REG	18.54	12.44	11.87
	EM	25.36	4.88	6.01
	MI	24.47	12.36	13.25
	kNN	76.34	31.87	50.98
Correct	INCOMPLETE	11.87	6.26	5.9
	MEAN	10.32	5.04	4.80
	REG	12.68	9.02	11.46
	EM	11.95	6.26	5.93
	MI	14.55	4.06	3.74
	kNN	**16.99**	**18.13**	**39.92**

Acknowledgment

The authors would like to express our appreciation to Universiti Teknikal Malaysia Melaka (UTeM) and the SLAB Programme by Ministry of Higher Education Malaysia (MOHE) for their invaluable supports either technically and financially in encouraging the authors to publish this paper.

References

[1] Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers, IEEE Std C57.104-2008 (Revision of IEEE Std C57.104-1991), 2009.

[2] IEC Publication 60599, “Mineral Oil-Impregnated Equipment in Service - Guide to the Interpretation of Dissolved and Free Gases Analysis, ” March 1999.

[3] M. Duval, “Dissolved gas analysis: It can save your transformer,” IEEE Electrical Insulation Magazine, vol. 5, Nov. 1989, pp. 22-27, doi: 10.1109/57.44605.

[4] S. W. Kim, S. J. Kim, H. D. Seo, J. R. Jung, H. J. Yang and M. Duval. (2013). “New methods of DGA diagnosis using IEC TC 10 and related databases Part 1: application of gas-ratio combinations,” IEEE Transactions on Dielectrics and Electrical Insulation, vol. 20, 2013, pp. 685-690.

[5] W. Q. Zhao, Y. L. Zhu, D. W. Wang, and X. M. Zhai, “A fault diagnosis model for power transformer based on statistical theory,” Proc. IEEE Conf. Wavelet Analysis and Pattern Recognition (ICWAPR'07), Nov. 2007 , pp. 962-966.

[6] M. Dong, D. K. Xu, M. H. Li, and Z. Yan, “Fault diagnosis model for power transformer based on statistical learning theory and dissolved gas analysis,” Proc. IEEE. Symp. Electrical Insulation, Sep. 2004, pp. 85-88.

[7] L. Zhong, Y. Jinsha, and S. Peng, “Fault Diagnosis of Power Transformer Based on Heuristic Reduction Algorithm ,” Proc. IEEE. Conf. Power and Energy Engineering (APPEEC 2009), Mar. 2009, pp. 1-4.

[8] R. Naresh, V. Sharma, and M. Vashisth, “An integrated neural fuzzy approach for fault diagnosis of transformers,” IEEE Transactions on Power Delivery, vol. 23, 2008, pp. 2017-2024.

[9] J. Liu, Y. Liang, and X. Sun, “Application of Learning Vector Quantization network in fault diagnosis of power transformer,” Proc. IEEE. Conf. Mechatronics and Automation (ICMA 2009), Aug. 2009, pp. 4435-4439.

[10] Z. Yong-li and G. Lan-qin, (2006, November). “Transformer fault diagnosis based on naive bayesian classifier and SVR,” Proc. IEEE. Conf. TENCON 2006, Nov. 2006 , pp. 1-4.

[11] B. Twala and M. Phorah, (2010). “Predicting incomplete gene microarray data with the use of supervised learning algorithms,” Pattern Recognition Letters, vol. 31, 2010, pp. 2061-2069, doi: 10.1016/j.patrec.2010.05.006.

[12] T. Yu, H. Peng, and W. Sun, (2011). “Incorporating nonlinear relationships in microarray missing value imputation,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, 2011, pp. 723-731, doi: 10.1109/tcbb.2010.73

[13] K. J. Janssen, A. R. T. Donders, F. E. Harrell Jr, Y. Vergouwe, Q. Chen, D. E. Grobbee and K. G. Moons, (2010). “Missing covariate data in medical research: to impute is better than to ignore,” Journal of clinical epidemiology, vol. 63, 2010, pp. 721-727.

[14] J. M. Jerez, I.. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín and L. Franco, “Missing data imputation using statistical and machine learning methods in a real breast cancer problem,” Artificial intelligence in Medicine, vol. 50, 2010, pp. 105-115.

[15] G. King, J. Honaker, A. Joseph and K. Scheve, “Analyzing incomplete political science data: An alternative algorithm for multiple imputation,” American Political Science Association, vol. 95, Cambridge University Press, 2001, pp. 49-69. .

[16] H. Peyre, A. Leplège and J. Coste, “Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey,” Quality of Life Research, vol. 20, 2011, pp. 287-300.

[17] J. L. Schafer and J. W. Graham, “Missing data: our view of the state of the art,” Psychological Methods, vol. 7, 2002, pp. 147-177.

[18] D. J. Stekhoven and P.Bühlmann, “MissForest—non-parametric missing value imputation for mixed-type data,” Bioinformatics, vol. 28, 2012, pp. 112-118, doi: 10.1093/bioinformatics/btr597.

[19] G. Batista and M.C. Monard, “A study of k-nearest neighbour as an imputation method,” Soft Computing Systems: Design, Management and Applications, 2002, pp. 251-260.

[20] N. Tsikriktsis, “A review of techniques for treating missing data in OM survey research,” Journal of Operations Management, vol. 24, 2005, pp. 53-62, doi: 10.1016/j.jom.2005.03.001.

[21] D.B. Rubin, “Inference and missing data,” Biometrika, vol. 63, 1976, pp. 581-592.

[22] S. Walks, “Moments and distributions of estimate of population parameters from fragments samples,” Ann. math. Stat.3, 1932, pp. 163-203.

[23] R. Little and D. Rubin, “Statistical Analysis with Missing Data,” second ed. John Wiley and Sons, New York, 2002.

[24] M. Glasser, “Linear regression analysis with missing observations among the independent variables,” Journal of the American Statistical Association 59.307,1964, pp. 834-844.

[25] A. A. Afifi and R. M. Elashoff. “Missing observations in multivariate statistics I. Review of the literature,” Journal of the American Statistical Association 61.315, 1966, pp. 595-604.

[26] C.H. Brown, “Asymptotic comparison of missing data procedures for estimating factor loadings,” Psychometrika 48.2 (1983): 269-291

[27] A. Farhangfar, L.A. Kurgan and W. Pedrycz, “A Novel Framework for Imputation of Missing Values in Databases,” IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, vol. 37, 2007, pp. 692-709.

[28] S. Zhang, Z. Jin and X. Zhu, “Missing data imputation by utilizing information within incomplete instances,” Journal of Systems and Software, vol. 84, 2011, pp. 452-459.

[29] Y.T. Mustafa, V.A. Tolpekin and A. Stein, “Application of the Expectation Maximization Algorithm to Estimate Missing Values in Gaussian Bayesian Network Modeling for Forest Growth,” IEEE Transactions on Geoscience and Remote Sensing, vol. 50, 2012, pp. 1821-1831.

[30] L. Lei, W. Naijun and L. Peng. “Applying sensitivity analysis to missing data in classifiers,” Proc. Conf. Services Systems and Services Management (ICSSSM '05), Jun. 2005, pp. 1051-1056.

[31] Q. Song, M. Shepperd, X. Chen, and J. Liu, “Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation,” J. Syst. Softw, vol. 81, 2008, pp. 2361-2370.

[32] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen and M. Kolehmainen, “Methods for imputation of missing values in air quality data sets,”Atmospheric Environment, vol. 3818, 2004, pp. 2895-2907

[33] P.D. Allison, “Missing data techniques for structural equation modeling,” Journal of abnormal psychology, vol. 112, 2003, pp. 545-557.

[34] N. M. L. A. P. Dempster, D. B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” Journal of the Royal Statistical Society, vol. 39, 1977, pp. 1-38.

[35] R. J. A. Litte and D.B. Rubin “Statistical analysis with missing data,” New York: Wiley, 1987.

[36] E. Acuna and C. Rodriguez, “The treatment of missing values and its effect inthe classifier accuracy,” In Banks,D. et al. (eds) Classification, Clustering and DataMining Applications. Springer-Verlag, Berlin, Heidelberg, 2004, pp. 639–648.

[37] O. Troyanskaya, “Missing value estimation methods for DNA microarrays ,” Bioinformatics, vol. 17, 2001, pp. 520–525.

[38] M. Duval and A. dePabla, “Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases,”, IEEE on Electrical Insulation Magazine, vol. 17, 2001, pp. 31-41.

[39] IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

[40] MATLAB, version 7.9.0 (R2009b). 2009, The MathWorks Inc.