Received November 16, 2020, accepted November 20, 2020, date of publication November 25, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040421

Classification of Partial Discharge Fault Sources on SF_6 Insulated Switchgear Based on Twelve By-Product Gases Random Forest Pattern Recognition

NOR ASIAH MUHAMAD[1], (Member, IEEE), IBRAHIM VISA MUSA[2], ZULKURNAIN ABDUL MALEK[3], (Senior Member, IEEE), AND AMMAR SALAH MAHDI[3]

[1]School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Engineering Campus, Nibong Tebal 14300, Malaysia
[2]Department of Electrical and Electronic Engineering, Modibbo Adama University of Technology, Yola, Nigeria
[3]Institute of High Voltage and High Current (IVAT), Universiti Teknologi Malaysia, Skudai 81310, Malaysia

Corresponding author: Nor Asiah Muhamad (norasiah.m@usm.my)

This work was supported in part by the Malaysian Ministry of Higher Education (MOHE), in part by the Universiti Sains Malaysia (USM) through USM Research University Incentive (RUI) under Grant 1001/PELECT/8014054, in part by the Universiti Teknologi Malaysia (UTM), and in part by the Tenaga Nasional Berhad (TNB) Research Sdn. Bhd.

ABSTRACT Sulphur hexafluoride (SF_6) gas insulated switchgear (GIS) is widely used in electrical power supply system and therefore needs regular preventive maintenance. Prediction and diagnosis analysis of faults in GIS using SF_6 gas by-products was introduced previously by using 4 to 8 types of by product gases. As latest development on gas analyser, more by-product gases can be detected and used for condition monitoring of the GIS. The type, number, concentration and chemical stability of by-product gases of SF_6 GIS are found to be closely correlated to the type of defect. However, the number of by-product gases used increases, the pattern for faults classification become more complex. Thus, further analysis on increasing number of by product gases using intelligent techniques such as pattern recognition is required. In this article, 12 significant by-products captured due to various sources of partial discharge fault in GIS were used. Random Forest (RF) was selected in this work as a multi-class classification technique. The analyses using RF pattern recognition with eight algorithms based on the presence and concentration of the gas by-products were carried out. The RF algorithm successfully recognises a given defect with an accuracy of 87.5% for all defects fault classification. The performance of the RF algorithm is 1.5 times better than the decision table algorithm which is the next best algorithm. This research illustrates the feasibility and applicability of an effective GIS diagnostic using gas by-products analyses, and in particular, using the RF pattern recognition.

INDEX TERMS Gas insulated switchgear (GIS), Sulphur hexafluoride (SF_6), partial discharge (PD), insulation, random forest (RF) pattern recognition.

I. INTRODUCTION

Partial discharge (PD) is the localised breakdown which occurs under high voltage stress in a small portion of solid, liquid or gas insulators. It does not completely bridge space between two conductor electrodes, and in gas, the localised breakdown is usually less than 1 mm [1], [2]. PD occur due to the defects of irregularities or protrusion on high voltage or earth electrodes, free or floating metallic particles, contamination on the surface of spacers, and void or gap at electrode/dielectric interface. These defects may course by the mechanical abrasion movement of the conductor during load cycling, error in manufacturing of gas insulated switchgear (GIS), vibration during shipment of GIS equipment and assembling, undetected scratch on electrodes, and poor electrical contacts [1], [3]–[6]. Partial discharge is accompanied by the emission of energy as electromagnetic emission in the form of the radio wave, light and heat; acoustic emission in the form of sound at ultrasonic ranges; and emission of ozone and nitrogen oxide. Although the

The associate editor coordinating the review of this manuscript and approving it for publication was Siddhartha Bhattacharyya.

magnitudes of partial discharge are small (about 5-10eV), they cause drastic reduction of onset corona and breakdown voltage, which leads to gradual progressive deterioration or degradation of insulation in GIS, eventually leading to ultimate failure of the GIS [1], [5], [7]–[14]. About eighty-five percent of GIS disruptive failure is caused by PD [3], [4], [15], therefore, there is a need to employ an effective method to identify the sources and to diagnose the activities of PD in GIS.

Some research on PD diagnostic sensing techniques namely, light, acoustic, electrical, ultra-high frequency, and chemical by-product techniques, has been conducted to diagnose PD in GIS [1], [16]–[18]. A key factor in the use of the chemical by-product diagnostic technique for PD detection is the gas analysis method. The presence of sulphur hexafluoride by-product gases is an indication of fault or discharges in the GIS and this may affect its health. SF_6 gas sample analysis is useful in determining the health condition of gas insulated systems. The degradation of sulphur hexafluoride gas as a result of PD leads to the generation of by-product gases such as SOF_4, SOF_2, SO_2F_2, SF_4, S_2F_{10}, CF_4, CO, COS, CO_2, H_2S, HF, SO_2 and SiF_4 [19]–[23]. These type of by-product gases and their concentrations are very useful information for gas analysis.

Previous researchers have used various algorithms, such as the Rogers ratio, the Doernenburge ratio, the IEC ratio, the Duval triangle, the artificial neural network and the key gas, in the dissolved gas analysis to diagnose faults in transformers. These methods can also be used for SF_6 gas analysis because their performance is good to a certain extent. However, when there are too many different types of gases, it is difficult to analyse [24], [25]. Moreover, with the emergence of infrared absorption spectrometry, gas chromatography and other mature detection, researchers shifted focus towards quantitative analysis of gas decomposition products under different discharge conditions like arc discharge, spark discharge and corona discharge [11]. More decomposition of SF6 gases can be detected now. Most of GIS decomposition diagnosis is based on 6-8 types of gases. As new technologies emerge, this work aims to use 12 by-product gases to conduct faults classification.

Various approaches have been used to identify the type of defect in GIS chambers by using SF6 decomposition products in previous research. Tang *et al.* [18] have used Space Vector Machine (SVM) to classify four types of defects. The parameters of the algorithm were optimized by using a Particle Swarm Optimization (PSO) algorithm. However, the performance of SVM depends significantly on its parameters. Moreover, the algorithm should be generalized when used for multi-class classification problems. Zeng *et al.* [26] have used Backpropagation Neural Network (BPNN) to identify four types of defects under DC PD. High accuracy rate was obtained when the concentration ratio was selected as a feature parameter. However, if BPNN is not generalized, overfitting can occur. The concentration ratio of decomposition products was selected as a feature parameter and applied to the decision tree algorithm to recognize four types of PD sources. The authors concluded that the decision tree performed well at PD recognition [17]. However, the change of training data can give variant results, hence, more than one tree is required to tackle the variance issue. Similar to the Duval triangle used in the diagnosing of transformers, a graphical triangle method is introduced by L. Zong *et al.* [27] to visualize defects in GIS. The three sides of the triangle represent the relative proportion of decomposition products, and the triangle is divided into zones, each representing the types of defect. The authors found that the methods are reliable to distinguish initial faults in GIS. However, this technique depends on typical decomposition products and defining the boundaries and zones of each fault.

In this article, we suggest Random Forest (RF) as a multiclass classification technique. In RF, the overfitting issue can be reduced by averaging the output of a large number of trees, making it able to predict the output of new data and reducing the variance issue. Also, as it is based on node splitting, the number of features is not important. In this research we employed the Waikato Environment for Knowledge Analysis (WEKA) machine learning workbench and data mining written in Java. This is because its performance is good for small data, perform fast and can be used for onsite diagnoses [17], [28]–[31].

The data used for this work is based on a simulation experiment of PD faults activities in SF_6 GIS. Eight types of PD defects were done (free conducting particle, electrode to dielectric void, electrode protrusion, fixed particle aluminium on spacer, fixed copper particle on spacer, electrode protrusion-fixed copper particle hybrid, electrode protrusion-free copper particle hybrid, and electrode to dielectric void-free copper particle hybrid). A total of 12 gas by-products significance gases were found. In this work, the pattern recognition of the defects can be divided in four categories. First, pattern recognition of sole defect (free conducting particle, electrode protrusion, electrode-dielectric void and fixed conducting particle). Second, pattern recognition of hybrid defects (electrode protrusion-fixed conducting particle, electrode protrusion-free conducting particle and electrode-dielectric void with free conducting particle). Third, pattern recognition of material dependent defect (fixed aluminium and fixed copper). Finally, in the fourth category, all the defects are combined and analysed at once using pattern recognition.

II. TREE-BASED MODEL LEARNING ALGORITHM

A tree-based learning algorithm is thought to be one of the best and most utilized administered learning techniques. A tree-based technique engages prescient models with high precision, security, stability and simplicity of elucidation. Unlike all linear models, it can map nonlinear relationships very well. It is versatile at tackling any sort of problem at hand (classification, prediction, inference or regression) [30], [31]. The tree-based model learning includes the decision tree

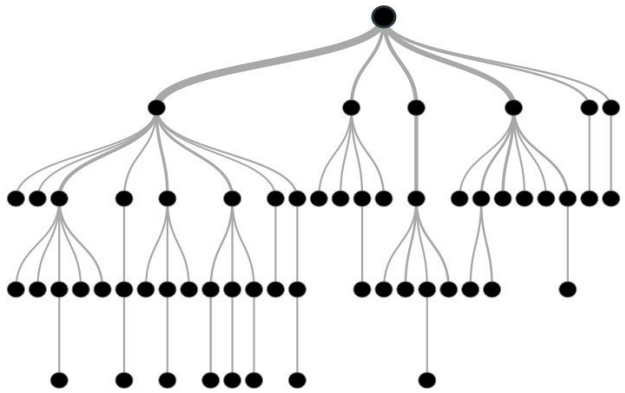

FIGURE 1. Decision tree structure [32].

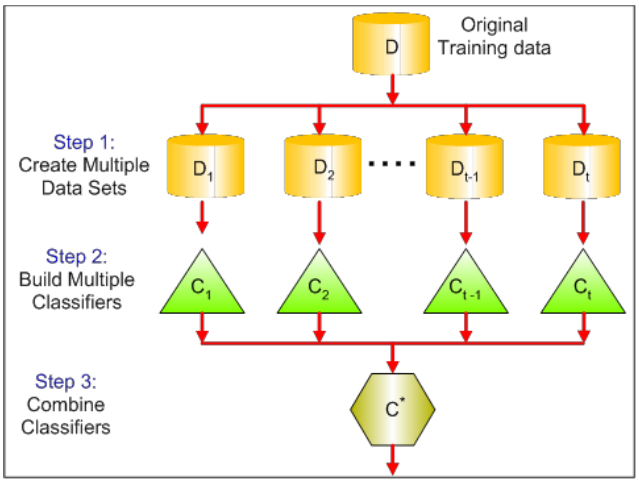

FIGURE 2. Model of random forest [32].

algorithm (J48), the random forest (ensemble decision tree) algorithm, the hoeffding tree, the rep tree, among others. Techniques like the decision tree and the random forest are by and large prevalently utilized in all kinds of data science problems [30]–[34].

A. DECISION TREE ALGORITHM

A decision tree structure is illustrated in Figure 1. A decision tree is an analytical decision support tool algorithm in machine learning that uses a tree flowchart-like structure (tree-like graph). It is a model of resolution and its feasible consequences include chance event outcome, utility and resource cost [17], [31]–[33].

The decision tree is divided into two types; categorical and continuous variable decision tree. The categorical variable decision tree is one with categorical target variables, such as Yes or No; while the continuous variable decision tree assumes the target, variable is continuous. The advantages of a decision tree are that it is easy to understand by people with a non-analytical background; it is fast and useful in data exploration with class data cleaning compared to other modelling algorithms; the data type is not constrained, that is, it handles both nominal and numerical variables; and it has no classifier structure or assumptions about the space distribution (non-parametric method). However, one of the disadvantages of using decision trees is overfitting and lack of fit to some continuous variable [30]–[32], [34].

B. RANDOM FOREST

The model of a random forest is shown in Figure 2. A random forest is a flexible ensemble machine learning algorithm equipped for performing both classification and regression tasks where a group of models is joined to frame an effective model (bagging technique). It also undertakes dimensional reduction approach (reduction of prediction variance), treats outlier values, treats missing values and other important steps of data analysis. A random forest is simply a group of classifiers that is made up of many decision trees and output the class by individual trees [30]–[32], [34]. A random tree algorithm has all the advantages of decision because it is the combination of the decision trees. The challenge of overfitting or not fitting to a continuous variable of the decision tree are overcome using random forest [30]–[32], [34].

III. PATTERN RECOGNITION USING WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS (WEKA)

WEKA is a machine learning workbench and data mining software written in Java under the GNU general public licence and was developed by the University of Waikato in New Zealand [31], [32]. It is a collection of machine language algorithms with main features that comprise a comprehensive set of data pre-processing tools, learning algorithms, evaluation methods, graphical user interface (GUI), including data visualization and a comparing learning algorithm environment. WEKA has a program for classification that includes the random forest, the J48 decision tree, the decision stump, the decision table, the Part, the one R (one rule), the multilayer perceptron, and sequential minimal optimization (SMO) and many others. Programs for numeric prediction include linear regression, multilayer perceptron, model tree generators, instance-based learners, locally weighted regression, decision tables and clustering programs, which include bagging, regression through classification, stacking, boosting, cost-sensitive classification, classification through regression and cobweb.

The following algorithms were used for this work: the random forest, the J48 decision tree, the decision stump, the decision table, the Part, the one R (one rule), the multilayer perceptron, and the SMO. Further detail on these algorithms can be found in [30]–[32], [34]. The following sections describe the details of WEKA parameters to be computed in each of the chosen algorithms. WEKA produces a set of parameters including ten-fold cross-validation, confusion matrix, accuracy, precision, recall, specificity, f-measure, threshold receiver operating characteristic (ROC) curve, Matthew correlation coefficient (MCC) and precision-recall curve (PRC), that can be used in the application of each algorithm. The procedure of pattern recognition using the workbench is shown in Figure 3.

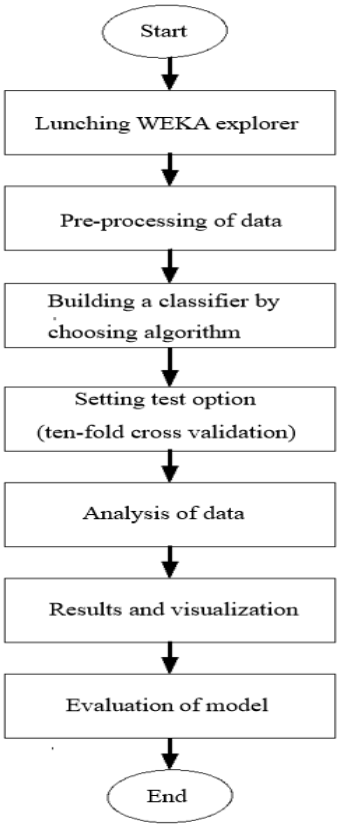

FIGURE 3. Pattern recognition procedure using WEKA workbench.

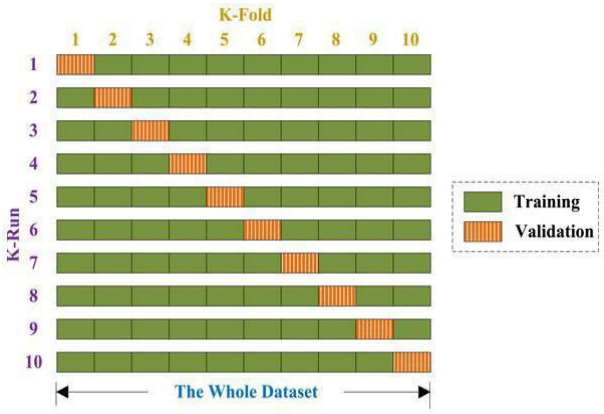

FIGURE 4. Ten-fold cross-validation process [29].

TABLE 1. Confusion matrix.

		Actual Value		
		Positive	Negatives	Total
Predicted Value	Positive	True Positive **(TP)**	False Positive **(FP)**	P`
	Negative	False Negative **(FN)**	True Negative **(TN)**	N`
	Total	P	N	P+N

A. TEN-FOLD CROSS-VALIDATION

The ten-fold cross-validation process is shown in Figure 4. Cross-validation is a default test or evaluation mode. It is the most widely recognized type of validation test, where the data are divided arbitrarily into z folds, Y1, Y2, Y3, …, Yk of approximately equal size in a k-fold cross-validation. Training and testing are carried out k times. In every iteration, one of the subsets is held as the test set while the others are utilized as the training set, that is in cycle 1, subset 1 is saved as the test set, the other partitions are utilized as the training set while in cycle k, subset k is saved as the test set and the rest are utilized as the training set.

The precision is obtained from the ratio of the total number of correctly classified instances in the k iterations to the total number of instances in the whole data. Cross-validation is worthwhile in that regardless of the number of partitions, each data set will be a test set precisely once and it will be in the training set k-1 times. On the other hand, calculation time takes as much time as the segment since training has to take place k times. Ten-fold cross-validation is mostly used where the data is parcelled into 10-folds and 10 times train on 9 folds, and test on the remaining one [29], [32].

B. CONFUSION MATRIX

The confusion matrix is also called an error matrix. It is a table representation of classification, prediction or inference results are shown in Table 1. The term confusion originates from the way that the table makes it simple to know whether a classifier is mistaking one class for the other. The confusion matrix enables the result to be seen easily, and it is suitable for analysis of the classifier in view of its acknowledgement on the class each instance belongs to. It portrays the performance of each classifier in light of the test data whose genuine class are known. The results are produced with the accompanying parameters [29], [32].

i. True Positive: Cases where YES is anticipated and the genuine class is YES
ii. True Negative: Cases where NO is anticipated and the genuine class is NO
iii. False Positive: Cases, where YES is anticipated yet the genuine class is NO
iv. False Negative: Cases where NO is anticipated yet the genuine class is YES

C. ACCURACY AND ERROR RATE

Accuracy is otherwise called recognition rate. It is a general term which is utilized to allude to the prescient capacities of the classifier and also is the rate of accurately characterized instances. The exactness of a classifier or built model cannot be ascertained based on the training data, but on the test data which have class-labelled instances that were not part of the training data. The adequacy of accuracy is shown when the distribution of the class is relatively balanced. The error rate is the opposite of accuracy. It is the percentage of classified

instances that are incorrect. The equation for inferring accuracy and error rate of a classifier from the disarray (confusion) matrix are given in equations (1) to (3) below.

$$Accuracy = \frac{TP + TN}{P + N} \tag{1}$$

$$Error\ Rate = \frac{FP + FN}{P + N} \tag{2}$$

or

$$Error\ Rate = 1 - Accuracy \tag{3}$$

It is not vital to assess a classifier only on its accuracy because infrequently, a classifier with lower accuracy may have higher prescient power. In this manner, there are other assessment measurements to validate a model [29].

D. PRECISION

Precision is a measure of accuracy. It demonstrates the rate of occasions named positive that are really positive. Precision depends on importance, that is, how pertinent are the occasions classified being positive [29]. The formula is given in equation (4) below.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

E. RECALL OR SENSITIVITY

Recall is the measure of culmination and also known as sensitivity. It is also known as positive acknowledgement rate. Recall is the proportion of positives accurately named positive or the measure of positives effectively distinguished accordingly. It portrays how much the classifier stays away from false negatives. The equation is given accordingly [29], [32] as in equation (5) below.

$$Recall = Sensitivity = \frac{TP}{TP + FN} = \frac{TP}{P} \tag{5}$$

F. SPECIFICITY

Specificity is known as true negative acknowledgement rate. It is the extent of negative occurrences which are effectively recognized in that capacity. Specificity portrays how much the classifier stays away from false positives [29]. Equation (6) for specificity is given below.

$$Specificity = \frac{TN}{N} \tag{6}$$

G. THRESHOLD RECEIVER OPERATING CHARACTERISTIC CURVE

Threshold ROC curve is a visually useful tool for machine learning that is used for comparing two classification models through graphical plotting. It illustrates the diagnostic ability of a model by plotting the true positive rate (TPR) known as the probability of detection in machine learning or sensitivity, against false positive rate (FPR) known as fall out or probability of false alarm. The Y axis is the TPR while the X axis is the FPR, while measure of the accuracy of a model is the area under the ROC curve [29], [32]. Figure 5 shows a sample illustration of a ROC curve.

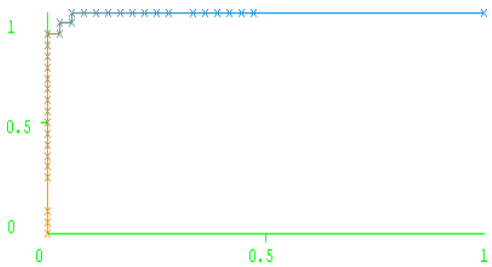

FIGURE 5. Sample of threshold curve for class 1, x and y axis are efficiency in per unit plot. (Area under ROC=0.9958).

H. MATTHEW CORRELATION COEFFCIENT (MCC)

MCC is a learning machine tool introduced in 1975 by the Biochemist Brian W. Matthew, and it is used as a measure of the quality of the classification of models. It is the balance measure that takes account of true and false positive and negatives, even if the size of the classes is not the same. MCC is the correlation classification coefficient of the model between the predicted and the observed model [31], [34].

I. PRECISION-RECALL CURVE (PRC)

PRC is the plotting of precision against recall for all potential cut offs for a test. It is the measure of the accuracy of a model through the area under the PRC [31], [34].

IV. SF_6 BY-PRODUCT GASES DATA

In this study, the gas analysis diagnostic technique with the method of Fourier Transform Infrared (FTIR) spectrometer as a test instrument, was adopted using prototype simulated coaxial gas insulated chamber similar to real life GIS. Eight artificial types of PD defects were used and eleven SF_6 by-products were detected in sole, hybrid and material dependent defect respectively. In total, twelve gases where detected in all the eight artificial defects as shown in Table 2.

V. PATTERN RECOGNITION PRE-PROCESS

The pattern recognition pre-process starts with the launching of the WEKA graphical environment (graphical user interface chooser) application called explorer for the four categories of pattern recognition: sole, hybrid, material dependent and overall defects.

A. PRE-PROCESS ON SOLE DEFECT

The pre-process of pattern recognition was carried out on four types or classes of sole defects: the free conducting particle, electrode protrusion, electrode- dielectric void and fixed conducting particle. The results of the pre-process on sole defects are represented in seven attributes or categories: number of by-product gases, identity of by-product gases and concentrations of by-product gases for all stress durations at 10, 20, 30, 40, and 50 hours.

TABLE 2. SF_6 by-products gases.

Defect	SF6 by-products
Free conducting particle	C_3F_8, C_2F_6
Electrode protrusion	SO_2, SOF_2, SO_2F_2, S_2F_{10}, SiF_4, CO, C_3F_8, C_2F_6
Electrode-dielectric void	SO_2, SO_2F_2, SiF_4, C_2F_6
Fixed conducting particle (Cu)	HF, SOF_2, SOF_4, SO_2F_2, S_2F_{10}, SiF_4 SO_2, CO, C_2F_6, CF_4
Fixed conducting particle (Al)	SO_2, SO_2F_2, C_3F_8, C_2F_6
Electrode-dielectric void with free conducting particle	C_3F_8, C_2F_6
Electrode protrusion-fixed conducting particle (Cu)	HF, SOF_2, SO_2F_2, SiF_4, SO_2, CO, COS, C_3F_8 C_2F_6 CF_4
Electrode protrusion-free conducting particle	SO_2, HF, SOF_2 SO_2F_2, S_2F_{10}, SiF_4, CO, C_2F_6

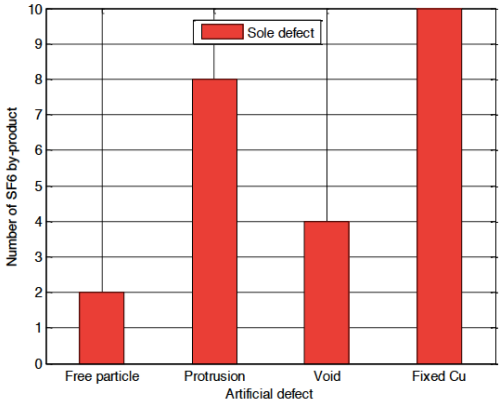

FIGURE 6. Number of by-product gases for each type of sole defect.

The first attribute is the class attribute, defined as the number of by-product gases produced by each defect. The attribute is based on all gases appearing in the whole experiment (50-hour stress duration) for a given defect. Figure 6 shows the number of SF6 by-products detected for all four types of defect in the sole defect category. As can be clearly seen, the fixed conducting particle defect has the highest number of gases (10). Therefore, it can be said that the fixed conducting particle defect produces the most severe effect among the types of the sole defect, followed by the electrode protrusion defect (8 gases), electrode-dielectric void defect (4 gases), and free conducting particle defect (2 gases).

The second attribute involves the identity of the by-product gases produced by each defect. The attribute is based on all gases appearing in the whole experiment (50- hour stress duration) for a given defect. This is shown in Figure 7.

As previously noted, the types of by-product gases produced by the partial discharge activity are dependent on the defect type. Also as previously discussed, the gas identification technique can be used to indicate the type of defect in which partial discharges occur.

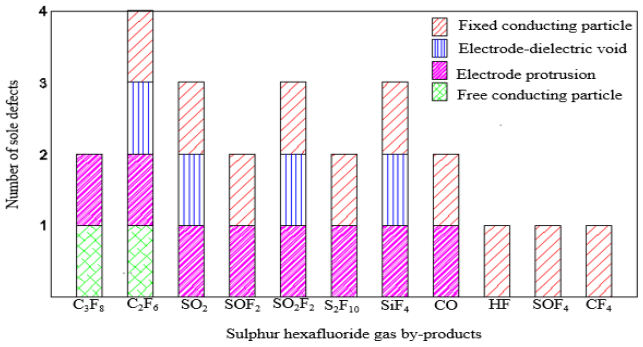

FIGURE 7. Types of by-product gases produced by all sole defects plotted as a stacked histogram.

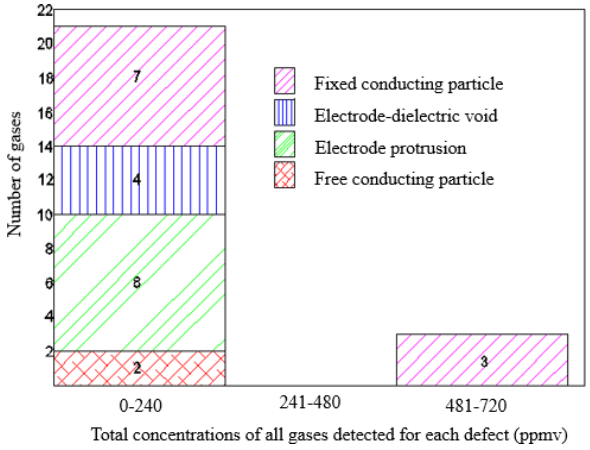

FIGURE 8. Frequency of the appearance of gas concentrations at 10-hour stress duration for all sole defects.

The third attribute is concentration of by-product gases. In this attribute a stacked histogram of the defect type as a function of the concentration of all produced by-product gases is plotted. The concentration of each of the produced by-product gases is given in three ranges of concentration. The exact ranges are dependent on the stress duration at which time the measurement was made. Figure 8 illustrates the frequency of the appearance of gases with a concentration falling in a given concentration range for the case of a 10-hour stress duration. The concentrations are given in three ranges: 0-240, 240-480 and 481-720 ppmv. Figure 8 shows that the fixed conducting particle defect produces 3 gases with the highest concentration range (481-720 ppmv). Other faults only produced gases at lower concentration and electrode protrusion with a higher number of gases compared to others at low concentration range.

The frequency of the appearance of gas concentrations at 20-hour stress duration is shown in Figure 9. Again, as in the case of the 10-hour stress duration, the fixed conducting particle defect also produces gases with the highest concentration (2 gases). The concentration of gases at this duration (maximum 3222 ppmv) had increased drastically compared to the 10-hour stress duration (maximum 720 ppmv). The concentration was divided into three ranges: 0-1074, 1075-2144 and 2145-3222 ppmv. At the

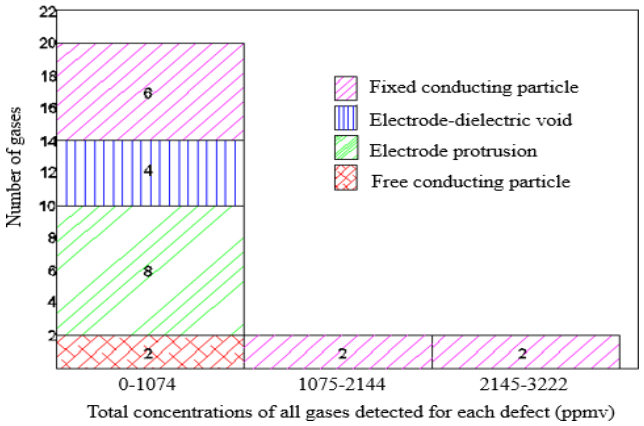

FIGURE 9. Frequency of the appearance of gas concentrations at 20-hour stress duration for all sole defects.

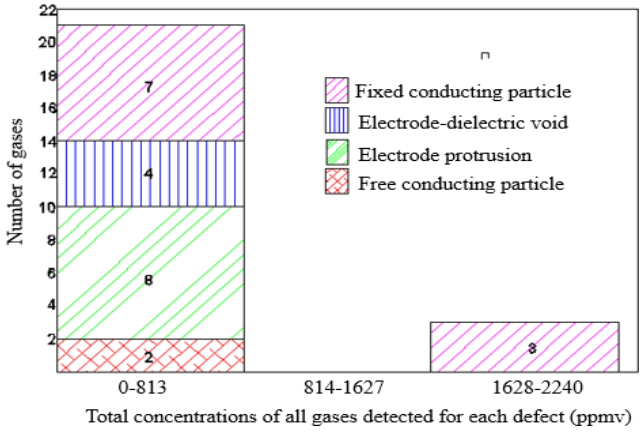

FIGURE 10. Frequency of the appearance of gas concentrations at 30-hour stress duration for all sole defects.

20-hour stress duration, the number of gases in the middle range for fixed conducting particle defect had increased from zero to 2 gases. This had reduced the number of gases for lower and higher range. For other types of faults, the ranges of gases concentration remained the same as in the 10-hour stress duration.

Figure 10 represents the frequency of the appearance of gas concentration for the 30-hour stress duration. As in the previous two cases, the highest concentration of gases is produced by the fixed conducting particle (Cu) defect. Figures 8 and 9 show same pattern of gases concentration detected. The results show that the higher severity of fixed conducting particle defect compared to other three sole defects for stress duration up to 30 hours.

The frequency of the appearance of gas concentration at 40-hour stress duration for all sole defects is shown in Figure 11. Starting from 30 hours stress duration, the concentration of gases for fixed conducting particles is not detected. The electrode protrusion defect produces 2 gases with the highest concentration range and have the highest number of gases detected at all ranges.

The frequency of the appearance of SF_6 by-products concentration at the 50-hour stress duration is shown

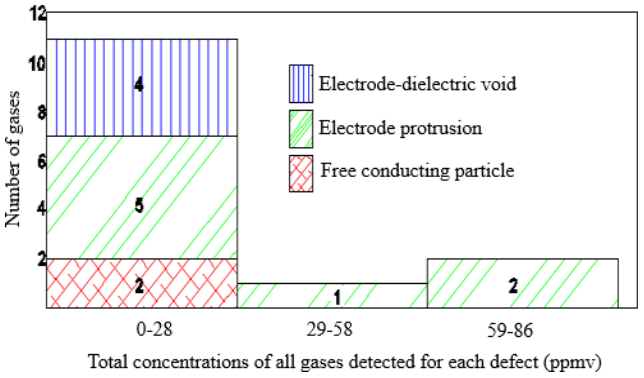

FIGURE 11. Frequency of the appearance of gas concentrations at 40-hour stress duration for all sole defects.

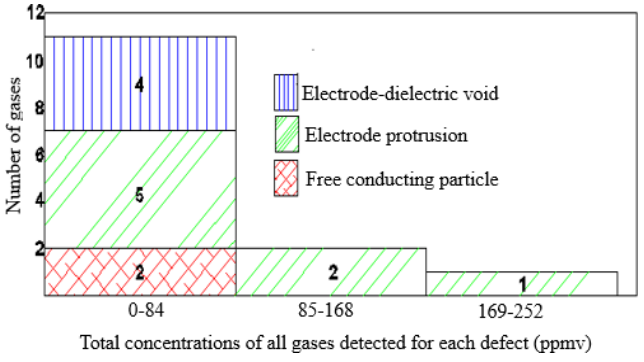

FIGURE 12. Frequency of the appearance of gas concentrations at 50-hour stress duration for all sole defects.

in Figure 12. For the 50-hour duration, the protrusion defect gives one gas at the highest concentration range (169-252 ppmv).

From the results shown above, it can be concluded that the severity of a discharge in SF6 medium is dependent on the type of by-product gases, the number of by-product gases and the frequency of the appearance of by-product gas concentrations. Based on the pre-process analyses, we conclude that the severity of a fault or discharge is caused by a single defect in a gas insulated, and this can be listed in the following order: fixed conducting particle (Cu), electrode protrusion, void and free conducting particle.

B. PRE-PROCESS ON HYBRID DEFECT

Three classes of hybrid defect namely, electrode protrusion-fixed conducting particle (Cu), electrode protrusion-free conducting particle and electrode-dielectric void with free conducting particle, were used in WEKA for pattern recognition. The results of pre-process on hybrid defect are represented in seven attributes as it was done for sole defect in section A above.

The number of by-product gases attributes is based on all gases appearing in the entire experiment (50-hour stress duration) for a given defect. The number of by-product gases detected in the hybrid defects is shown in Figure 13. The highest number of by-product gases was produced by the protrusion-fixed Cu particle defect; thus, the severity of

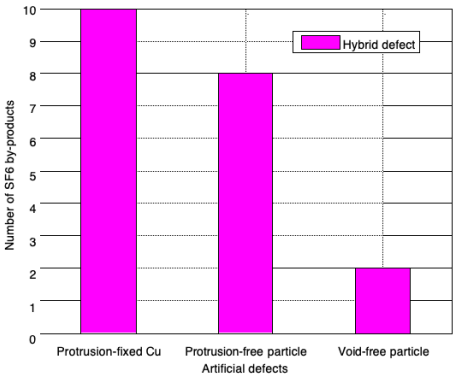

FIGURE 13. Number of by-product gases as a function of hybrid defect type.

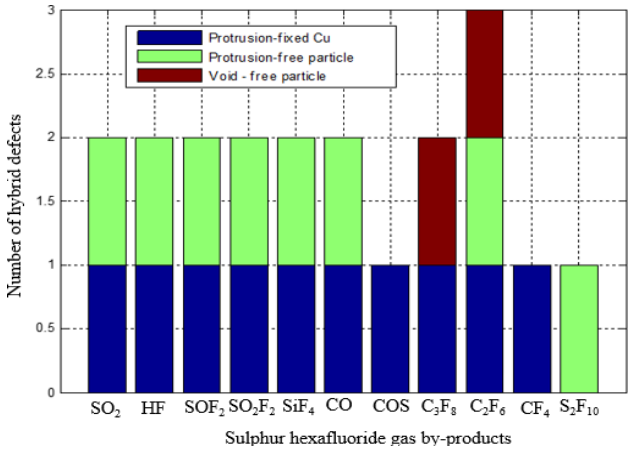

FIGURE 14. Types of by-product gases for the hybrid defects.

this fault was the worst among all defects in this hybrid category.

Figure 14 shows the stacked histogram of the second attribute, which is the identity of the by-product gases produced by all defects in the hybrid category. Again, the type of by-product gases varies with the defect used in the PD experiment.

Figure 15 shows the frequency of the appearance of various concentrations of the SF_6 by-products at the 10-hour stress duration. The groupings of the concentration show that the electrode protrusion-fixed Cu particle defect produces gases with the highest concentration.

The frequency of the appearance of various concentrations of the SF_6 by-products for the 20-hour stress duration is illustrated in Figure 16. At 20-hour duration, the plot shows that the electrode protrusion-fixed copper particle defect produces 1 gas with the highest concentration (587-878 ppmv).

Figure 17 shows the frequency of the appearance of various gas concentrations at the 30-hour stress duration for the three types of hybrid defects. The figure shows that the electrode protrusion-fixed copper particle produces one gas with the highest concentration (511-764 ppmv).

For the 40-hour stress duration, the frequency of the appearance of various gas concentrations is illustrated in Figure 18. Again, the electrode protrusion-fixed Cu particle defect produces 2 gases with the highest concentration (589-881 ppmv).

Figure 19 shows the frequency of the appearance of various gas concentrations at the 50-hour stress duration for the three types of hybrid defects. As Figure 19 shows, the severity of

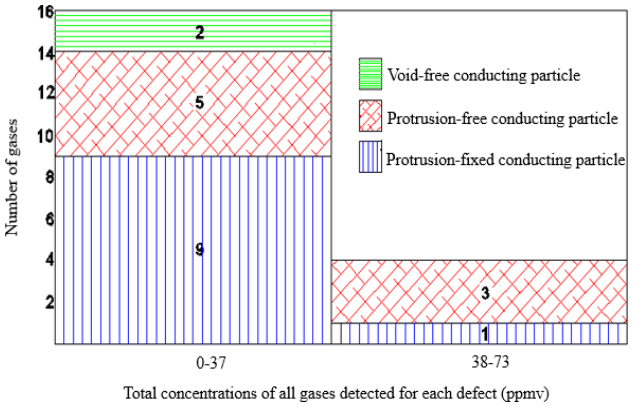

FIGURE 15. Frequency of the appearance of gas concentrations at 10-hour stress duration for all hybrid defects.

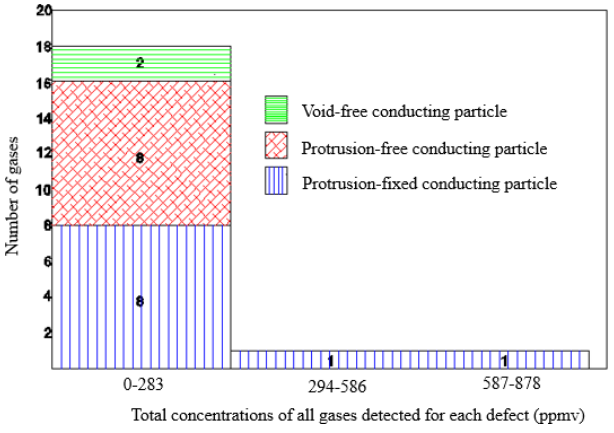

FIGURE 16. Frequency of the appearance of gas concentrations at 20-hour stress duration for all hybrid defects.

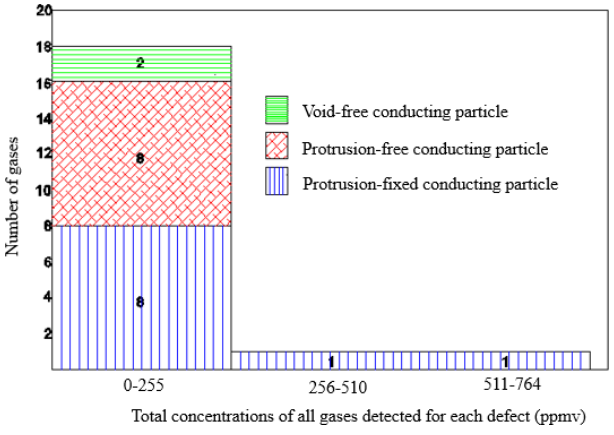

FIGURE 17. Frequency of the appearance of gas concentrations at 30-hour stress duration for all hybrid defects.

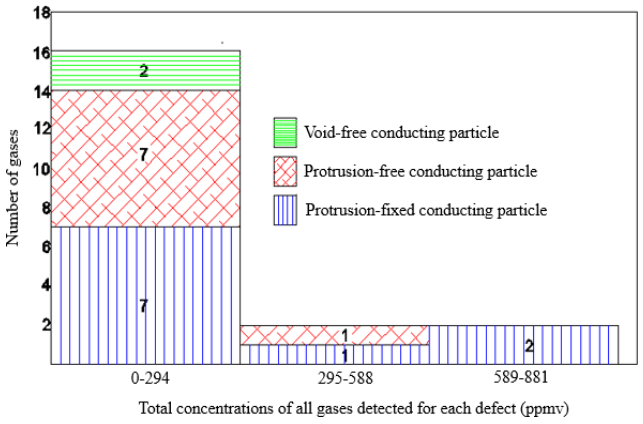

FIGURE 18. Frequency of the appearance of gas concentrations at 40-hour stress duration for all hybrid defects.

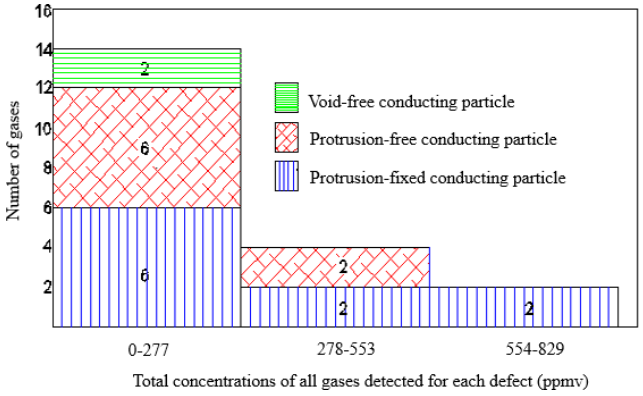

FIGURE 19. Frequency of the appearance of gas concentrations at 50-hour stress duration for all hybrid defects.

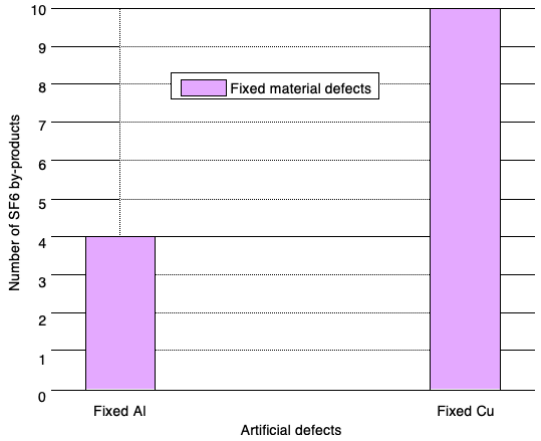

FIGURE 20. The number of by-product gases for both material dependent defects.

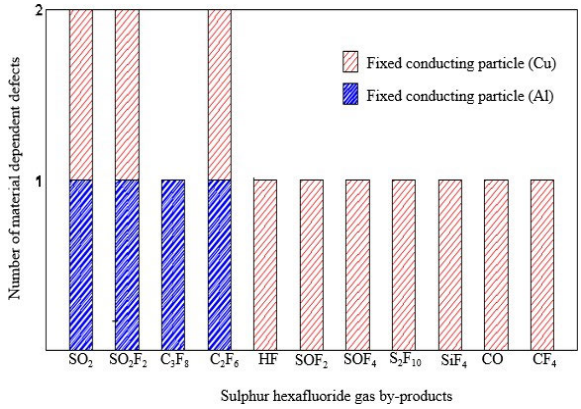

FIGURE 21. Type of SF6 by-products of two material dependent defects.

the electrode protrusion-fixed Cu particle is the highest with 2 gases appearing at the highest concentration range.

From the above results, the pre-process analysis of the hybrid defects shows that the order of severity of the discharges caused by a hybrid defect in a gas insulated system is as follows, starting from the most severe: the electrode protrusion-fixed conducting particle defect, the electrode protrusion-free conducting particle defect, and the electrode to dielectric void-free conducting particle defect.

C. PRE-PROCESS ON MATERIAL DEPENDENT DEFECT

The pattern recognition pre-process of by-product gases produced in the partial discharge experiments using material dependent defect was conducted. Two defects were used, namely, the fixed aluminium particle and the fixed copper particle. The results of the pre-process are described in terms of the number and identity of gas by-products, as well as the concentration variation with stress duration.

The number of by-product gases attributes is based on all gases appearing in the entire experiment (30-hour stress duration) for a given defect. The graph showing the number of by-product gases in the material dependent defect is shown in Figure 20. The number of by-product gases detected in the fixed aluminium (Al) particle and the fixed Cu particle defects are 4 and 10, respectively. Based on this result, it can be deduced that the fixed Cu particle defect produces a more severe PD effect than the fixed Al defect.

The identity of by-product gases attributes is based on all gases appearing in the entire experiment (30-hour stress duration) for a given defect. The result on by-product gases produced by two types of material dependents defect is shown in Figure 21. The two material dependent defects collectively produced 11 by-product gases. It can be presumed that the generation of these by-product gases is due to the material dependent defects.

A stacked histogram of by-product gases of the material dependent defects as a function of concentration range (10-hour stress duration) is shown in Figure 22. As seen on the concentration plot, the fixed Cu particle defect produces 3 gases with the highest concentration (361-720 ppmv). Similar results were also obtained for 20-hour and 30-hour stress durations. Of note is that the severity of fault caused by the fixed copper particle defect in a gas insulated system is higher compared to that of the fixed Al particle defect.

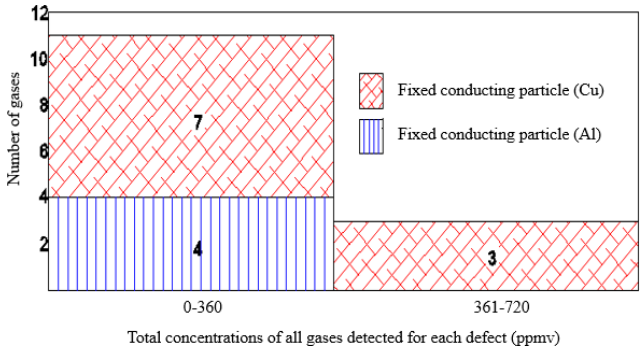

FIGURE 22. Frequency of the appearance of gas concentrations at 10-hour stress duration for both material dependent defects.

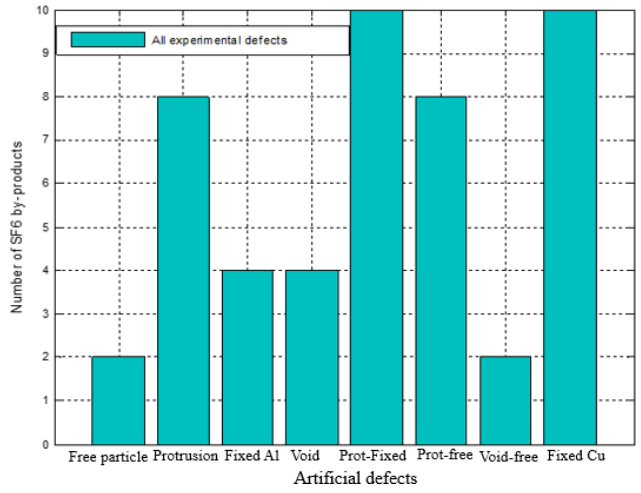

FIGURE 23. Number of by-product gases for all defects.

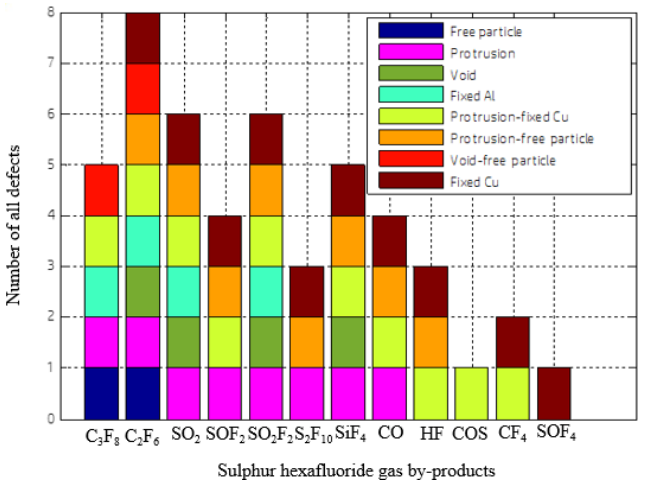

FIGURE 24. Types of partial discharge by-products for all tested defects.

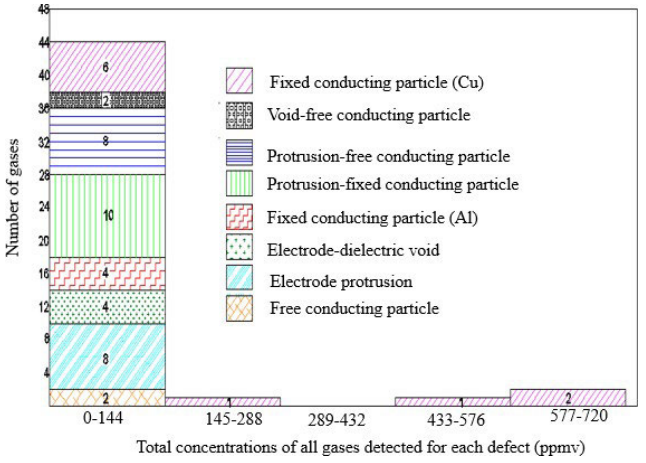

FIGURE 25. Frequency of the appearance of gas concentrations at 10-hour stress duration for all defects.

D. PRE-PROCESS ON THE OVERALL DEFECT

The results of the pattern recognition analyses on all defects studied in this work are reported. There is a total of eight defects from the three categories of defects. The defects are listed here again as: free conducting particle, electrode protrusion, electrode to dielectric void, fixed copper particle, fixed aluminium particle, electrode protrusion-fixed copper particle, electrode protrusion-free conducting particle, and electrode to dielectric void-free conducting particle defects. The results are discussed according to number and identification of by-product gases, and the by-product gas concentration for varying stress durations.

Figure 23 is a plot of the number of by-product gases as a function of the defect type. As can be seen in the figure, the fixed Cu particle from the sole defect category and the electrode protrusion-fixed copper particle from the hybrid category have the highest number of by-product gases (10 gases).

The attribute is based on all gases appearing in the entire experiment (50-hour stress duration) for a given defect. The types of by-product gases produced by the partial discharge activity in all defects are illustrated in Figure 24. The observations from this figure is that, the type of gases can be directly correlated with the type of defect firstly then by the type of fault.

Figures 25 through 29 show the frequency of the appearance of gas concentrations of the SF_6 by-product at 10-hour to 50-hours duration.

The above pattern recognition pre-process analyses reveal that the order of severity of a fault or discharge caused by all defects studied in this study can be listed in the following order (starting with the most severe): the electrode protrusion-fixed copper particle defect, the fixed copper particle defect, the electrode protrusion-free conducting particle defect, the electrode protrusion defect, the fixed aluminium particle defect, the electrode to dielectric void-free conducting particle defect, the electrode to dielectric void defect, and the free conducting particle defects.

VI. CLASSIFICATION OF DEFECTS USING RANDOM FOREST ALGORITHM

Based on the pre-process analyses, the following categories of defects are as follows: sole defect, hybrid defect, material

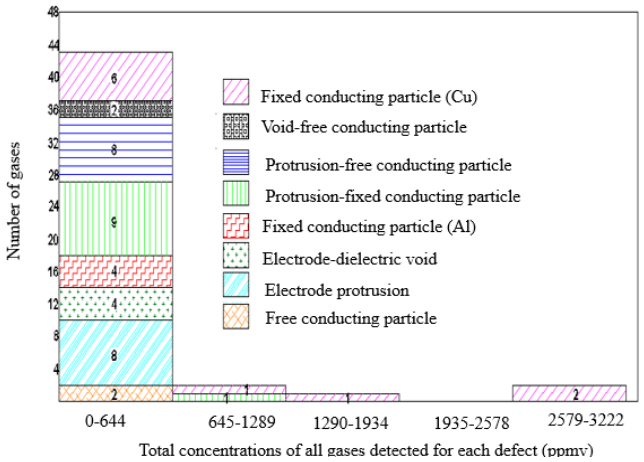

FIGURE 26. Frequency of the appearance of gas concentrations at 20-hour stress duration for all defects.

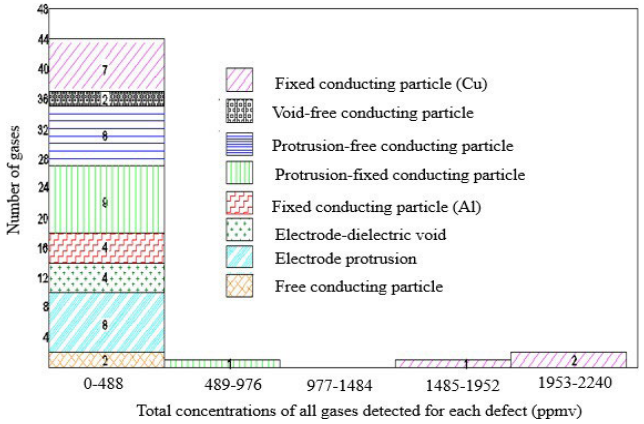

FIGURE 27. Frequency of the appearance of gas concentrations at 30-hour stress duration for all defects.

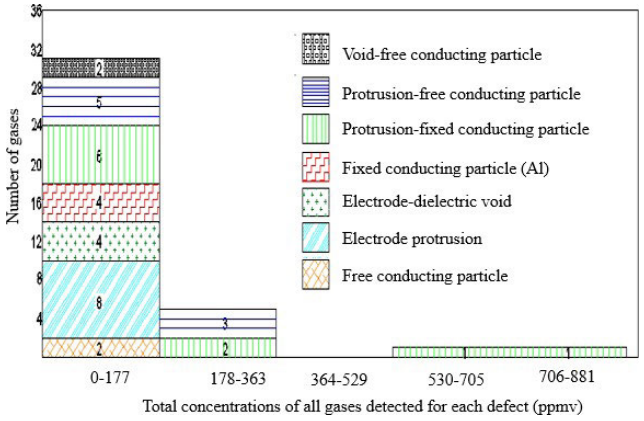

FIGURE 28. Frequency of the appearance of gas concentrations at 40-hour stress duration for all defects.

dependent defect, and all defects. The classification of the four categories of defects was carried out using the random forest algorithm. The validation of the random forest algorithm was carried out using test mode ten-fold cross-validation.

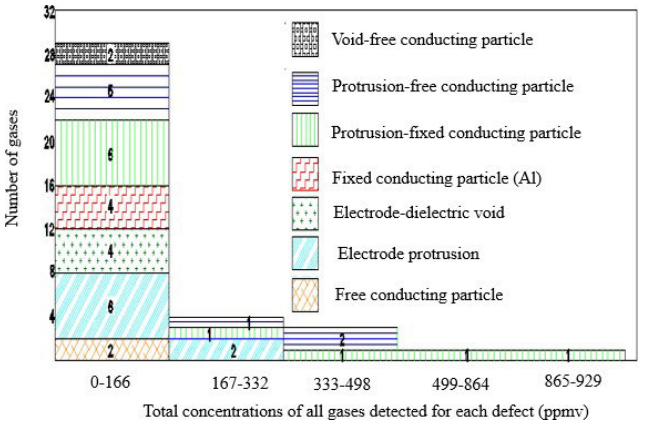

FIGURE 29. Frequency of the appearance of gas concentrations at 50-hour stress duration for all defects.

TABLE 3. Classification accuracy of random forest algorithm for the sole defect.

Class	Free conducting particle	Protrusion	Void	Fixed Cu	Weighted Average
TP	1.000	0.813	1.000	1.000	0.938
FP	0.023	0.000	0.000	0.071	0.032
Precision	0.800	1.000	1.000	0.909	0.945
Recall	1.000	0.813	1.000	1.000	0.938
F–measure	0.889	0.897	1.000	0.952	0.936
MMC	0.884	0.862	1.000	0.919	0.910
ROC area	1.000	0.951	1.000	1.000	0.984
RPC area	1.000	0.944	1.000	1.000	0.981

Note: TP: True Positive, FP: False Positive, MMC: Matthew Correlation Coefficient, ROC: Receiver Operating Characteristic Curve and PRC: Precision-Recall Curve. The percentage accuracy by class is the true positive (TP) rate of that class and the overall classification accuracy of the random forest algorithm is the weighted average.

A. CLASSIFICATION OF SOLE DEFECT

The pre-process results of the sole defect category were further analysed using the random forest algorithm. Table 3 shows the accuracy of the classification. The classification accuracy of the analysis done on the free conducting particle defect is 100%, on the electrode protrusion is 81.3%, on the electrode to dielectric void is 100%, and on the fixed conducting particle is 100%. These give an overall classification accuracy of 93.8%, which is also known as the weighted average of the classes for sole defect. The results show that the applied random forest algorithm results in well performing pattern recognition.

Table 4 shows the random forest confusion matrix for the sole defect classification. This table shows that the random forest algorithm successfully and correctly classified the free conducting particle defect, the void defect and the fixed Cu

TABLE 4. Confusion matrix table for sole defect.

Actual Defect	Classified Defect			
	Free conducting particle	Protrusion	Void	Fixed Cu
Free conducting particle	4	0	0	0
Protrusion	1	13	0	2
Void	0	0	8	0
Fixed Cu	0	0	0	20

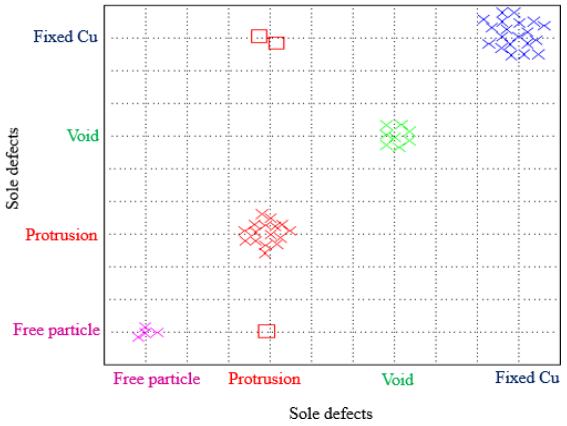

FIGURE 30. Confusion matrix plot for the sole defect.

TABLE 5. Classification result of sole defect.

Classification	Instances	Percentage
Correctly classified	45	93.8%
Incorrectly classified	3	6.3%

particle defect. However, for the electrode protrusion defect, 3 instances of wrong classification occurred (two instances are wrongly classified as fixed Cu particle defect and one instance is wrongly classified as free conducting particle defect).

The graph of classification error is shown in Figure 30. The actual defect is plotted in the x-axis and the classification results in the y-axis. The square boxes in Figure 30 show the three instances of the wrong classification for the protrusion defect as described above. The summary of the classification results is shown in Table 5. The table shows that 45 out of 48 instances were correctly classified. This is equivalent to 93.8%.

B. CLASSIFICATION OF HYBRID DEFECT

Table 6 shows the classification results for the hybrid defect using the forest algorithm. The overall classification accuracy of the random forest algorithm is 80%.

TABLE 6. Classification accuracy of random forest algorithm for the hybrid defect.

Defect	TP rate	FP rate	Precision	Recall	F-measure	MCC	ROC area	PRC area
Protrusion – fixed Cu	0.850	0.200	0.810	0.850	0.829	0.651	0.935	0.940
Protrusion – free conducting particle	0.813	0.167	0.765	0.813	0.788	0.640	0.943	0.906
Void – free conducting particle	0.500	0.000	1.000	0.500	0.667	0.688	0.993	0.917
Weighted Average	0.800	0.167	0.811	0.800	0.796	0.650	0.944	0.924

TABLE 7. Confusion matrix table for hybrid defect.

Actual Defect	Classified Defect		
	Protrusion – fixed Cu	Protrusion – free conducting particle	Void–free conducting particle
Protrusion – fixed Cu	17	3	0
Protrusion – free conducting particle	3	13	0
Void-free conducting particle	1	1	2

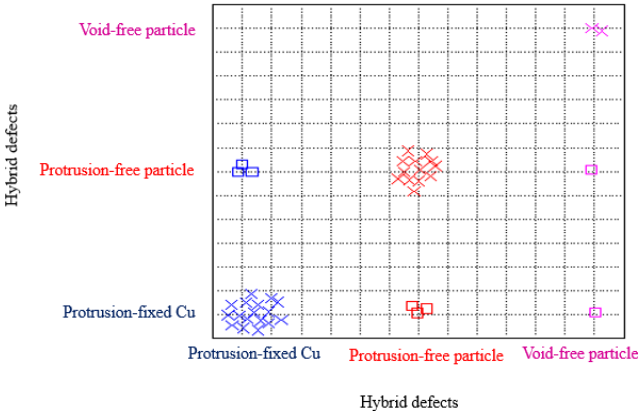

FIGURE 31. Confusion matrix plot for the hybrid defects.

TABLE 8. Classification result of hybrid defect.

Classification	Instances	Percentage
Correctly classified	32	80%
Incorrectly classified	8	20%

The confusion matrix for the classification of the hybrid defect is shown in Table 7. Figure 31 illustrates the corresponding plot for the confusion matrix. Table 8 represents the summary of the classification result. The total percentage of the correctly classified instances is 80%.

TABLE 9. Classification accuracy of random forest algorithm for the material dependent defect.

Class	TP rate	FP rate	Precision	Recall	F-measure	MCC	ROC area	PRC area
Fixed Al	1.000	0.050	0.889	1.000	0.941	0.919	1.000	1.000
Fixed Cu	0.950	0.000	1.000	0.950	0.974	0.919	1.000	1.000
Weighted Average	0.964	0.014	0.968	0.964	0.965	0.919	1.000	1.000

TABLE 10. Confusion matrix table for material dependent defect.

Defect	Classification	
	Fixed aluminium	Fixed copper
Fixed aluminium	8	0
Fixed copper	1	19

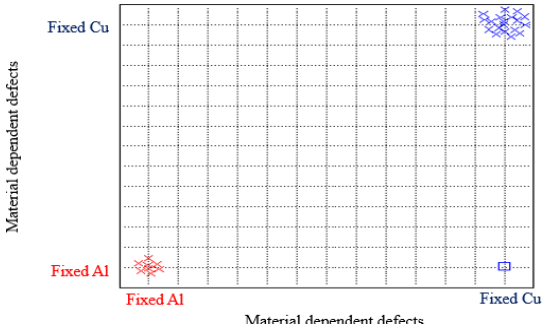

FIGURE 32. Confusion matrix plot for material dependent defect.

TABLE 11. Classification result of material dependent defect.

Classification	Instances	Percentage
Correctly classified	27	96.4%
Incorrectly classified	1	3.6%

C. CLASSIFICATION OF MATERIAL DEPENDENT DEFECT

The detailed classification results of material dependent defects using random forest algorithm is shown in Table 9. The overall accuracy of the model is 96.4% (TP rate), which shows the applied random forest has a very good performance on the recognition of the material dependent defects.

Table 10 shows the confusion matrix of the classification for the material dependent defects. The corresponding graph of the confusion matrix is shown in Figure 32. The summary of the classification results is shown in Table 11. The percentage of the correctly classified instances is 96.4%. Again, it can be concluded that the applied random forest on the recognition of material dependent defects has a very good performance.

D. CLASSIFICATION OF ALL DEFECTS

The classification of all defects using the random forest algorithm is shown in Table 12. The overall classification

TABLE 12. Classification accuracy of random forest algorithm for all defects.

Defect	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Free conductin particle	1.000	0.011	0.800	1.000	0.889	0.890	1.000	1.000
Protrusior	0.875	0.000	1.000	0.875	0.933	0.924	0.994	0.964
Void	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
Fixed Al	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
Protrusion fixed Cu	0.850	0.079	0.739	0.850	0.791	0.734	0.906	0.843
Protrusion free conductin particle	0.688	0.000	1.000	0.688	0.815	0.804	0.790	0.782
Void-free conductin particle	1.000	0.022	0.667	1.000	0.800	0.808	0.997	0.917
Fixed Cu	0.900	0.039	0.857	0.900	0.878	0.845	0.966	0.923
Weighted Average	0.875	0.026	0.894	0.875	0.876	0.854	0.937	0.906

TABLE 13. Confusion matrix table for all defects.

Defect	Classification							
	FCP	P	V	FAl	PFCu	PFCP	VFCP	FCu
FCP	4	0	0	0	0	0	0	0
P	1	14	0	0	1	0	0	0
V	0	0	8	0	0	0	0	0
FAl	0	0	0	8	0	0	0	0
PFCu	0	0	0	0	17	0	2	1
PFCP	0	0	0	0	3	11	0	2
VFCP	0	0	0	0	0	0	4	0
FCu	0	0	0	0	2	0	0	18

FCP = Free Conduction Particle, P = Protrusion, V = Void, FAl = Fixed Al, PFCu = Protrusion Fixed Cu, PFCP = Protrusion Free Conducting

TABLE 14. Classification result of all defects.

Classification	Instances	Percentage
Correctly classified	84	87.5%
Incorrectly classified	12	12.5%

accuracy (87.5%) shows that the applied random forest performs very well at defects recognition.

Table 13 shows the random forest confusion matrix for all defects. The corresponding graph of the confusion matrix is shown in Figure 33. The summary of the classification results is shown in Table 14. The total percentage of the correctly classified instances is 87.5%. Therefore, we can conclude that the applied random forest has a very good performance when applied for the recognition of all defects.

VII. COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER ALGORITHMS

Table 15 shows a comparison for the accuracy score of eight classification algorithms when solo, hybrid, material and all of the defects have applied for classification. The tabulated accuracy results show the remarkable superiority

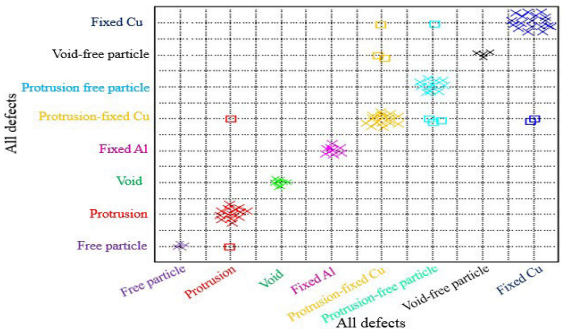

FIGURE 33. Confusion matrix plot for all defects.

TABLE 15. Comparison of the classification accuracy of various algorithms.

Algorithm	Classification Accuracy (%)			
	Sole defect	Hybrid defect	Material dependent	All defects
Random forest	93.8	80.0	96.4	87.5
Decision tree	75.0	50.0	-	22.9
Decision stump	68.8	42.5	-	50.0
Decision table	81.3	42.5	-	58.3
The part	68.8	45.0	92.9	27.1
One rule	83.3	52.5	92.9	48.9
Multilayer perceptron	83.3	67.5	85.7	51.0

of Random Forest to identify the type of defect in all categories. To ensure the comparability of classification results, all algorithms were carried out using the Weka tool. The most important to consider in this table is the accuracy score when all defects have classified, supposing that categories of the defects are unknown, besides when the classes number are diverse, the classification will be more complicated and multiclass classification algorithm will be required.

An inspection for the results, by taking all defect category as a reference to compare the reduction and increasing of accuracy score. It can be concluded that Random Forest accuracy percent has less variation when compared to others. For instance, the reduction difference between solo defect and all defect was 6.3 percent from 93.8 to 87.5 which is small when compared to other algorithms, taking into consideration the very high accuracy score obtained by Random Forest. In the same manner, there was an increase to 7.5 percent between hybrid and all defect categories.

Furthermore, the algorithms used are diverse as tree-based such as random forest, decision tree and decision stump, and gradient-based such as multilayer perceptron and rule-based algorithm such as one rule. This diversity provides a comparison for pattern recognition algorithms based on how decision are made. For some algorithms, data are required to be normalized or standardized, such as for distance-based or gradient-based algorithms, which can highly influence the results. However, for Random Forest, there is no scaling required owing to the decision are made based on node split. Hence, reducing the steps of data preprocessing.

Moreover, Random Forest did not only show a higher accuracy, but it also less prone to overfitting that may happen in a decision tree, and able to predict new input data as large number of trees are used to construct the algorithm. In this article, the generated products have used as feature parameters without considering specific gases as characteristics quantities. However, Random Forest will use features that have more information, providing dimensionality reduction feature during classification. Generally, as performance metrics in this article, the accuracy score has shown the suitability of Random Forest to classify a variety of defect categories.

Other factors should also be considered when comparing the performance of the classification algorithm, such as memory required, computational time, as well as the ability to be used for a multi-class classification problem. However, these can vary depending on the amount of dataset of the problem applied. Also, more performance metrics can be utilized such as sensitivity, precision and ROC curve to provide better evaluation for the classification model.

VIII. CONCLUSION

The use of FTIR for gas analysis has enabled more by-product gases to be detected. A total of twelve gas by-products due to partial discharge activity in all defects were detected. Arranged according to significance, these are hexafluoroethane (C_2F_6), sulphur dioxide (SO_2), sulfuryl fluoride (SO_2F_2), octafluoropropane (C_3F_8), silicon tetrafluoride (SiF_4), thionyl fluoride (SOF_2), carbon monoxide (CO), disulfur decafluoride (S_2F_{10}), hydrogen fluoride (HF), tetrafluoromethane (CF_4), carbonyl sulphide (COS) and tetrafluoride (SOF_4). More gases detected help in detail data for accurate faults classification, better diagnosis on GIS condition. Indeed, by detecting more by product gases the safety operation of the GIS can be ensure as the presence of CO, COS, SiF_4 and HF gases can be harmful to the GIS system due to their flammable and corrosive nature.

The defect classification using 12-by product gases recognition pattern have found that the type, number, concentration, and chemical stability of by-product gases are closely correlated to the type of defect. Generally, the number and concentration of the by-product gases increases with electrical stress duration and the presence of the by-product gas and its concentration can be said to be an indication of a fault in GIS and the fault is harmful to the GIS. Further analyses using pattern recognition with eight algorithms based on the presence and concentration of the gas by-products were carried out. The random forest algorithm based on these 12 by-product gases had successfully recognises a given defect with an accuracy of 87.5% for all defects algorithm which is far better than others method as highlighted in Table 15 which is 1.5 times better than the decision table algorithm (the 2nd highest percentage of accuracy). This research illustrates the feasibility and applicability of an effective GIS diagnostic using gas by-products analyses, in particular, using the random forest pattern recognition.

Furthermore, from the analyses using Waikato Environment for Knowledge Analysis (WEKA) workbench machine learning and data mining, in particular, the random forest algorithm of pattern recognition, the defect classification of sole, hybrid, and material dependent were successfully obtained with classification accuracies of 93.8%, 80%, and 96.4%, respectively. Therefore, the random forest algorithm can be applied as a very good tool for pattern recognition and prediction of multi-fault in a gas insulated system.

ACKNOWLEDGMENT

The authors cordially thank all TNB Research Sdn. Bhd. staff involved in this research.

REFERENCES

[1] I. A. Metwally, "Status review on partial discharge measurement techniques in gas-insulated switchgear/lines," *Electr. Power Syst. Res.*, vol. 69, no. 1, pp. 25–36, Apr. 2004.

[2] J. S. Pearson, O. Farish, B. F. Hampton, M. D. Judd, D. Templeton, B. W. Pryor, and I. M. Welch, "Partial discharge diagnostics for gas insulated substations," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 2, no. 5, pp. 893–905, Oct. 1995.

[3] M. S. Naidu and V. Kamaraju, *High Voltage Engineering*, 3rd ed. New Delhi, India: Tata McGraw-Hill Professional Publishing Company Limited, 1996, pp. 1–9.

[4] W. Ding, R. Hayashi, K. Ochi, J. Suehiro, K. Imasaka, M. Hara, N. Sano, E. Nagao, and T. Minagawa, "Analysis of PD-generated SF_6 decomposition gases adsorbed on carbon nanotubes," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 13, no. 6, pp. 1200–1207, Dec. 2006.

[5] R. Baumgartner, B. Fruth, W. Lanz, and K. Pettersson, "Partial discharge. X. PD in gas-insulated substations-measurement and practical considerations," *IEEE Elect. Insul. Mag.*, vol. 8, no. 1, pp. 16–27, Jan./Feb. 1992.

[6] L. E. Lundgaard, K. Ljokelsoy, W. Hansen, A. Schei, and L. Hofstad, "Acoustic insulation analyzer for periodic condition monitoring of insulation systems such as GIS, cable terminations and joints ETS," TransiNor As, Trondheim, Norway, Tech. Rep., 1997.

[7] V. Aaradhi and K. Gaidhani, "Special problems in gas insulated substations (GIS) and their effects on Indian power system," in *Proc. IEEE Int. Conf. Power Syst. Technol. (POWERCON)*, Oct. 2012, pp. 1–5.

[8] O. Soppart, J. I. Baumbach, S. M. Alberti, and D. Klockow, "On-site quality assessment of SF_6 using ion mobility spectrometry," in *Proc. Conf. 10th Int. Symp High Voltage Eng.*, Montreal, QC, Canada, vol. 4, 1997, pp. 147–150.

[9] G. C. Stone, "Partial discharge diagnostics and electrical equipment insulation condition assessment," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 12, no. 5, pp. 891–904, Oct. 2005.

[10] S. A. Boggs, "Partial discharge. III. Cavity-induced PD in solid dielectrics," *IEEE Elect. Insul. Mag.*, vol. 6, no. 6, pp. 11–16, Nov. 1990.

[11] M. Ren, M. Dong, and J. Liu, "Statistical analysis of partial discharges in SF6 gas via optical detection in various spectral ranges," *Energies*, vol. 9, no. 3, p. 152, Mar. 2016.

[12] T. Hattori, M. Honda, H. Aoyagi, N. Kobayashi, and K. Terasaka, "A study on effects of conducting particles in SF_6 gas and test methods for GIS," *IEEE Trans. Power Del.*, vol. 3, no. 1, pp. 197–204, 1988.

[13] T. Yamigiwa, T. Ishikawa, and F. Endo, "Particle-initiated breakdown characteristics on a ribbed spacer surface for SF_6 gas insulated switchgear," *IEEE Trans. Power Del.*, vol. 3, no. 3, pp. 954–960, Jul. 1988.

[14] *Knowledge Rules for Partial Discharge Diagnosis in Service. Electra*, Standard CIGRE, T. 15.11/33.03. 02, 2003, pp. 63–66.

[15] *Partial Discharge, for 24/7 Gas Insulated Switchgear Monitoring System*, EA Technol. Innovators Power Eng., Chester, U.K., 2008, pp. 1–8.

[16] J. Tang, F. Liu, X. Zhang, Q. Meng, and J. Zhou, "Partial discharge recognition through an analysis of SF_6 decomposition products Part 1: Decomposition characteristics of SF_6 under four different partial discharges," *IEEE Trans. Dielectr. Elect. Insul.*, vol. 19, no. 1, pp. 29–36, Feb. 2012.

[17] J. Tang, F. Liu, Q. Meng, X. Zhang, and J. Tao, "Partial discharge recognition through an analysis of SF_6 decomposition products part 2: Feature extraction and decision tree-based pattern recognition," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 19, no. 1, pp. 37–44, Feb. 2012.

[18] J. Tang, F. Liu, X. Zhang, X. Liang, and Q. Fan, "Partial discharge recognition based on SF_6 decomposition products and support vector machine," *IET Sci., Meas. Technol.*, vol. 6, no. 4, pp. 198–204, 2012.

[19] J. Tang, X. Yang, G. Ye, Q. Yao, Y. Miao, and F. Zeng, "Decomposition characteristics of SF_6 and partial discharge recognition under negative DC conditions," *Energies*, vol. 10, no. 4, p. 556, Apr. 2017.

[20] U. S. Environmental Protection Agency EPA). (Jan. 2002). *Byproducts of Sulfur Hexafluoride (SF_6) Use in the Electric Power Industry*. Accessed: Feb. 15, 2016. [Online]. Available: https://www.epa.gov/sites/production/files

[21] Y. Wang, L. Li, and W. Yao, "SF_6 by-products in high-humidity environment: An experimental evaluation between 200°C and 500°C," *J. Electromagn. Anal. Appl.*, vol. 3, no. 6, p. 179, 2011.

[22] C. T. Dervos and P. Vassiliou, "Sulfur hexafluoride (SF_6): Global environmental effects and toxic byproduct formation," *J. Air Waste Manage. Assoc.*, vol. 50, no. 1, pp. 137–141, Jan. 2000.

[23] A. Kreider, "US EPA's SF_6 emissions reduction partnership for electric power systems: Results and prospects," in *Gaseous Dielectrics IX*. Cham, Switzerland: Springer, 2001, pp. 593–596.

[24] L. Hao and P. Lewin, "Partial discharge source discrimination using a support vector machine," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 17, no. 1, pp. 189–197, Feb. 2010.

[25] D. Evagorou, P. L. Lewin, V. Efthymiou, A. Kyprianou, G. E. Georghiou, A. Stavrou, and A. C. Metaxas, "Feature extraction of partial discharge signals using the wavelet packet transform and classification with a probabilistic neural network," *IET Sci., Meas. Technol.*, vol. 4, no. 3, pp. 177–192, May 2010.

[26] F. Zeng, S. Wu, X. Yang, Z. Wan, J. Tang, M. Zhang, and Q. Yao, "Fault diagnosis and condition division criterion of DC gas insulating equipment based on SF6 partial discharge decomposition characteristics," *IEEE Access*, vol. 7, pp. 29869–29881, 2019.

[27] L. Zhong, S. Ji, K. Liu, Q. Xiong, and L. Zhu, "Decomposition characteristics of SF_6 under three typical defects and the diagnostic application of triangle method," *IEEE Trans. Dielectr. Elect. Insul.*, vol. 23, no. 5, pp. 2594–2606, Oct. 2016.

[28] H. Hirose, M. Hikita, S. Ohtsuka, S.-I. Tsuru, and J. Ichimaru, "Diagnosis of electric power apparatus using the decision tree method," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 15, no. 5, pp. 1252–1260, Oct. 2008.

[29] N. J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed. San Mateo, CA, USA: Morgan Kaufmann, 2012, pp. 327–392.

[30] E. Frank, "Machine learning with WEKA," Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, Tech. Rep. 97/6, 1999.

[31] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann Publishers. 2016.

[32] L. Rokach and O. Maimon, *Data Mining With Decision Trees: Theory and Applications*, 2nd ed. Singapore: World Science publishing Co. Pt. Ltd, 2014, pp. 31–165.

[33] G. Kaur and A. Chhabra, "Improved J48 classification algorithm for the prediction of diabetes," *Int. J. Comput. Appl.*, vol. 98, no. 22, pp. 13–17, Jul. 2014.

[34] H. Ian and F. Eibe, *Data Mining, Practical Machine Learning Tools and Techniques*, 2nd ed. San Mateo, CA, USA: Morgan Kaufmann, 2005.

NOR ASIAH MUHAMAD (Member, IEEE) received the bachelor's degree in electrical and electronic engineering from the Universiti Teknologi Petronas, Malaysia, in 2002, the master's degree in electrical power engineering from the University of South Australia, in 2006, and the Ph.D. degree from the University of New South Wales, Australia, in 2009. She was a Researcher and a Senior Lecturer with the Faculty of Electrical Engineering, Institute of High Voltage and High Current (IVAT), Universiti Teknologi Malaysia. She is currently serving as a Senior Lecturer with the School of Electrical and Electronic Engineering, Universiti Sains Malaysia. Her research interest includes power system equipment monitoring, in particular, insulation diagnosis and the development of new systems for condition monitoring. Besides doing research in HV area, she also works in energy efficiency and integration for domestic and industrial area in Malaysia.

IBRAHIM VISA MUSA received the Bachelor of Engineering (B.Eng.) degree in electrical and electronics from the University of Maiduguri Nigeria, in 1996, the M.Eng. degree in electrical power system engineering from the University of Benin Nigeria, in 2005, and the Ph.D. degree in high voltage engineering from Universiti Teknologi Malaysia (UTM), in 2018. He is currently a Senior Lecturer, a Researcher, and the Head of the Department of Electrical and Electronics Engineering, Modibbo Adama University of Technology, Yola, Nigeria, formerly known as the Federal University of Technology. His research interests include condition monitoring of power system equipment and diagnosis using artificial intelligent. Apart from carrying out research in high voltage engineering, he is also working as a Consultant Services Designer of domestic and industrial electrification in Nigeria.

ZULKURNAIN ABDUL MALEK (Senior Member, IEEE) received the B.E. degree from Monash University, Melbourne, Australia, in 1989, and the M.Sc. degree in electrical and electromagnetic engineering with industrial applications and the Ph.D. degree in high voltage engineering from Cardiff University, in 1995 and 1999, respectively. He has been with Universiti Teknologi Malaysia, for 31 years, where he is currently a Professor and the Director of the Institute of High Voltage and High Current. He has published more than 150 papers in various technical journals and conference proceedings. His research interests include surge arrester intelligent condition monitoring, lightning related EMC, warning systems, lightning protection systems, electrical discharge, condition monitoring, and nanodielectrics. He is actively involved in many committees.

AMMAR SALAH MAHDI received the bachelor's degree in electrical engineering from the University of Basra, Iraq, in 2009, and the master's degree in electrical engineering from the Universiti Tun Hussein Onn Malaysia (UTHM), in 2019. He is currently pursuing the Ph.D. degree with the Institute of High Voltage and High Current (IVAT), Universiti Technology Malaysia (UTM). His research interest includes pattern recognition for GIS status under partial discharge fault.