New Dissolved Gas Analysis Diagnostics Framework for Substation Transformers using Random Forest Algorithm and IEEE C57.104 – 2019TM Guide

Jestoni P. Tan and Wilen Melsedec O. Narvios*
Department of Electrical Engineering
Cebu Technological University
Cebu City, 6014 Philippines
wilenmelsedec.narvios@ctu.edu.ph

Date received: March 14, 2023
Revision accepted: May 30, 2025

Abstract

Dissolved Gas Analysis (DGA) is a practical, non-intrusive test to check transformer health status, as it is widely used in the field. However, the traditional methods of DGA-based diagnostics have intrinsic weaknesses. For example, the Rogers ratio method is limited only to gases involved in the computation. The interpretation of the IEC Ratio method can be unknown at some point. The Duval triangle method cannot diagnose healthy degradation of oil from faulty ones. All traditional methods were subject to expert subjective judgment. To fill these gaps, this paper introduces the two-layer framework using a random forest algorithm with the IEEE C57.104 – 2019TM guide as a watchdog (layer 1) for unhealthy oil degradation versus normal ones. The prediction model (layer 2) used the random forest algorithm. Using the 277 DGA datasets from Distribution Utilities from different parts of the Philippines, the framework surpassed the accuracy of traditional methods (Duval triangle method, IEC ratio, Doernunberg method) with an accuracy of 100%. The Duval triangle got 98.92% accuracy, the IEC ratio had 28.32% accuracy, and the Doernunberg method had an accuracy of 27.50%. Other ML algorithms, such as ANN (MLP), K-nearest neighbors, SVM (linear), and J48, were also used for comparison. The ANN (MLP), K-Nearest neighbor, and SVM (linear) got 78.6%, 85.7%, and 78.6% accuracy, respectively. The random forest got the highest cross-validation score (89.14% ave.) among all ML methods. Further evaluations were used for J48, DT, and Random Forest since all got 100% accuracy. RF algorithm still got the highest PR-AUC (94%, 89%) and ROC-AUC (95%, 97%) scores among the J48 and DT in the 70/30 and 80/20 data split.

***Keywords**:* dissolved gas analysis, machine learning, random forest, supervised learning, transformer health

1. Introduction

1.1 Condition Assessment of Transformers

Condition Assessment procedures (CA) are vital tools to evaluate the physical condition of equipment to make informed maintenance decisions and in asset management. Condition assessment consists of a systematic inspection, review, and report of the state of the equipment (de Castro-Cros *et al*., 2021). The goal of CA in the context of power transformers was to maximize the overall lifecycle of an aging asset while minimizing possible operational costs. CA procedures are subject to expert opinion. The problem with expert opinion is that it differs from one expert to another, and manual assessment is time-consuming (Sholevar *et al*., 2022). This leads to inconsistency on the receiving end of the interpretation of the data. Power assets were one of the most complex assets to manage and protect. Power transformers are indispensable assets in the power system. Asset managers had two tasks: to know the underlying incipient fault to protect the transformers and to estimate their remnant life (Aminifar *et al*., 2022). These two tasks are significant challenges themselves since transformers are composed of complex systems (L. Sun *et al*., 2016). In effect, various tests and surveys were needed, and factors to consider, such as loading history, maintenance record, environmental condition, and fault history, to do condition assessment. Early and accurate diagnosis is critical for these machines to prevent unexpected breakdowns that can cause catastrophic damage. Moreover, replacing damaged transformers is costly and time-consuming. However, with the advancement in machine learning, automated and unbiased decision-making processes can significantly enhance the accuracy of diagnosis.

1.2 Health Index

One key aspect of machine health assessment is the computation of a health index, which quantifies the machine's overall condition based on various parameters and sensor readings. This index provides a standardized metric for comparing the health of different machines and tracking changes over time. The HI ranges from 0% to 100%. The lesser the transformer's HI, say <40%, the more likely the transformer was in good condition, and 100% indicates the transformer is in "poor" condition. Although the HI approach cannot reflect the status of any specific component of a transformer, it measures the level of overall long-term deterioration (Murugan and Ramasamy, 2019). Abu-Elanien *et al*. (2011) coined the word “health index” (HI) as a general transformer health indicator using a feed-forward artificial neural network

(FFANN), which proved successful for all transformers. The index approach was based on the lab test results, on-site physical examination, and historical fault summary of the transformers' operation. The index approach harmonized the results of these tests with qualitative physical surveys to have an overall transformer evaluation (Aizpurua *et al*., 2019; Azmi *et al*., 2017). Machine learning approaches such as evidence theory, gray clustering decision, matter element theory, and Bayesian network normalized by the fuzzy logic algorithm were employed in the HI approach (Da Silva *et al*., 2021). Guo, H. and Guo, L. (2022) developed a practical HI index approach by averaging two HI values: (HI1 and HI2) for the condition assessment applied to 55 power transformers at 330 kV. This considers not only the traditional data, such as lab tests and fault history, but also the environmental operating conditions of the transformer. The factor of the age of the transformer was also taken into account. HI approach of ranking and weights was effective in minimizing maintenance costs in the field. However, in the context of the proactive HI approach, which was based on data and expert opinion, the computation might vary from one expert to another. To overcome this limitation, intelligent algorithms such as fuzzy logic, ANN, and SVM were developed to eliminate expert decisions for a more consistent index. Though it does eliminate subjective evaluation, it does not change the fact that the source data is still inconsistent in each research advancement. Some transformer test data were included in one study, and some were not. Moreover, other tests are costly, and the method is rigorous (Wong *et al*., 2022). As many machine learning (ML) algorithms and varied datasets aided the HI approach, which gave more power and vigor to the method, it also became more complex and inconsistent (Murugan and Ramasamy, 2019). In effect, irregularities in the Health Index approach existed. Furthermore, the index approach, as a general “health” indicator, could not be specific about what fault the transformer was at risk of. Thus, engineers and field practitioners could not implement maintenance procedures to mitigate the real and specific “health” issue of aging transformers.

1.3 Dissolved Gas Analysis

Dissolved Gas Analysis (DGA) monitors the dissolved gases in the transformer oil. The purpose of the DGA test is to evaluate the incipient faults. Incipient faults include electrical, thermal, partial discharge, and stray gassing (Faiz and Soleimani, 2017). The traditional methods of DGA interpretation are the Key Gas method, Rogers’ ratio method, Dornenburg ratio method, IEC method, Nomograph method, IEEE method, and the Duval triangle method (C. Sun *et al*., 2017). Even though these are established methods in the field

of DGA, they have inherent limitations. In the case of the Key Gas method, which is prone to misdiagnosis, many experts do not use this method (Anil and Archana, 2017). The Dornenburg ratio method may have undefined results when the gases are outside the threshold limits of gas. Roger's ratio method uses only three gas ratios (CH_4/H_2, C_2H_2/C_2H_4, and C_2H_4/C_2H_6), and its limitation is that it cannot diagnose other gas combinations. The IEC method also cannot have a diagnosis if it does not fit the listed code classification (Wani *et al*., 2021). The Duval triangle method (DTM) is a graphical triangular method using three gas percentages (CH_4, C_2H_2, and C_2H_4) to indicate six faults. One of the setbacks of DTM is that it does not identify the normal state of the transformer.

1.4 Dissolved Gas Analysis (DGA) Condition Assessment

Aside from the HI approach, other studies also used a popular and reliable condition assessment for transformers. The test was called the Dissolved Gas Analysis (DGA). This test was used to monitor the dissolved gases in the transformer oil. DGA has been recognized for over 50 years for improving reliability and lowering transformer asset maintenance costs. DGA test differed from other routine tests as it could be done more than once a year or daily, depending on the necessity. Oil radicals such as hydrogen (H_2), methane (CH_4), ethane (C_2H_6), ethylene (C_2H_4), acetylene (C_2H_2), carbon monoxide (CO), carbon dioxide (CO_2), oxygen (O_2) and Nitrogen (N_2) in the DGA test were used to evaluate the incipient faults in the transformer. Graphically presented in Figure 1 is the fault gas generation chart by the United States Department of the Interior Bureau of Reclamation (Liu and Bao, 2022). Incipient faults include electrical, thermal, partial discharge, and stray gassing (Faiz and Soleimani, 2017). This was based on the fact that the presence of gas radicals correlates with the presence of mechanical, electrical, and thermal faults in the transformer. Unlike the HI approach, there was no need to include other oil and component tests. Therefore, consistency of data was achieved. Consequently, it is a more practical approach than the index. The DGA data was more accessible on the premise that DGA is a common test among electric utilities. Furthermore, DGA was fault-specific, and engineers could do prompt mitigations to that fault diagnosis. However, DGA is not a pure science and can have errors in the oil sampling and logistics (ASTM D3613), especially in remote areas. Nevertheless, it is logical to say, based on the literature, that it is still a well-known method for CA in transformers. The DGA-based transformer health approach has been proven effective for decades in the power engineering field as it became part of the IEC 60599 and IEEE C57.104 standard. The established traditional methods of DGA interpretation were the

Key Gas Method, Rogers Ratio Method, Dornenburg Ratio Method, IEC Method, Nomograph Method, IEEE Method, and the Duval Triangle Method. These methods had inherent limitations. In the case of the Key Gas Method, which was prone to misdiagnosis, many experts did not recommend the method (Anil and Archana, 2017). The Dornenburg Ratio method may have undefined results in some cases. The Rogers Ratio method uses only three gas ratios (CH_4/H2, C_2H_2/C_2H_4, and C_2H_4/C_2H_6), and its limitation is that it cannot diagnose other gas combinations. IEC Method also had the limitation to have no diagnosis if it doesn't fit the listed code classification (Wani *et al*., 2021). The IEEE Method used the Total Dissolved Concentration Gas (TDCG) formula and R (an increase of TDCG value in millimeters/day) to evaluate the condition of the transformer. The latter method could not be used without a series of TDCG datasets. The Duval Triangle Method (DTM) was a graphical triangular method using three gas percentages (CH_4, C_2H_2, and C_2H_4) to indicate six faults. The graph of the DTM is shown in Figure 2. One of the setbacks of DTM was that it could not identify the normal state of oil degradation of the transformer. A nomograph used a series of logarithmic scales of individual hydrocarbon gases based on a certain model to diagnose faults based on gas ratios.

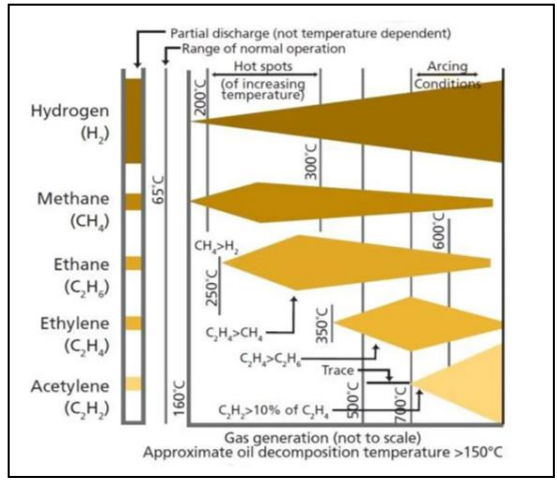


Figure 1. DGA Fault Gas Generation Chart (Temple and Duncan, 1989)

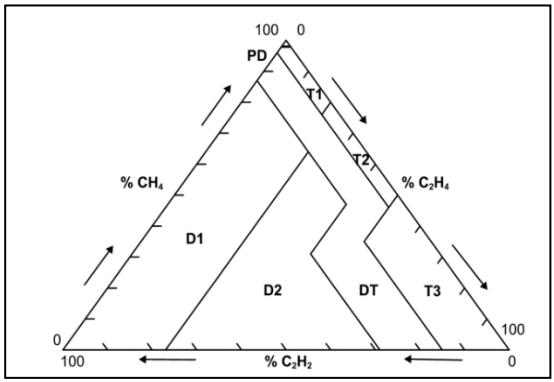


Figure 2. The Duval triangle method graph with fault codes (IEEE, 2019)

1.5 Machine Learning in Dissolved Gas Analysis Approach

As machine learning methods were employed in the HI approach to improve its performance, DGA interpretation was no exception. The machine learning methods aided in overcoming the limitations of traditional methods of DGA evaluation and improved their performance. For example, the Fuzzy logic model, pattern recognition, and Extension Theory aided the limitations of the IEC Method by overcoming uncertainties in the method. The extended version of the Duval Triangle Method (DTM) using Fuzzy Logic provided multiple fault differentiation that the conventional counterpart cannot do (Wani *et al*., 2019). Though Fuzzy logic is a good method for improving conventional DGA methods, its parameters should be tuned correctly to avoid discrepancies in the output. The capabilities of Artificial Neural Networks (ANN) in solving uncertainties, ANN extended the limitations of conventional methods. ANN algorithm fused with ratio methods improved performance compared to their original versions. But amidst the computing power of ANN models, they were constrained by many factors, like training time and network structure (computational cost). They were also solely dependent on the quality of data. Therefore, compromising data quality using ANN leads to an incorrect diagnosis. Aside from ANN and Fuzzy logic, other intelligent methods such as Adaptive Neuro-Fuzzy Inference System (ANFIS) and Support Vector Machine Technique (SVM) were also employed. Malik and Mishra (2016) proposed Gene Expression Programming (GEP) using three ratios of the IEC method as inputs. Same to ANN and Fuzzy, the three methods (ANFIS, SVM, and GEP) have limitations. They were limited by adequate data samples (ANFIS), kernel function identification (SVM), and fitness function selection (GEP) in practical settings. Grey Clustering Analysis (GCA) and Deep Belief

Network were also used for diagnosis. Ibrahim *et al*. (2018) also developed the DGA Lab software to compare traditional methods with their AI-enhanced equivalents, helping save time when assessing AI performance methods. Correlative analysis of these AI methods showed there was no best method among them and should be considered complementary to one another (Faiz and Soleimani, 2018). Since DGA Interpretation was a complicated task, many studies employed hybrid intelligent algorithms such as using fuzzy logic, SVM, Wavelet Networks (WN), Artificial Neural Network (ANN) models coupled with evolutionary programming (EP), Particle Swarm Optimization (PSO), and Genetic Algorithm (GA) (Senoussaoui *et al*., 2018). Comparing hybrid and non-hybrid intelligent classifiers, the hybrid versions gave a better diagnosis. The setback of these systems was the complexity of parameter tuning, which limited their practical application. Some studies used the combination of one or more traditional methods aided by ML methods (Ibrahim *et al*., 2018; Li *et al*., 2018; Wani *et al*., 2019). However, these techniques did not add to the knowledge about the fault information.

1.6 Random Forest (RF) Algorithm in Dissolved Gas Analysis Method

The random forest (RF) algorithm was an ensemble machine learning algorithm proposed by Breiman in 2001. This was an improvement of his boosting ensemble method in 1994. Ensemble means it was a combination of different decision models (known as trees) to perform as a whole. It was on the premise that weak model learners are joined to be strong learners. The Random Forest classifier used bagging or bootstrap aggregating (getting subsets of training samples through replacement) to form an ensemble of classification trees. In effect, the same sample could be selected several times, while others might not be selected at all (Belgiu and Drăguţ, 2016). The final classification decision was taken by averaging (using the arithmetic mean) the class assignment probabilities calculated by all produced trees. A new unlabeled data input was thus evaluated against all decision trees created in the ensemble, and each tree votes for class membership. The membership class with the maximum votes would be the one that is finally selected. RF algorithms were better ML methods than decision tree classifiers. RF was robust and insensitive to overfitting. Though RF was sensitive to data sampling and a little bit slower in computation, it was nevertheless a reliable and highly accurate ML classifier. The RF algorithm is usually exploited in remote data sensing tasks where it performs best, as in hyperspectral data classification and land cover (LC) classification of Enhanced Thematic Mapper (ETM+) or Multispectral Scanner (MSS) and Digital Elevation Model (DEM) data (Zhao *et al*., 2023). RF algorithm was also applied to statistical

tasks such as quantile estimation, causal inference, and survival analysis for coronary artery disease, and it proved to be very competitive (Wager and Athey, 2018). Many algorithms for DGA interpretation were employed using a total of 4580 DGA samples and the IEC TC 10 database, and proved that ensemble classifiers such as Random Forest are better with DGA fault prediction (Rao *et al*., 2021). Ekojono *et al*. (2022) investigated algorithms such as decision tree, support vector machine, random forest (RF), neural network, Naïve Bayes, and AdaBoost that could best aid the Duval Triangle Method for DGA interpretation. The RF Classifier performed best among the mentioned algorithms based on classification accuracy, the area under the curve, F1, Precision, and Recall. Kumar and Haque (2022) used an RF classifier for a modified Duval pentagon method using the density-based clustering (DBSCAN) approach, which proved to have a high classification accuracy. Dai *et al*. (2017) used kernel principal component analysis KPCA and a random forest RF Classifier to diagnose faults of traction transformers using the DGA dataset, achieving 100% accuracy. Jamshed *et al*. (2021) used the RF Classifier with the ten gas ratio combinations of hydrogen, ethane, methane, acetylene, and ethylene (DGA gases) for fault diagnosis. The paper achieved 89% accuracy and was effective in detecting Partial Discharge (PD)

This study developed a framework based on DGA data using the Random Forest Algorithm and the IEEE C57.104 – 2019TM Guide to improve the diagnostic method of DGA interpretation. The proposed framework utilized the primary data collected from the Distribution Utilities of Cebu and other contractors. The framework will eliminate the subjectivity of expert opinion. The framework's performance was compared to established traditional and machine learning methods.

2. Methodology

The study encompasses the evaluation of transformer health through Dissolved Gas Analysis solely and does not include other parameters for transformer health evaluation such as loading history, operational conditions, and other routine tests dataset. Only five input gases (as features) were included in the study namely methane (CH_4,) ethane (C_2H_6), acetylene (C_2H_2), ethylene (C_2H_4) and hydrogen (H_2) were the features of the dataset. These five gases were common features of DGA traditional methods interpretation. The fault types (the labels of the dataset) were based on the IEC 60599 standard,

which are partial discharge (PD), low energy discharge (D1), high energy discharge (D2), and thermal faults 1, 2, and 3 (T1, T2, T3). The study only uses 227 DGA datasets with an uneven number of fault output data. Moreover, only mineral oil-filled transformers will be studied in this paper. Lastly, stray gassing in the transformer oil is not considered in the DGA interpretation. The voltage level for the primary DGA data was not uniform, as it is taken from different parts of Cebu.

2.1 The Two Layer DGA diagnostic framework

As shown in Figure 3, the proposed two-layer DGA diagnostic framework will consist of an IEEE C57.104 Guide and a random forest (RF) algorithm prediction model. The former will serve as the "fault sensor or watchdog" to detect normal from the faulty transformer. The latter will diagnose the type of incipient fault present in the transformer. DGA gases in ppm values with labeled faults, specifically hydrogen, methane, ethane, ethylene, and acetylene, will be fed to the "fault sensor" layer. The threshold of fault gases, the IEEE thresholds, will be the barometer of whether the said DGA dataset represents a faulty transformer. If the dataset is within the limits of the IEEE Guide, the DGA dataset will no longer proceed to layer 2. However, if it exceeds the limit, it will proceed to the second layer, the "classification layer." The prediction model of the said layer will process the faulty dataset. Then the prediction model will classify the specific fault based on its learned insight using a random forest algorithm.

As shown in Figure 3, the developed two-layer DGA diagnostic framework consisted of an IEEE C57.104 Guide and a random forest (RF) algorithm prediction model. The former served as the "fault sensor or watchdog" to detect "high risk" from "low risk" transformers. If there were a minimum of two key gases that surpass the limit set by the IEEE Guide, it would be considered a "high risk" transformer dataset. All "low risk" datasets would be for recordkeeping and subject to another DGA testing after six months. All "high risk" datasets would proceed to the second layer, which was the machine learning (random forest algorithm) layer. The second layer would diagnose the type of incipient fault present in the transformer using the learned insights from the input dataset itself.

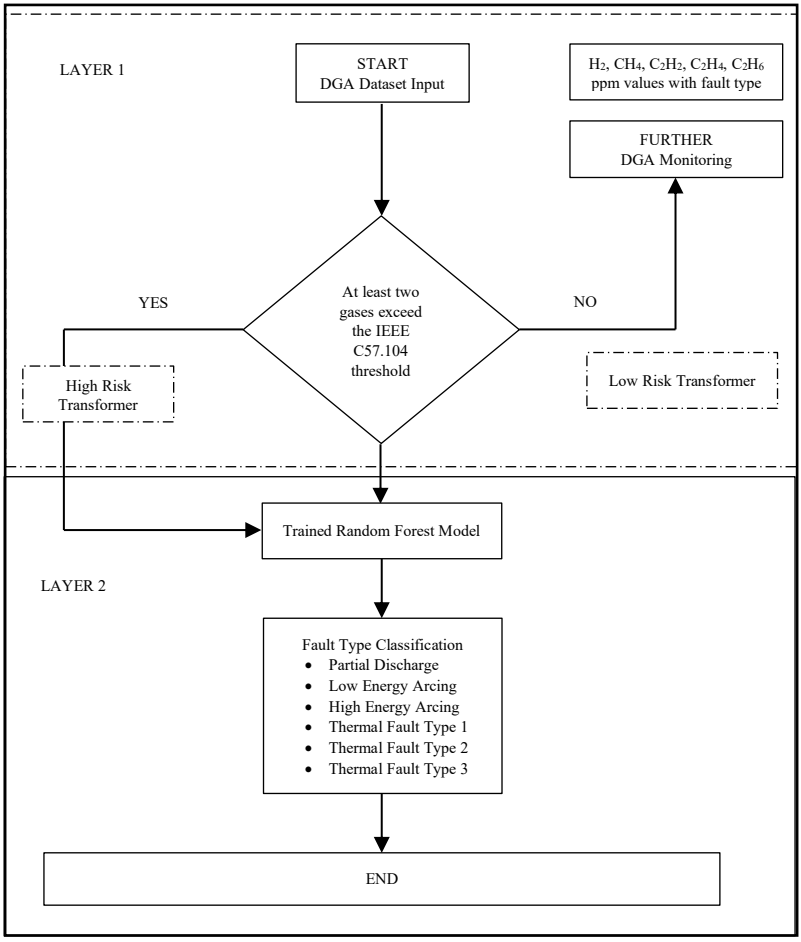


Figure 3. The Developed Two-Layer Framework

The model has two layers. The first layer filters the low-risk (not yet faulted) from the high- risk (possibly faulted). The simplified threshold for DGA levels using IEEE C59.104 is shown in Table 2. This is a simplified version than using the values of Table 1. The second layer, classifies the high-risk fault type. The second layer was classifying six specific fault types based on IEC 60599 standard (as shown in Figure 3). CO and CO_2 are not included in the DGA data since the CO/CO_2 ratio can only be significant if the ratio reached 10 (Banovic *et al*., 2015). From the overall dataset only 20% reached the said ratio so this study did not include the CO/CO_2 ratio as feature of the research dataset.

2.2 Sampling Rate

The development of the prediction model followed the standard machine learning protocol. As shown in Figure 4, the labeled DGA dataset samples were tested using the remaining 20% for validation (10%) and testing (10%).

Due to non-disclosure agreements (NDAs) with industries that own the dataset, the parties were legally bound to protect the confidentiality of the original data. This confidentiality was essential for maintaining trust and compliance with legal agreements, thereby safeguarding proprietary information, trade secrets, and potentially sensitive or confidential data points within the dataset.

If the accuracy of the random forest prediction model was low, the paper used the Random Search CV from sklearn library as the hyperparameter optimization method for this study. Random Search CV used the hard and fast range of hyperparameter settings sampled from mere chance distributions. As shown in Figure 4, compared to grid search CV, random search CV yields better results because of its high dimensionality in selecting hyperparameters. Overfitting was checked during the implementation using the stratified cross validation library in google colaboratory platform.

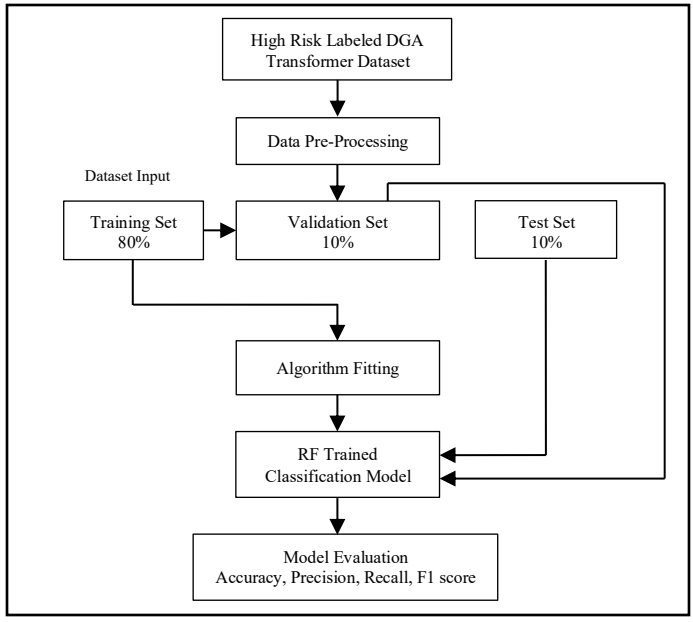


Figure 4. Classification Model Development and Evaluation

2.3 The Random Forest Algorithm Classifier

The Random Forest Algorithm is an ensemble algorithm of decision trees. The original training dataset is being resampled randomly and turned to n number of bootstrapped training sets which corresponds to the n decision trees that it would feed upon. The n predictions or outputs of n decision trees would be selected through majority voting thus arriving at one outcome as shown in Figure 5. This classifier tends not to overfit since there is no redundancy of dataset features in each n decision tree.

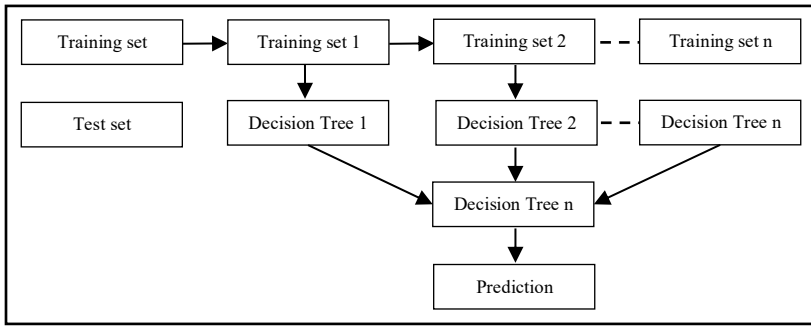


Figure 5. Random Forest Classifier Flow Chart (Ahmad *et al*., 2022)

2.4 Stratified K Fold Validation

While fitting and evaluating the classification model, another cross-validation was needed to ensure there is no overfitting during training. This study used the stratified K-fold validation because of the imbalanced characteristics of the DGA dataset. In K fold Cross-validation, the training datasets were grouped ("folded") into desired parts and trained individually. After which, the average of accuracies in each fold would be averaged as shown in Figure 6.

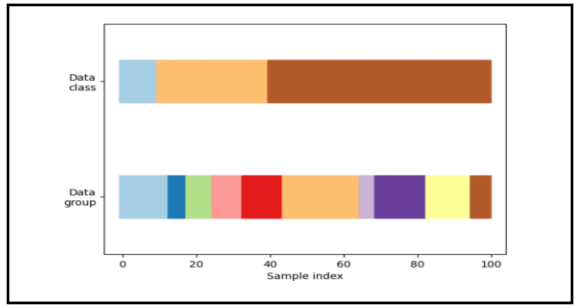


Figure 6. Sample Visualization of K Fold CV in Python
(The scikit-learn developers, 2025)

The band colors of the bar at the top represent the dataset's classification types (3 classes) in percentage. The band colors of the bar below represent the dataset's features (10 features), also in percentage.

2.5 Prediction Model Evaluation

2.5.1 Confusion Matrix and Classification Report Metrics

Standard metrics, such as accuracy and confusion matrix derivatives, would evaluate the prediction model. The confusion matrix is a fundamental metric in classification tasks. It identifies true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These values compute essential classification report metrics such as precision, recall, F1 score, and accuracy, which are calculated using Equations 2, 3, 4, and 1, respectively. Precision (Positive Predictive Value) measures the proportion of optimistic predictions that are correct. High precision indicates that the rate of positive class predictions was usually correct. Recall (Sensitivity/True Positive Rate), on the other hand, measures the proportion of actual positive instances that the model correctly identified. High recall means fewer false negative predictions. F score (F1 score) computes the harmonic mean of precision and recall. It provides a better metric of incorrectly classified cases than accuracy:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN \quad (1)$$

$$Precision = (TP)/(TP + FP) \quad (2)$$

$$Recall = (TP)/(TP + FN) \quad (3)$$

$$F1\text{-}score = 2*(Precision*Recall)/(Recall + Precision) \quad (4)$$

where, TP is the true positive predictions; TN = true negative predictions; FP = false positive predictions; TP = true positive prediction.

2.5.2 Area under the Precision Recall Curve (AUC-PR)

The Precision-Recall (PR) curve is a graphical representation used for binary classifications, but can also be used for multiclass classification using the One versus All library in the IDE. One class is treated as positive and the other classes as negative, thus treating them still as binary representation. The graph compares the precision to the recall performance of the algorithm at different thresholds.PR curve is useful when there is a data imbalance. Since this is a multiclassification task at hand, it is imperative to average the individual binary scores in the one vs all setup either by macro or micro averaging.

2.5.2.1 Micro Averaging over Macro Averaging

Macro averaging is averaging equally the class prediction in the PR curve. In an imbalanced dataset, precision and recall for minority classes are minimal due to fewer positive examples. This results in lower macro averaged scores in those classes, giving a biased overall view of the model's performance. The micro average calculates the metrics by aggregating the true positives, false positives, and false negatives across all classes before calculating the AUC PR score. Since the official dataset is quite imbalanced, it is beneficial for the study. The micro-averaging setup treats all predictions across all classes equally. This helps to accurately evaluate the algorithm based on individuality rather than the majority classes.

2.5.3 ROC Curve (Receiver Operating Characteristics Curve)

ROC Curve is also an established graphical method, like the PR Curve, but focuses on trade-offs between sensitivity (true positive rate) and specificity (false positive rate). It aids in understanding the model's ability to correctly identify positive cases while avoiding false alarms across all possible thresholds.

2.5.3.1 Macro Averaging over Micro Averaging of ROC Curve

Micro averaging aggregates the positives and negatives across all classes before computing the ROC Curve, which biases the curve to the majority classes. Macro averaging gives us more insight than micro averaging in this case. It prevents the majority class from dominating the overall metric; thus, minority classes were also given importance in the calculation. Therefore, this study used macro averaging in using the ROC Curve. The closer the AUC ROC value to 1, the better the model discriminates between positive and negative classes.

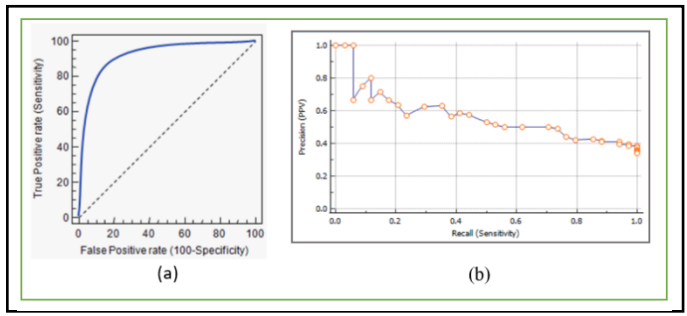


Figure 7. Sample ROC Curve and Precision Recall Curve (MedCalc Software, n.d.)

A model with an ROC curve closer to the top left corner is better able to differentiate between the positive and negative classes. Consequently, in a PR Curve, a model with a curve closer to the top right corner is a better model.

2.6 Traditional DGA Methods for Comparison

The Doernenburg Method, Duval Triangle, Rogers Ratio, and IEC Ratio were the conventional methods compared against the developed method. These methods were used for comparison with the proposed method as they were part of the IEEE C59.104 and IEC 60599, which govern the DGA interpretation of the DGA data of oil-filled transformers.

2.7 Machine Learning Methods for Comparison

Decision tree algorithm, J48 decision tree algorithm, SVM, artificial neural nets (ANN), and K nearest neighbor algorithm (KNN) were compared against the developed method. These algorithms represent different approaches to machine learning, offering a broad spectrum of techniques to compare. Decision trees provide a clear, interpretable model that mimics human decision-making. ANN models complex relationships through layers of interconnected nodes. SVM finds the optimal hyperplane that separates data into classes. The J48 algorithm is often used for its simplicity and efficiency. KNN is a non-parametric method that classifies data based on the closest training examples. These algorithms have different strengths and weaknesses, making them ideal for comparison. Decision trees and J48 were easy to understand and interpret, but can overfit complex datasets. ANN is good for capturing nonlinear relationships but requires substantial computational resources and tuning. SVM is effective in high-dimensional spaces but can be computationally intensive. Lastly, KNN is simple and intuitive, but can be slow with large datasets and sensitive to the k value and distance metrics. These algorithms are well-established in the literature, with extensive studies on their performance characteristics. This makes them reliable benchmarks for new algorithms or improvements.

2.8 Hyperparameter Tuning using RandomSearchCV

The accuracy of the classification model (random forest) after execution could be improved using hyperparameter tuning techniques. These popular techniques were gridsearchCV and randomsearchCV. Both were useful for tuning; however, randomsearchCV was used because of its advantages compared to gridsearchCV. RandomsearchCV takes less time to find the

"best" parameters as it does not follow a uniform search and can handle high dimensionality, as shown in Figure 8.

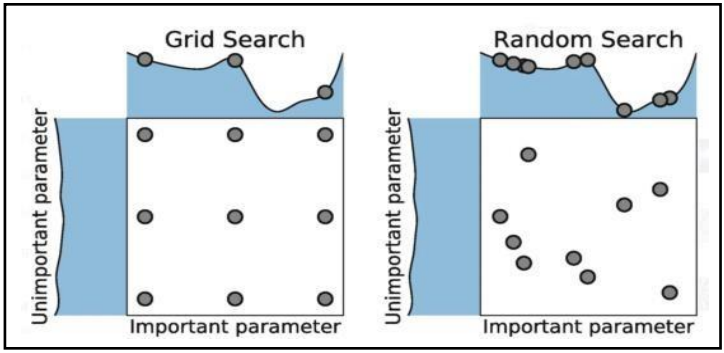


Figure 8. RamdomSearchCV Dimensionality (Bergstra *et al*., 2012)

Higher dimensionality means more combinations to evaluate, which can increase the search time exponentially. Therefore, it is important to strike a balance between exploring a wide range of hyperparameters and managing the computational cost. RandomizedSearchCV helps address this by randomly sampling a subset of combinations from the search space, making it more efficient than an exhaustive grid search.

Table 1. IEEE C57.104 2019 Gas Threshold for a Normal Transformer

		O_2/N_2 Ratio ≤ 0.2				O_2/N_2 Ratio > 0.2			
		Transformer Age in Years				Transformer Age in Years			
		Unknown	1-9	10-30	> 30	Unknown	1-9	10-30	>30
Gas	Hydrogen (H_2)	80	75		100	40	40		
	Methane (CH_4)	90	45	90	110	20	20		
	Ethane (C_2H_6)	90	30	90	150	15	15		
	Ethylene (C_2H_4)	50	20	50	90	50	25	60	
	Acetylene (C_2H_2)	1	1			2	2		
	Carbon monoxide (CO)	900	900			500	500		
	Carbon dioxide (CO_2)	9000	5000	10000		5000	3500	5500	

Table 2. Simplified IEEE C57.104 20 19 Gas Threshold for a Normal Transformer

H_2	CH_4	C_2H_2	C_2H_4	C_2H_6
100	110	2	90	150

2.9 Datasets Characteristics

The total number of labeled datasets for the framework is 277. Table 3 shows the breakdown of the fault types. The faults partial discharge (PD), low energy discharge (D1), and high energy discharge (D2) accounted for only 2.16%, 2.88%, and 1.8%, respectively, of the total data. The fault types mentioned above are difficult for dissolved gas analysis to detect because they are instantaneous faults. This confirms the literature about dissolved gas analysis, which is not effective in detecting these partial discharges, flashovers, and other instantaneous faults. Thermal faults 1, 2, and 3 constituted 7.22%, 19.49%, and 66.79% of the total dataset, respectively.

Table 3. Dataset Distribution

Fault Type	No. of dataset	Percentage
Partial Discharge (PD)	5	1.81
Low Energy Discharge (D1)	8	2.89
High Energy Discharge (D2)	5	1.81
Thermal Fault 1 (T1)	20	7.22
Thermal Fault 2 (T2)	54	19.49
Thermal Fault 3 (T3)	185	66.79
Total	277	100

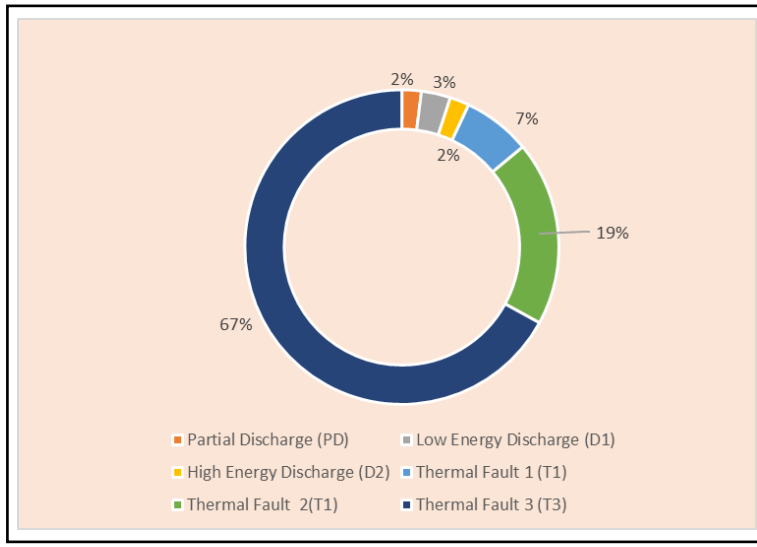


Figure 9. Dataset Distribution Chart

The dataset was mostly comprised of thermal faults, as shown in Figure 9. Thermal faults were more common in DGA interpretations because they developed gradually due to everyday stresses like overloading or cooling failures, leading to consistent gas production that is easily detected. In contrast, electrical faults like arcing and partial discharge were often sudden and catastrophic, so they are less frequently captured in DGA datasets.

	H2	CH4	C2H6	C2H4	C2H2
count	277.000000	277.000000	277.000000	277.000000	277.000000
mean	632.628159	81.581227	53.913357	78.018051	27.190614
std	4440.024596	322.490801	121.891968	185.578247	155.397037
min	3.000000	2.000000	1.000000	1.000000	0.000000
25%	15.000000	7.000000	13.000000	11.000000	0.000000
50%	20.000000	23.000000	26.000000	18.000000	0.000000
75%	26.000000	64.000000	40.000000	56.000000	0.500000
max	40280.000000	4704.000000	1060.000000	1629.000000	1880.000000

Figure 10. DGA dataset Statistics

Hydrogen (H_2) has a higher mean in Figure 10 because it is produced under diverse conditions, from normal operations to various fault types. Its standard deviation is high (4440.62) due to the significant variability in the amounts produced across different transformers, depending on the nature and severity of their conditions. The mean concentration of methane (CH_4) is lower than that of hydrogen because it is associated with specific types of faults, not the broad range of conditions that generate hydrogen. The standard deviation of methane is moderate, reflecting variations in the severity of partial discharges or thermal faults across different transformers. However, it is lower than hydrogen due to its more specific fault associations. The mean value of ethane (C_2H_6) is often lower than both hydrogen and methane, as significant thermal faults are less common across the entire population of transformers. The standard deviation for ethane is lower than that of hydrogen because its production is tied to more specific conditions. However, it can increase if the dataset includes transformers with varying degrees of thermal stress. The mean concentration of ethylene (C_2H_4) can be higher in datasets with older transformers or those operating under high stress. However, it is generally lower than hydrogen due to the more specific and severe conditions required for its production, as evident in Ethylene's standard deviation (Figure 10), which is usually higher than ethane but lower than hydrogen. This reflects that severe overheating is less common, but when it does occur, it can lead to significant variability in ethylene levels.

The mean value of acetylene (C_2H_2) is typically low because high-energy electrical discharges are relatively rare in transformers. When present, it indicates serious faults. Acetylene usually has a high standard deviation because its concentration can spike dramatically during severe faults, leading to significant variability in the dataset.

Figure 11 shows the pair plot or scatter plot matrix. This is the relationship in terms of distribution between two features of the DGA dataset. This is one of the visualizations of the dataset in Python. Vertical and horizontal dot patterns mean that the pair has no correlation, as in the case of C_2H_6 versus H_2 and also C_2H_6 versus C_2H_2. While patterns with a positive slope mean a positive correlation between C_2H_4 and C_2H_6. This scatterplot provides a unique overview of the dataset correlation and distribution.

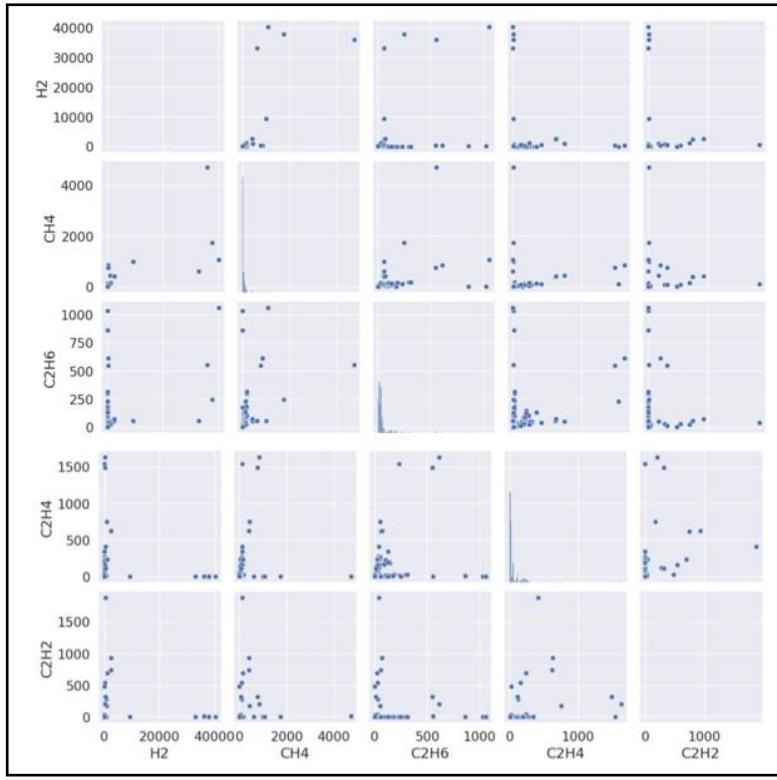


Figure 11. Pairplot of the DGA dataset

Figure 12 shows another visualization of the official dataset. This is a feature analysis. In each feature (gas), the distribution of gas values in frequency values is displayed. The first row shows the distribution of gases in terms of numerical value and their corresponding frequency. In the case of H_2, the values are more concentrated on the left side, which means the dataset values are not different from each other, similar to CH_4. While C_2H_6 and C_2H_4, their datapoints have some variations in terms of frequency. The second row is a pairplot matrix. The third row is the inverse representation of row one, as these graphs interchange the frequency and the gas values in the graph. Feature analysis provides deeper insights into the dataset and illustrates how these characteristics influence the framework into which the datasets are integrated.

While Figure 11 and 12 are graphical relationships and feature visualizations, Figure 13 shows the numerical correlation values in terms of two gases concerned. The negative correlation values indicate that the two gases have inverse relationships in the distribution as in the case of H_2-C_2H_4 pair. The positive correlation values show directly proportional relationship in the distribution of values. Strong association are in the case of H_2-CH_4 (0.76), followed by C_2H_6-CH_4 (0.45). Weak association were in the case of CH_4-C_2H_4, C_2H_6-C_2H_4.These means that they are least likely proportional in terms of distribution.

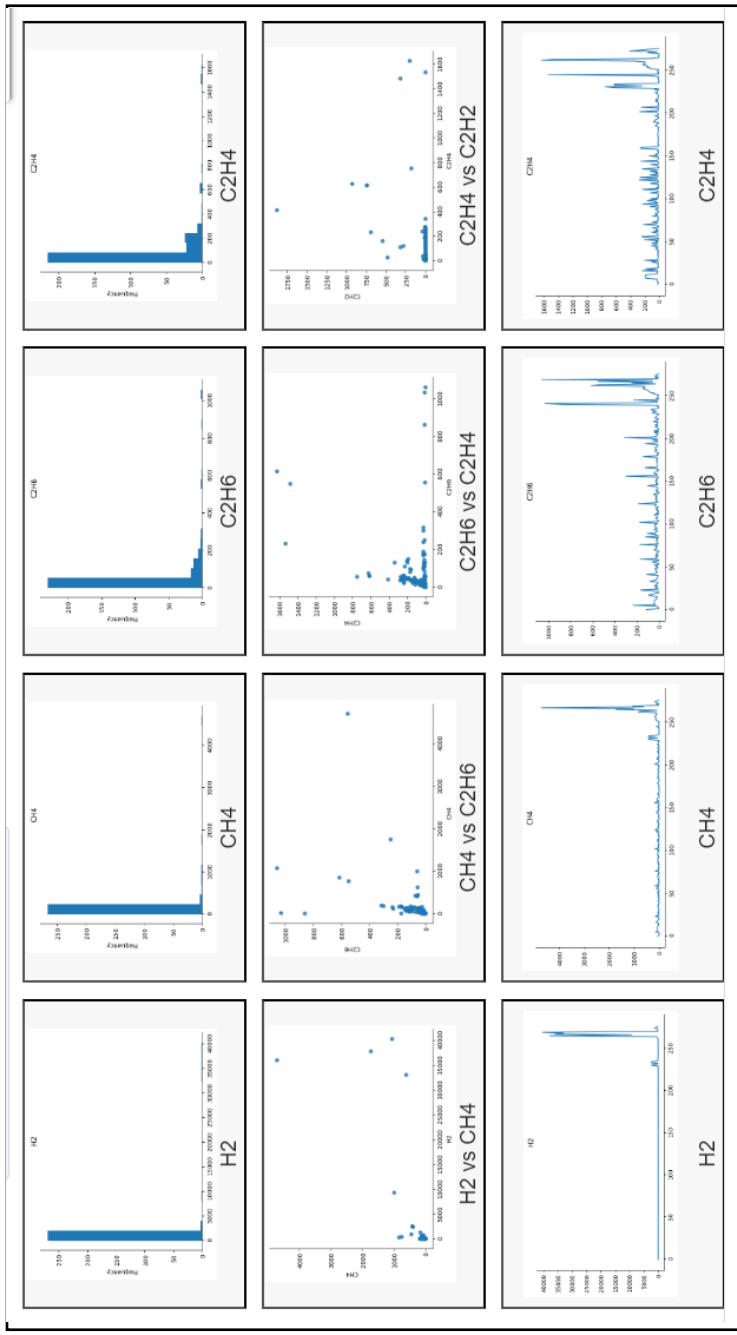


Figure 12. Graphical Distribution Feature Analysis of the Official Dataset

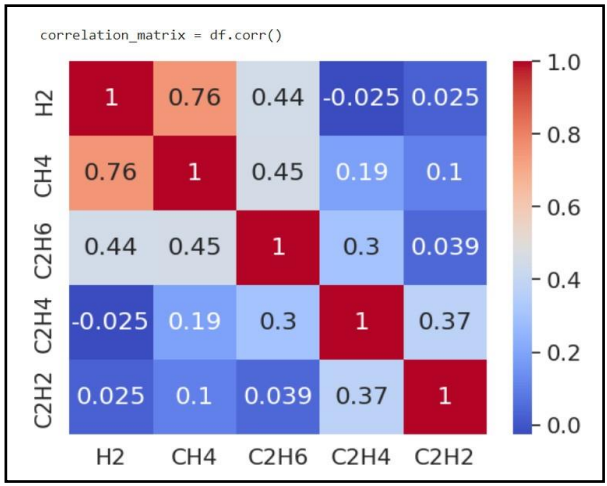


Figure 13. Correlation matrix of the DGA dataset

2.10 Stratified Sampling to handle Dataset Imbalance

Since there is an imbalance in the dataset of the study, this study used stratified sampling to address the issue. In stratified sampling, the selection of samples within each stratum is typically done randomly. This ensures that each unit in the dataset has an equal chance of being selected for the sample within its respective stratum. Therefore, while the overall process of stratified sampling involves deliberate grouping of the population into strata based on certain characteristics, the selection of samples within each stratum is random, which helps to reduce bias and ensure the representativeness of the sample.

2.11 Feature Engineering

2.11.1 Standardization of dataset

Standardization is one of the methods of feature scaling in a dataset. Dataset should be scaled in comparable limits so the machine learning algorithms would not tend to weigh greater values than lower ones. This method calculates the mean and standard deviation (SD) of the data and subtract the mean value from each entry and divide by the standard deviation (σ) as calculated in Equation 1. This method normalizes the data with a mean of zero and SD of 1:

$$X_{scaled} = \frac{X_i - X_{mean}}{\sigma} \quad (5)$$

Where, Xscaled = scaled feature, Xi = feature sample, Xmean = mean of the features, σ = standard deviation.

2.11.2 Binarization of Labels

The dataset labels were categorical values such as "PD" (partial discharge), DI (low energy discharge), D2 (high energy discharge) and three thermal faults T1, T2, T3. This nature of labels does not coincide to the numerical values of the feature gases which is in ppm. Though it is still possible to have categorical non-numerical labels in the fitting of ML methods, it is common practice and to use the *LabelBinarizer* library in scikit learn for all operations and evaluations that the dataset will undergo. This study is an example of this set up. This study used binarization in the labels (as calculated in Equation 2) for ease of navigation in the operations of datasets in the Google Colab platform. Below is the mathematical expression for binarization.

A set of possible classes $C = C_1, C_2, \ldots C_n$ and a label $\boldsymbol{y_j}$. The binarized vector y' for the label y_j would be expressed as:

$$y_i' = \delta\left(y_j, C_i\right) \tag{6}$$

where y_i' is the value at the ith position in the binarized vector. $\delta\,(y_j, C_i)$ is the Kronecker delta function.

2.12 Data Analysis

Figure 14 shows the data analysis flow of this study, which used the developmental method of data analysis. Figure 15 shows the detailed process flow of the study.

2.12.1 Gathering of Data

This section outlines and explains the processes involved in gathering data during and after designing and implementing the developed framework. Four phases were identified to facilitate data collection: the preliminary, design, implementation, and evaluation phases.

2.12.2 Preliminary Phase

In this phase communications: transmittal letters were sent to the respective Distribution Utilities (DUs) and electrical contractors requesting to entrust their DGA laboratory results of power transformers. This phase intended to

collect sufficient primary DGA datasets to be run through the algorithm in the next phase of this study. This phase prepared the necessary documents to get the primary data through the Notarized Non-Disclosure Agreement (NDA) with the power companies.

2.12.3 Design Phase

The design of the developed framework and the appropriate algorithm for the task were finalized. This phase completed the framework and determined which fault gases would be included in the final preprocessed data.

2.12.4 Implementation Phase

This study executed the algorithm (random forest algorithm) in a machine learning platform using the preprocessed DGA dataset.

2.12.5 Evaluation Phase

This study evaluated the performance of the trained classification model using a random forest algorithm using accuracy metrics, precision, recall, and F1 score. This phase also checked the overfitting tendency using stratified K-fold validation. Optimization took place also in this phase if necessary.

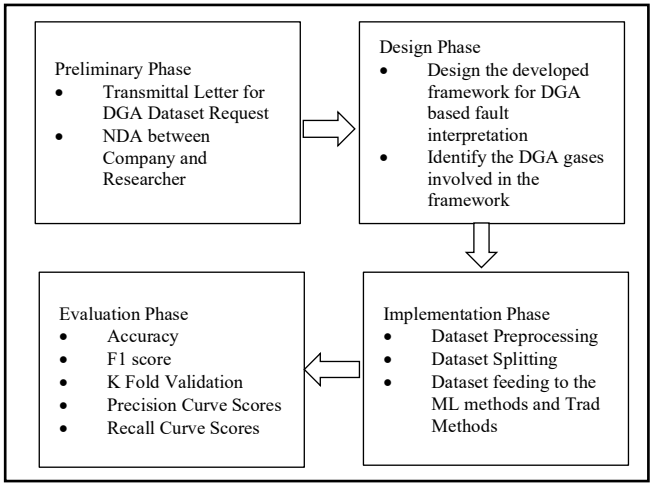


Figure 14. Data Analysis Flow

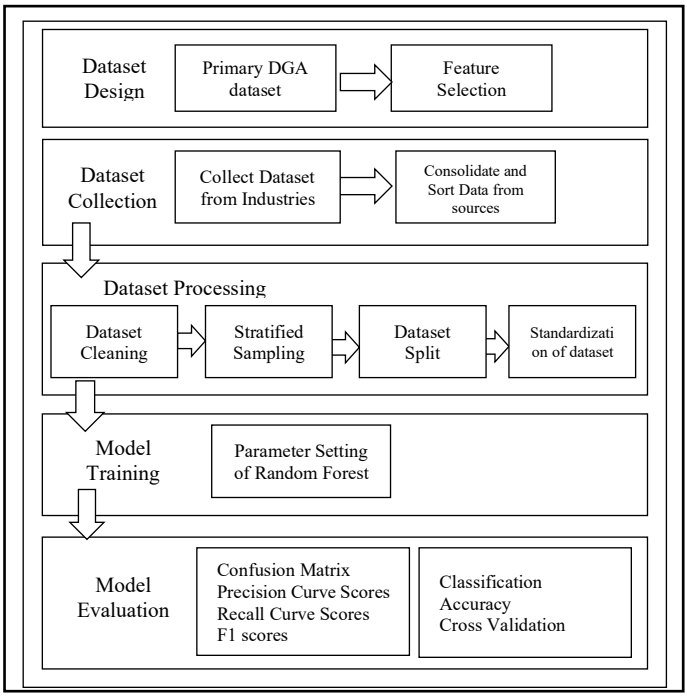


Figure 15. Dataset Architecture and Process Flow

3. Results and Discussion

3.1 Performance of DGA Traditional Methods

As shown in Figure 16, Duval Triangle method got the highest accuracy among other traditional methods in the table. This confirms the literature that Duval was the most efficient trad method in comparison with the methods in the list. The developed RF model surpassed the Duval Triangle Method accuracy by 1.08%. The developed model is 72.50% higher than the Doernenburg Method. In the case of the Rogers Ratio method and IEC Ratio Method, the model is 67.99% and 71.68% higher, respectively.

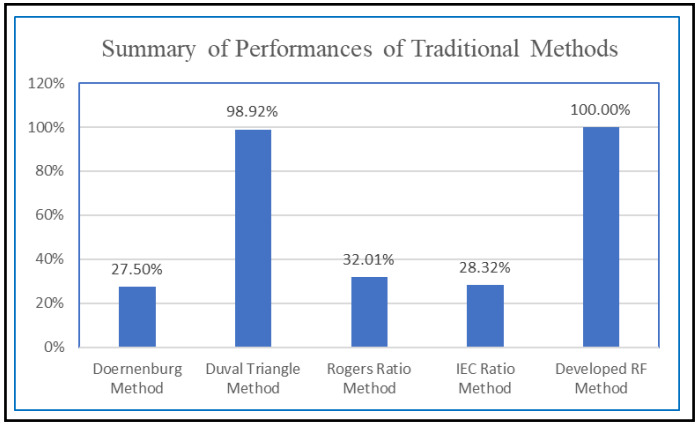


Figure 16. Performance Summary of Traditional Methods

3.2 Performance of Machine Learning Methods

In Figure 17 the J48, decision tree and random forest got the superior accuracy (100%) when compared to several well-established machine learning algorithms used in this paper, including k-Nearest Neighbors (KNN) (85.7%), Support Vector Machine (SVM) Linear (78.6%) and Artificial Neural Network (MLP classifier) (78.60%). Among the top three ML performers the random forest got the highest validation accuracy of 85.7%. This means that RF got the best generalization capabilities: learning underlying patterns that generalize to new data.

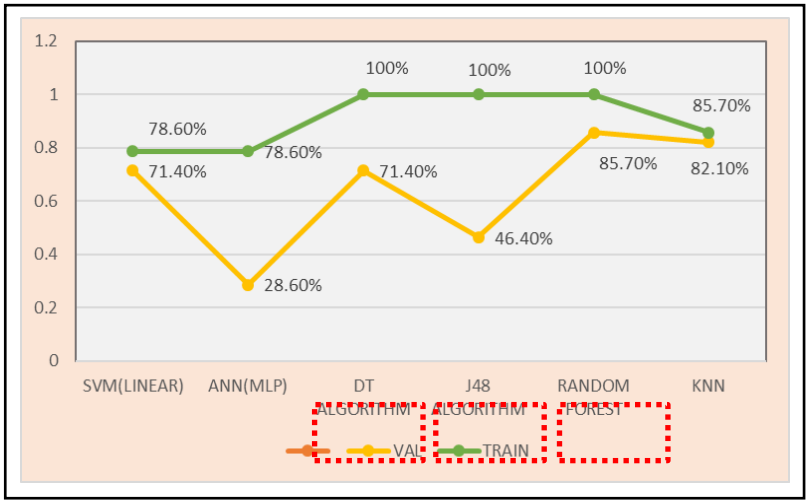


Figure 17. Accuracy Summary of Machine Learning DGA Methods

Table 4. Cross Validation Scores of ML Methods

ML Methods	5 fold Cross Validation Scores (80/10/10 Split)					Average CV Scores	Accuracy
KNN	0.888889	0.840909	0.795455	0.772727	0.886364	0.836869	Val:82.1% Test: 85.7%
SVM (linear)	0.755556	0.727273	0.750000	0.727273	0.727273	0.737475	Val: 71% Test:78.6%
ANN (MLP Classifier)	0.755556	0.795455	0.750000	0.750000	0.727272	0.755657	Val:28.6% Test:78.6%
DT Algorithm	0.911111	0.772727	0.863636	0.909091	0.954545	0.882222	Val:71.4% Test: 100%
J48 DT Algorithm (Developed Method)	0.866667	0.886364	0.863636	0.818182	0.886364	0.864242	Val: 46.4% Test: 100%
Random Forest	0.888889	0.863636	0.931818	0.863636	0.909009	0.891414	Val: 85.7% Test: 100%

Classification accuracy is not guaranteed to be the sole parameter for evaluating ML methods. Cross-validation is an essential tool to validate that there is no overfitting during the training process of identifying transformer faults. In a 5-fold cross-validation, k-fold cross-validation is used in this study. In this case, 5-fold CV was used. As shown in Table 4, the average accuracy across all 5 tests was calculated to get an average accuracy which in this case of Random Forest is 89.14%. This score is the highest among other ML algorithms used in comparison in this study. Good scores across different folds indicate that the model is stable and robust.

Table 5. Evaluations and Visualizations of ML Methods

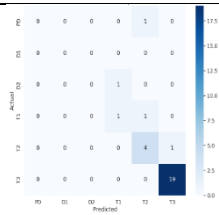
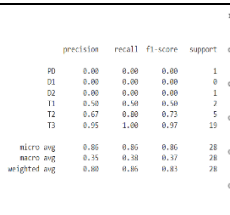
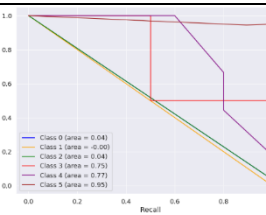
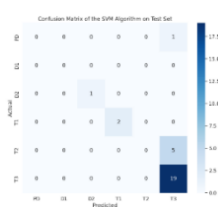
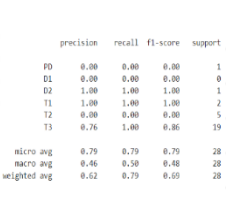
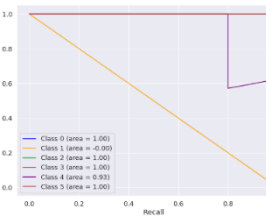
ML	Confusion matrix	Classification Report	Precision Recall Curve (AUC)
KNN			
SVM (linear)			

Table continued.

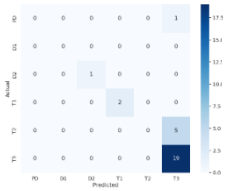
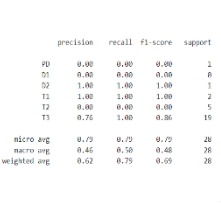
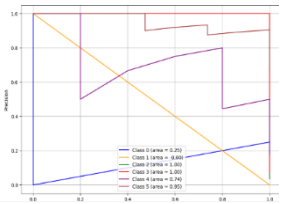
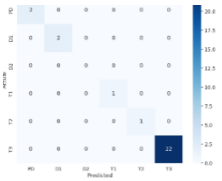
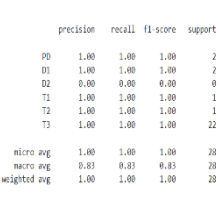
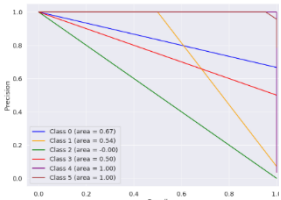
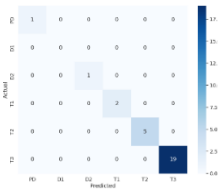
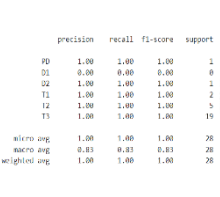
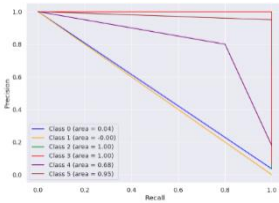
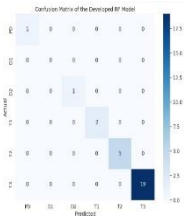
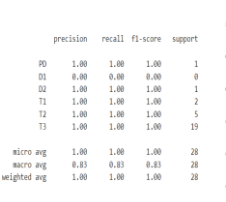
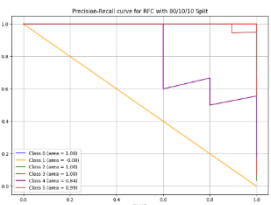
ANN MLP Classifier)			
DT Algorithm			
J48 Algorithm			
Random Forest			

Table 6. F1 score and PR AUC scores of ML methods

ML Method	Weighted F1 score	PR Curve (AUC)
KNN	83.00%	42.5%
SVM(linear)	69%	82.2%
ANN (MLP Classifier)	69.00%	65.7%
DT Algorithm	100%	61.8%
J48 DT Algorithm	100%	61.2%
(Developed Method) Random Forest	100%	80.67%

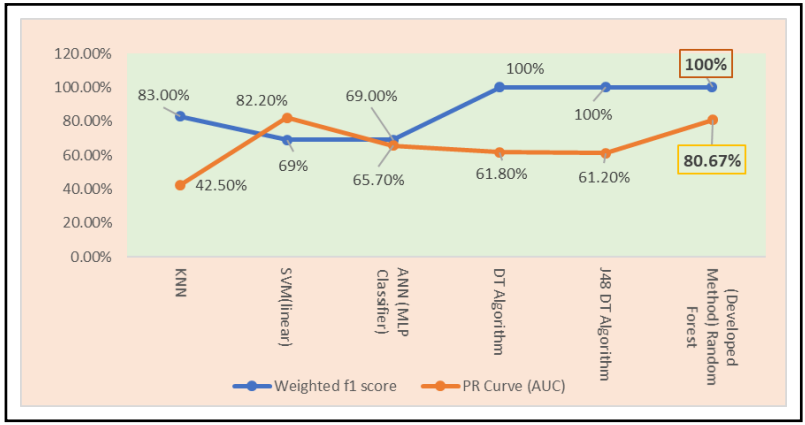


Figure 18. F1 score and PR Curve scores of ML Methods (Graphical)

In a diagnostic ML such as the framework of this study, precision recall and F1 score were crucial metrics due to their ability to provide insights into the model's performance. Moreover these metrics were useful for an imbalanced dataset. The F1 score of the ML methods and the average scores of the PR AUC Curve are calculated. DT, J48 and RF still got the highest F1 score. Table 5 showed the visuals of the confusion matrix, classification reports and PR Curves of the ML methods. Random Forest also got the high PR_AUC score which is 80.67% as shown in Table 6. SVM got the highest PR AUC Score (82.2%) but fall short on the F1 score (69%). Figure 18 showed the graphical visualization of the F1 scores and PR Curves (AUC).

3.3 Further Evaluation of top three ML Methods (J48, DT, Random forest)

Since J48, DT, and Random Forest were almost the same in their accuracy, cross-validation, and F1 score performances. The behavior of these three ML methods was further analyzed using two different data splits and reevaluated their accuracies and PR-AUC Curve, as shown in Table 7. Figures 19 and 20 show the bar graph version of Table 7 to better emphasize the value difference of scores. Table 8 shows the actual ROC, PR, and classification reports of the 80/20 and 70/30 splits of the top three ML methods. This time ROC_AUC Curve was added for evaluation. Through these further evaluations, it can be concluded which ML performs better given the dataset changes in training and testing.

Table 7. Precision Recall (AUC) Curve Scores and ROC (AUC) Curve Scores

ML Method	Data Split	PR AUC Curve Score (micro average)	ROC AUC Curve Score (macro average)
Decision Tree Classifier	70% train 30% test	84.00%	91.00%
	original split	92.00%	94.00%
	80% train 20% test	87.00%	96.00%
J48 DT Classifier	70% train 30% test	88.00%	90.00%
	original split	92%	83%
	80% train 20% test	87.00%	95.00%
Random Forest Algorithm (Developed Method)	70% train 30% test	94%	95%
	original split	96%	94%
	80% train 20% test	89%	97%

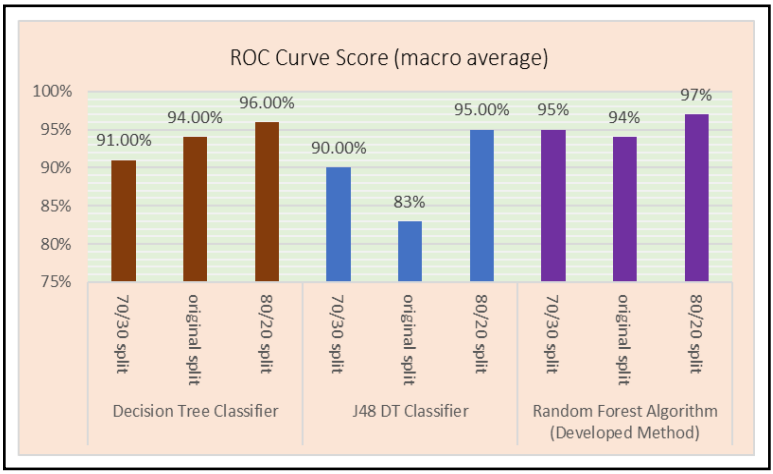


Figure 19. ROC (AUC) Curve for different datasplit

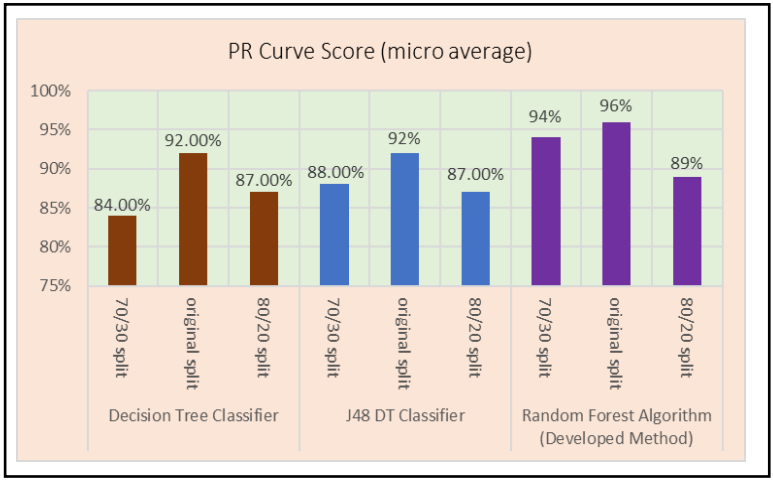


Figure 20. PR (AUC) Curve scores (diff. split)

During the execution of the three ML methods in two different data splits, all got a classification accuracy of 100%. This is expected since the dataset is small and easier for an ML to learn and make correct predictions. The random forest got the overall highest scores among DT and J48 both for ROC-AUC and PR AUC Curve scores for three data splits as shown in Table 9. The Random forest got an average of 95.33% (ROC_AUC score) and 93.00% (PR -AUC Curve) across multiple splits. This ML is followed by DT algorithm 93.66(ROC_AUC score) and 87.67(PR-AUC Score). J48 got an average of 89.33% (ROC_AUC score) and 89.00% (PR AUC Curve).

A high ROC curve (AUC) score means that the model has excellent capacity in discriminating between positive and negative classes especially in imbalanced datasets. On the other hand high PR Curve (AUC) score signifies that the model is very good in predicting positives accurately which is essential in diagnostic settings similar to the proposed framework of the study.

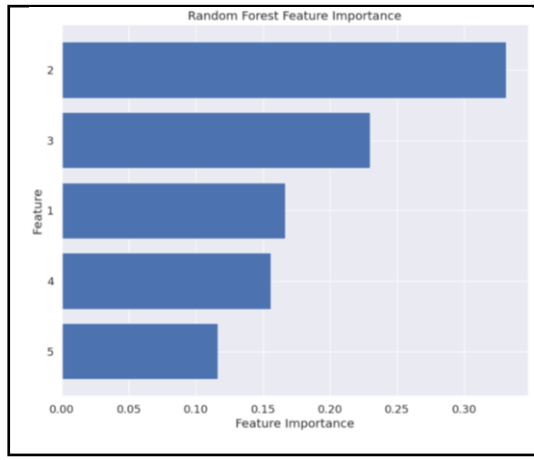


Figure 21. Feature Importance Analysis Graph

Feature importance analysis visualization shows the importance of each feature in making predictions. It helps identify which features are most influential in the model's decision-making process. In Figure 21, It showed that the methane gas (CH_4) (feature 2) gas has the highest influence feature among all gases comprised followed by C_2H_6 (feature 3:ethane), H_2(feature1: hydrogen), C_2H_4 (feature 4:ethylene) and C_2H_2 (feature 5: acetylene) respectively.

Table 8. Evaluation Visuals based on 80/20 Split

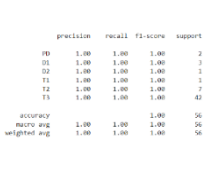
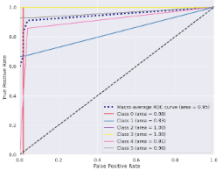
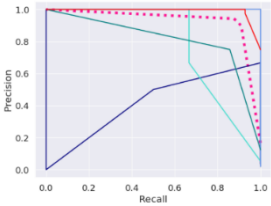
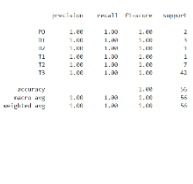
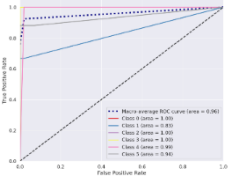
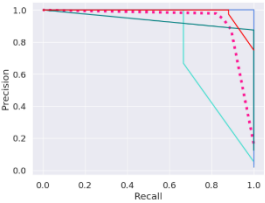
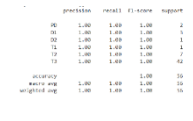
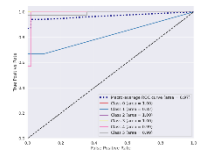
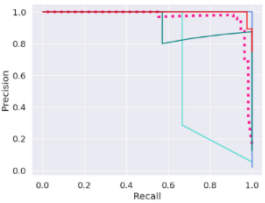
Evaluation for 80/20 SPLIT					
ML	Classification Report	ROC Curve	Precision (AUC)	Recall	Curve
J48					
Table continued.					
DT					
Random Forest Algorithm					

Table 9. Evaluation Visuals based on 70/30 Split

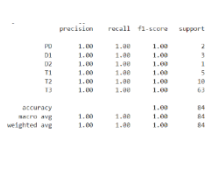
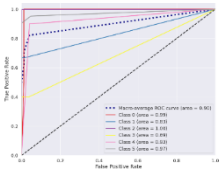
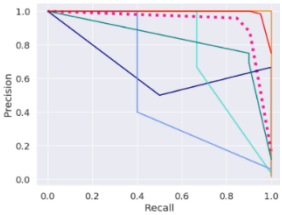
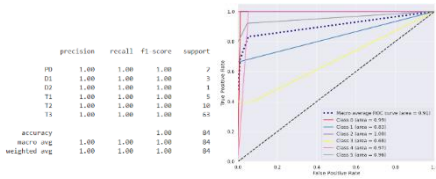
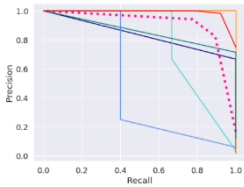
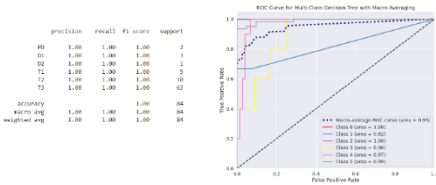
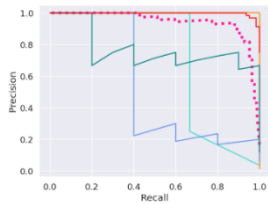
Evaluation Result for 70/30 SPLIT					
ML	Classification Report	ROC Curve (AUC)	Precision (AUC)	Recall	Curve
J48					

Table continued.

DT				
Random Forest Algorithm				

3.4 Some Limitations of the Proposed Framework

Though the framework had the accuracy of 100% it does not correlate to a perfect model. The dataset was limited and imbalance in nature which made the developed method more advantageous in its implementation.

Interpretation of DGA thresholds that the IEEE C59.104 set is more likely an art than science since interpretation of gas ratios can be ambiguous, sometimes leading to false positives or negatives regarding the health of transformers. The standard may not detect minor faults early enough, as it is more effective for identifying severe issues. Fixed threshold values may not account for variations in different transformer designs and operating conditions. Lastly, requires expert knowledge for accurate interpretation, which can limit its usability in the field.

Random forest algorithm, though powerful, has several limitations. Its complexity and multiple decision trees make it difficult to interpret. Training requires significant computational resources, making it time-consuming and memory-intensive. Overfitting remains a risk if hyperparameters are not properly tuned. Feature importance estimates can be biased, especially with varying feature scales or categorical levels. The algorithm struggles with imbalanced data unless special techniques are applied as in the case of the dataset of the study. While somewhat robust to noise, performance can still degrade with noisy data. Scaling to very large datasets is challenging due to high computational demands. Lastly, though the absence of expert opinion was the feature of this framework, the important rule of good engineering

judgment was still an indispensable factor in the fault diagnosis of transformers.

3.5 Challenges in using the Framework in Real World Setting

Most industries may not have easy access to DGA equipment or expertise therefore the proposed framework cannot be fully utilized. Secondly, there is a lack of machine learning background skills in most of power industry technical teams. The integration concerns of ML frameworks in the existing maintenance workflows seen to be a challenge also. Lastly, the developed model raise ethical concerns related to data privacy, bias, fairness, and accountability.

4. Conclusion and Recommendation

Based on the results of the study, Random Forest proved to be the most powerful and robust ML method in the study in comparison to the other ML used. It surpassed as well the traditional methods such as the Doernenburg method (27.50% accu.), Rogers Ratio (32.01% accu.), and IEC Ratio method (28.32% accu.). Moreover, it performed very well with its ML counterparts. These results confirmed the literature about the good performance of random forest algorithms in DGA-based fault classification tasks (Belgiu and Drăguţ, 2016; Wager and Athey, 2018).There was also no overfitting during the execution of the algorithm because of the good scores during cross-validation. This RF's CV score is the highest among the compared ML methods in the study. The RF performed better than the J48 algorithm and Decision tree Algorithm in terms of AUC curve scores (ROC Curve and PR Curve). This confirmed the robustness and versatility of the random forest classifier in the proposed two-layer framework together with the IEEE C59-104 Standard in diagnosing incipient faults of transformers.

In light of the results and the findings of the study, these were the made recommendations. It was recommended to use a greater number of primary DGA datasets to get a more conclusive output for the prediction model. It was also better to include stray gassing factors that may affect DGA diagnosis. There is also a concern in terms of balancing the dataset representation of fault types as it is not addressed in this paper. It is also other qualitative methods and tests results of transformer parts to have a holistic approach in the context of transformer health monitoring.

5. Acknowledgement

Immeasurable appreciation and deepest gratitude are extended to those who, in one way or another, contributed to making this study possible. Special thanks to Engr. Jayson Jueco, who was always ready to assist whenever clarification related to this study was needed. Gratitude is also expressed to Emerita, for her support and words of encouragement.

6. References

Abu-Elanien, A.E.B., Salama, M.M.A., & Ibrahim, M. (2011). Determination of transformer health condition using artificial neural networks. *2011 International Symposium on Innovations in Intelligent Systems and Applications*, 1–5. https://doi.org/10.1109/INISTA.2011.5946173

Ahmad, G.N., Ullah, S., Algethami, A., Fatima, H., & Akhter, S. Md. H. (2022). Comparative Study of Optimum Medical Diagnosis of Human Heart Disease Using Machine Learning Technique With and Without Sequential Feature Selection. *IEEE Access*, *10*, 23808–23828. https://doi.org/10.1109/ACCESS.2022.3153047

Aizpurua, J.I., Stewart, B.G., McArthur, S.D.J., Lambert, B., Cross, J.G., & Catterson, V.M. (2019). Improved power transformer condition monitoring under uncertainty through soft computing and probabilistic health index. *Applied Soft Computing*, *85*, 105530. https://doi.org/10.1016/j.asoc.2019.105530

Aminifar, F., Abedini, M., Amraee, T., Jafarian, P., Samimi, M.H., & Shahidehpour, M. (2022). A review of power system protection and asset management with machine learning techniques. *Energy Systems*, *13*(4), 855–892. https://doi.org/10.1007/s12667-021-00448-6

Anil, A.K., & Archana, R. (2017). *A Review of Dissolved Gas Analysis and Transformer Health Condition*. *2*, 2395–4396. www.ijariie.com

Azmi, A., Jasni, J., Azis, N., & Kadir, M.Z.A. Ab. (2017). Evolution of transformer health index in the form of mathematical equation. *Renewable and Sustainable Energy Reviews*, *76*, 687–700. https://doi.org/10.1016/j.rser.2017.03.094

Banovic, M., Ramachandran, P., Rego, N., & Justiz P. (2015) Transformer Magazine 2(1), 34. Retrieved from https://transformers-magazine.com

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bergstra, J., Ca, J.B., & Ca, Y.B. (2012). Random Search for Hyper-Parameter Optimization Yoshua Bengio. In *Journal of Machine Learning Research* (Vol. 13). http://scikit-learn.sourceforge.net.

Da Silva, D.G.T., Braga Da Silva, H.J., Marafão, F.P., Paredes, H.K.M., & Gonçalves, F.A.S. (2021). Enhanced health index for power transformers diagnosis. *Engineering Failure Analysis*, *126*, 105427. https://doi.org/10.1016/j.engfailanal.2021.105427

Dai, J., Song, H., Sheng, G., & Jiang, X. (2017). Dissolved gas analysis of insulating oil for power transformer fault diagnosis with deep belief network. *IEEE Transactions on Dielectrics and Electrical Insulation*, *24*(5), 2828–2835. https://doi.org/10.1109/TDEI.2017.006727

de Castro-Cros, M., Velasco, M., & Angulo, C. (2021). Machine-Learning-Based Condition Assessment of Gas Turbines—A Review. *Energies*, *14*(24), 8468. https://doi.org/10.3390/en14248468

Ekojono, Prasojo, R.A., Apriyani, M.E., & Rahmanto, A.N. (2022). Investigation on machine learning algorithms to support transformer dissolved gas analysis fault identification. *Electrical Engineering*, *104*(5), 3037–3047. https://doi.org/10.1007/s00202-022-01532-5

Faiz, J., & Soleimani, M. (2017). Dissolved gas analysis evaluation in electric power transformers using conventional methods a review. *IEEE Transactions on Dielectrics and Electrical Insulation*, *24*(2), 1239–1248. https://doi.org/10.1109/TDEI.2017.005959

Faiz, J., & Soleimani, M. (2018). Assessment of computational intelligence and conventional dissolved gas analysis methods for transformer fault diagnosis. *IEEE Transactions on Dielectrics and Electrical Insulation*, *25*(5), 1798–1806. https://doi.org/10.1109/TDEI.2018.007191

Guo, H., & Guo, L. (2022). Health index for power transformer condition assessment based on operation history and test data. *Energy Reports*, *8*, 9038–9045. https://doi.org/10.1016/j.egyr.2022.07.041

Ibrahim, S.I., Ghoneim, S.S.M., & Taha, I.B.M. (2018). DGALab: an extensible software implementation for DGA. *IET Generation, Transmission & Distribution*, *12*(18), 4117–4124. https://doi.org/10.1049/iet-gtd.2018.5564

IEEE. (2019). *IEEE guide for the interpretation of gases generated in oil-immersed transformer IEEE Std C57. 104™-2019*. IEEE Power Energy Society.

Jamshed, A., Chatterjee, K., & Haque, N. (2021). Random Forest Classifier based Dissolved Gas Analysis for Identification of Power Transformer faults using Gas Ratio Data. *2021 2nd International Conference for Emerging Technology (INCET)*, 1–5. https://doi.org/10.1109/INCET51464.2021.9456256

Kumar, K.R., & Haque, N. (2022). A Comparative Analysis of Different Machine Learning Algorithms for Classification of Partial Discharge Signals under HVDC. *2022 3rd International Conference for Emerging Technology (INCET)*, 1–5. https://doi.org/10.1109/INCET54531.2022.9824127

Li, H., Wang, Y., Liang, X., He, Y., & Zhao, Y. (2018). Nonparametric Kernel Density Estimation Model of Transformer Health Based on Dissolved Gases in Oil. *2018 IEEE Electrical Insulation Conference (EIC)*, 236–239. https://doi.org/10.1109/EIC.2018.8481032

Liu, Y., & Bao, Y. (2022). Review on automated condition assessment of pipelines with machine learning. *Advanced Engineering Informatics*, *53*, 101687. https://doi.org/10.1016/j.aei.2022.101687

Malik, H., & Mishra, S. (2016). Application of Gene Expression Programming (GEP) in Power Transformers Fault Diagnosis Using DGA. *IEEE Transactions on Industry Applications*, *52*(6), 4556–4565. https://doi.org/10.1109/TIA.2016.2598677

MedCalc Software. (n.d.). *ROC curve analysis, method comparison and quality control tools*. Retrieved July 5, 2025, from https://www.medcalc.org/features/roccurves.php

Murugan, R., & Ramasamy, R. (2019). Understanding the power transformer component failures for health index-based maintenance planning in electric utilities. *Engineering Failure Analysis*, *96*, 274–288. https://doi.org/10.1016/j.engfailanal.2018.10.011

Rao, U.M., Fofana, I., Rajesh, K.N.V.P.S., & Picher, P. (2021). Identification and Application of Machine Learning Algorithms for Transformer Dissolved Gas Analysis. *IEEE Transactions on Dielectrics and Electrical Insulation*, *28*(5), 1828–1835. https://doi.org/10.1109/TDEI.2021.009770

Senoussaoui, M.E.A., Brahami, M., & Fofana, I. (2018). Combining and comparing various machine-learning algorithms to improve dissolved gas analysis interpretation. *IET Generation, Transmission & Distribution*, *12*(15), 3673–3679. https://doi.org/10.1049/iet-gtd.2018.0059

Sholevar, N., Golroo, A., & Esfahani, S.R. (2022). Machine learning techniques for pavement condition evaluation. *Automation in Construction*, *136*, 104190. https://doi.org/10.1016/j.autcon.2022.104190

Sun, C., Ohodnicki, P.R., & Stewart, E.M. (2017). Chemical Sensing Strategies for Real-Time Monitoring of Transformer Oil: A Review. *IEEE Sensors Journal*, *17*(18), 5786–5806. https://doi.org/10.1109/JSEN.2017.2735193

Sun, L., Ma, Z., Shang, Y., Liu, Y., Yuan, H., & Wu, G. (2016). Research on multi-attribute decision-making in condition evaluation for power transformer using fuzzy AHP and modified weighted averaging combination. *IET Generation, Transmission & Distribution*, *10*(15), 3855–3864. https://doi.org/10.1049/iet-gtd.2016.0381

Temple, D.D., & Duncan, W. (1989). *Facilities instructions, standards, & techniques* (1st ed., Vol. 2). United States Department of the Interior.

The scikit-learn developers. (2025). *Visualizing cross-validation behavior in scikit-learn*. https://scikit-learn.org/stable/auto_examples/model_selection/plot_cv_indices.html

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Wani, S.A., Gupta, D., Farooque, Md. U., & Khan, S.A. (2019). Multiple incipient fault classification approach for enhancing the accuracy of dissolved gas analysis (DGA). *IET Science, Measurement & Technology*, *13*(7), 959–967. https://doi.org/10.1049/iet-smt.2018.5135

Wani, S.A., Rana, A.S., Sohail, S., Rahman, O., Parveen, S., & Khan, S.A. (2021). Advances in DGA based condition monitoring of transformers: A review. *Renewable and Sustainable Energy Reviews*, *149*, 111347. https://doi.org/10.1016/j.rser.2021.111347

Wong, S.Y., Ye, X., Guo, F., & Goh, H.H. (2022). Computational intelligence for preventive maintenance of power transformers. *Applied Soft Computing*, *114*, 108129. https://doi.org/10.1016/j.asoc.2021.108129

Zhao, C., Xu, X., Chen, H., Wang, F., Li, P., He, C., Shi, Q., Yi, Y., Li, X., Li, S., & He, D. (2023). Exploring the Complexities of Dissolved Organic Matter Photochemistry from the Molecular Level by Using Machine Learning Approaches. *Environmental Science & Technology*, *57*(46), 17889–17899. https://doi.org/10.1021/acs.est.3c00199