International Journal Of
Industrial Electronics Control and Optimization
Print ISSN: 3517-2645
Online ISSN: 3568-2645
Homepage: https://ieco.usb.ac.ir/

A Comparative Analysis of AI Algorithms for Power Transformer Fault Diagnosis Using Dissolved Gas Analysis

Hamid Reza Sezavar[1] | Hamid Karimi[1] | Navid Fahimi[2]

Department of Electrical and Computer Engineering, Qom University of Technology, Qom, Iran.[1]
Department of Electrical and Computer Engineering, Iran University of Technology, Tehran, Iran.[2]
Corresponding author's email: sezavar@qut.ac.ir

Article Info

Article type:
Research Article

Article history:
Received: **********
Received in revised form: **********
Accepted: **********
Published online: **********

Keywords: [3 to 5 Keywords]
Word,
Word,
Word,
Word.
[Authors are suggested to enter keywords or phrases in alphabetical order, separated by commas. For a list of suggested keywords by the IEEE Taxonomy, visit]

ABSTRACT

A comparative approach is pretend in this paper that evaluates different Artificial Intelligence (AI) methods for diagnosing power transformer faults using Dissolved Gas Analysis (DGA). Traditional approaches like the Rogers Ratio Method and Duval Triangle have been used for many years, but offer unreliable results for complex cases. Although, newer AI methods present better results, but still vary in how well they work. In this paper, several AI approaches are evaluated including Support Vector Machines (SVMs), Random Forest (RF), Gradient Boosting Machines (GBMs), Deep Neural Networks (DNNs) and a new combinational model is proposed based on comparing results. A real DGA dataset is used for covering six different fault types for the proposed testing. The results show that while all AI methods do better than traditional approaches, the combinational approach performs the best with 92.3% accuracy. This is found 20.2% better than traditional methods and 4.8% better than the best single AI model. Rational explanation is provided for how each method works and practical recommendation is presented for choosing the right approach based on particular requirements and available resources in real practices.

I. Introduction

Power transformers represent critical infrastructure components in which their operational reliability directly impacts power system stability. Transformer malfunctions, resulting from electrical, thermal, or mechanical stresses on their insulation systems (comprising from oil and paper) can lead to system disconnections with substantial consequences for electricity utilities. Dissolved Gas Analysis (DGA) has gained widespread interest as a primary diagnostic methodology for identifying incipient transformer faults. Traditional DGA interpretation techniques, including the IEC Code, Rogers Ratio Method, and Duval Triangle, have demonstrated diagnostic limitations in both accuracy and reliability. Research by [1] highlights these limitations through the development of a Transformer Fault Diagnosis Intelligent System (TFDIS) that integrates outputs from four distinct DGA methodologies [2]: Code Tree 2020, Modified IEC, Rogers' Ratio, and Neural Pattern Recognition. Their hybrid intelligent system achieved a diagnostic accuracy of 89.12%, surpassing the highest individual method accuracy of 86.01% from neural pattern recognition. This work demonstrates the potential of integrated diagnostic approaches but also reveals persistent accuracy limitations below 90%.

While theoretical development of DGA interpretation methods has progressed significantly, but a substantial research gap still exists regarding empirical field-verified correlations between DGA results and physical inspection findings across diverse fault categories. In [3], this gap is addressed through an integrated diagnostic approach

combining DGA with physical inspection across three industrial transformer case studies encompassing high-energy electrical discharge faults and severe thermal faults exceeding 700°C. Their investigation employed established DGA methods including Rogers ratios, Doernenburg ratios, basic gas ratios, and Duval Triangle techniques [2, 4]. Long-term DGA trend monitoring in one experimental study revealed progressing faults, enabling timely preventive maintenance (PM). Physical inspections consistently confirmed DGA-based fault predictions for all cases, demonstrating robust diagnostic consistency under actual operating conditions. This research empirically validated established DGA frameworks through integrated field evidence and multi-method cross-verification that verified DGA as a cost-effective PM tool while highlighting the importance of field validation studies.

Artificial Intelligence (AI) applications to DGA interpretation have been remarkably evolved in recent years, with diverse Machine Learning (ML) approaches that represented varying degrees of success. Initial implementations focused on shallow learning algorithms [5], particularly Artificial Neural Networks (ANNs) that employes various architectures including feedforward networks, radial basis function networks, or probabilistic neural networks. Support Vector Machines (SVMs) and Self-Organizing Maps (SOMs) with diverse kernel functions have been extensively evaluated for DGA applications as well [6]. Additional shallow learning approaches applied to DGA include decision trees [7], belief networks [8], Fuzzy Logic (FL) systems [9], capsule networks [10], extreme learning machines [11], and classifier ensembles. More recently, Deep Learning (DL) algorithms have been increasingly adopted for DGA analysis, even though comprehensive performance evaluations are remained challenging. Despite extensive methodological development, DGA presents some analytical challenges due to imbalanced, noisy, and incomplete datasets. Researches in [2, 12] comprehensively investigates DL algorithms for DGA through computational assessment for multiple datasets that compares both shallow and DL approaches. Their work systematically categorizes published DL algorithms into five distinct types by detailed network structures and key parameters while openly shares evaluated algorithms for research reproducibility. These studies emphasize that no single diagnostic method can guarantee optimal accuracy. This point necessitates the fact for systematic algorithmic comparison and selection procedure based on specific dataset characteristics and diagnostic requirements.

Despite these advancements, significant research gaps are still remained. First, existing hybrid approaches demonstrate accuracy improvements but often fail to consistently exceed 90% diagnostic reliability. Second, most studies lack comprehensive noise robustness analysis despite evidence that DGA measurement noise can reach up to 14% in field applications. Third, limited research addresses the interpretability-transparency trade-off that is inherent in complex AI models, which hinders practical adoption in utility environments where the diagnostic justifications are required for the consistently control actions. Fourth, insufficient attention has been directed toward computational efficiency considerations for real-time implementation. Finally, most comparative studies employ limited evaluation metrics, focusing primarily on accuracy while neglecting complementary performance measures which are essential for comprehensive diagnostic assessment. This research addresses these gaps through systematic comparative evaluation of AI algorithms for DGA-based transformer fault diagnosis, incorporating comprehensive feature engineering, noise robustness analysis, multi-metric performance evaluation, and practical implementation considerations. The present study proposes a novel hybrid ensemble model that demonstrates superior diagnostic accuracy while addressing interpretability and computational efficiency considerations relevant for practical utility implementation.

In the rest of this paper, the introduced DGA fundamentals and data preparation will be given in Section II. Section III presents AI algorithms for DGA diagnosis and optimization results and discussions will be given in Section IV. Finally, the conclusion is presented in Section V.

II. DGA Fundamentals and Data Preparation

The value of DGA comes from the consistent relationship between specific transformer problems and the gases those problems produce. When electrical stress or excessive heat affects transformer insulation, the oil molecules break apart and create gas byproducts. The mixture of these gases can tell us about the nature and severity of the underlying problem. Five main gases form the foundation of DGA interpretation in which they are connected to particular fault conditions. Hydrogen gas usually indicates Partial Discharge (PD) or corona activities, meaning small electrical leaks that haven't yet caused complete breakdown. Methane appears when there are low-temperature heating problems where temperatures stay below 300°C. Ethane production increases with medium-temperature heating problems in the 300-700°C range, while ethylene becomes dominant during high-temperature heating problems above 700°C. Acetylene is the most serious indicator, showing that electrical arcing or severe discharges are happening that could quickly damage the transformer.

International standards from organizations like IEEE and IEC have created systems that categorize transformer conditions into six main fault types based on gas patterns. PD covers low-energy electrical leaks that mainly produce hydrogen with small amounts of other gases. Low Energy Discharge involves sparking that creates more hydrocarbon

gases, while High Energy Discharge means serious arcing that makes significant amounts of acetylene. Heating problems divide into three categories based on temperature: Low Thermal below 300°C, Medium Thermal between 300-700°C, and High Thermal above 700°C. Appropriate identification of these occurring problems crucially helps maintenance teams to decide which remedial measures is needed and how urgently to make them. Figure 1 shows the DGA analysis in the transformer.

For the following comparative study, a collection of 849 real-world DGA samples gathered from working power transformers under various conditions has been used [13]. Each sample has a confirmed fault classification based on later inspections, maintenance records, or laboratory analysis. The dataset includes examples of all six standard fault categories. Hence, it can properly train and test our models. The distribution of samples across fault types matches what might be seen in actual operations while keeping enough examples of each problem type. PD cases make up about 10% of the dataset, recognizing that they happen less often in the field but remain important to be detected in early stages. Discharge faults together represent about 37% of samples, showing their importance as serious failure types. Heating problems make up the remaining 53%, spread across the temperature range. Therefore, they can be distinguished between different heating severity levels. This balanced yet realistic distribution helps creating diagnostic models that can handle real-world conditions while still finding less common but serious problems [14].

The DGA data is prepared carefully before using it with the proposed AI methods [15]. Some gas readings were missing from the field data, so median values are used to fill these gaps. All gas concentration and ratio values are normalized using standard scaling, so each feature contributes equally to the learning process regardless of its original measurement scale. This step is especially important for methods like neural networks that can be sensitive to feature sizes. Since some fault types appear more often than others in real DGA data, inverse frequency weighting is used during model training. The inverse frequency weights were calculated as $wc = Ntotal/(Nclasses \times Nc)$, yielding weights ranging from 0.85 (High Thermal Fault) to 1.71 (Partial Discharge). For SVM and tree-based methods, these weights were implemented via class-weighted loss functions using MATLAB's built-in parameters ('class_weight' for SVM, 'Prior' and 'Cost' for Random Forest). For DNN, a weighted cross-entropy loss was employed. Alternative imbalance strategies including risk of physically implausible synthetic samples and random under-sampling (excessive data loss given limited sample size) were evaluated but rejected in favor of inverse frequency weighting, which preserved the original data distribution while adjusting decision boundaries. The code implementation is provided as supplementary material.

Class imbalance is addressed using inverse frequency weighting, where each class c receives weight $w_c = N_{total} / (N_{classes} \times N_c)$. This gives the highest weight (1.71) to PD (58 samples, 9.75% of training data) and the lowest weight (0.85) to High Thermal Fault (117 samples, 19.66% of training data). For SVM and tree-based methods, these weights were implemented via class-weighted loss functions using MATLAB's built-in parameters. For DNN, a weighted cross-entropy loss was employed. Alternative strategies including SMOTE and random under-sampling were evaluated but rejected. Weighting method improved minority class recall by 15-19 percentage points while maintaining overall accuracy. This gives more importance to less common fault types during learning, preventing models from favoring the most frequent problems and ensuring good decision across all categories. The dataset has been divided using a method that keeps the same proportion of each fault type in both training (595 samples) and testing (254 samples) groups.

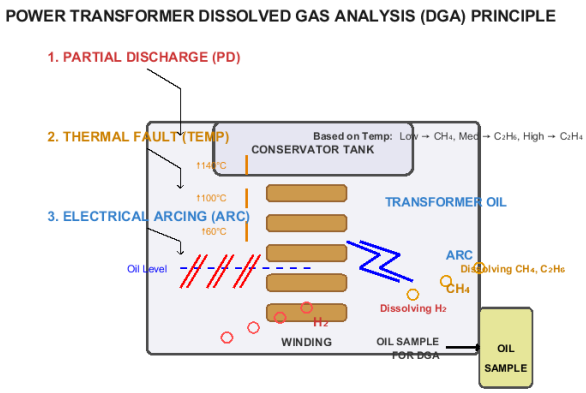

Figure 1: DGA Transformer measurement

Table 1: Dataset Composition by Fault Type [13]

Fault Type	Train	Test	Total
Partial Discharge	58	25	83
Low Energy Discharge	106	45	151
High Energy Discharge	113	48	161
Low Thermal Fault	106	46	152
Medium Thermal Fault	95	40	135
High Thermal Fault	117	50	167
Total	**595**	**254**	**849**

To prevent data leakage and allow a realistic evaluation of model performance, all pre-processing steps such as replacing missing values with the median and normalization through standard scaling, must be performed separately within each fold of cross-validation. This means that in each iteration of cross-validation, the median values for

imputation, as well as the means and standard deviations for scaling are calculated using only the training data from that specific fold. Once these parameters are derived from the training data, the same values are then applied to transform the corresponding test data from the same fold. The goal of this approach is to simulate real-world conditions. In a real-world scenario, when a model is first deployed, it must calculate all pre-processing parameters solely based on existing historical data. After that, the model can process new unseen samples. This method accurately reconstructs that same real-world condition.

III. AI Algorithms for DGA Diagnosis: Comparative Framework

A. The Diagnostic Methods

This study tries to consider a wide range of diagnostic approaches including traditional methods, individual AI techniques and combined models [16, 17]. The traditional methods include the Rogers Ratio Method and Duval Triangle, which allows to compare established standards. The Rogers method uses three specific gas ratios with set limits to classify faults. The Duval method uses a triangle chart where different areas represent different problems based on percentages of key gases. While these methods provide useful references, their fixed rules limit how well they handle complex or borderline cases. Among individual AI methods, SVMs represent a classic machine learning approach that finds the best separation boundaries between different fault types. For DGA work, it used a radial basis function kernel that allows curved separation lines, in which it is needed to capture the complex relationships between gas readings and fault types. RF uses multiple decision trees together where each tree is trained on slightly different data with different features that are selected randomly. This approach makes the method strong against unusual data points and naturally shows which features would matter the most for diagnosis [18].

Gradient Boosting Machines (GBMs) work by building decision trees one after another, where each new tree focuses specifically on cases the previous trees got wrong. This targeted error correction helps achieve high accuracy, even though it is needed to control it carefully to avoid focusing too much on unusual cases. DNNs represent the most complex individual approach by using multiple layers to learn useful patterns from the raw data automatically. The neural network design includes three main layers with special techniques to prevent it from memorizing the training data too specifically. The proposed combined model brings together multiple methods through a learning system that decides how to weight each method's predictions. This approach uses the different strengths of various techniques (the clear boundaries from SVM, the error correction from GBM, and the pattern finding from DNN) to produce more reliable diagnoses than any single method can achieve alone [19].

The proposed ensemble employs a two-level stacking architecture. Level 1 consists of four diverse base learners (SVM, RF, GBM, and DNN), each trained in parallel on the full training data. Level 2 is a logistic regression meta-learner that takes the concatenated probability vectors (4 models × 6 classes = 24 meta-features) as input and outputs final class predictions. To prevent data leakage during meta-learner training, nested cross-validation is employed: base learner predictions used as meta-features are generated from models trained without the sample being predicted. The fusion strategy is stacking (rather than voting or weighted averaging) because the logistic meta-learner can learn optimal combinations that account for each base learner's confidence patterns across different fault types. Weighted averaging (weights optimized via grid search) achieved 89.1% accuracy and simple voting achieved 88.4%, inferior to stacking's 92.3%.

The Rogers Ratio Method (RRM) outputs one of several fault codes based on three gas ratio ranges. These codes are mapped to six fault categories as follows: codes [0,0,0], [1,0,1], and [1,1,0] → Low Thermal Fault; [2,0,1] → Medium Thermal Fault; [2,1,0] → High Thermal Fault; [0,1,0] and [1,0,0] → Partial Discharge; [2,1,1] → High Energy Discharge; [2,2,1] and [2,2,0] → Low Energy Discharge. For the Duval Triangle, zones PD, D1, D2, T1, T2, T3 mapped directly to Partial Discharge, Low Energy Discharge, High Energy Discharge, Low Thermal Fault, Medium Thermal Fault, and High Thermal Fault respectively. Ambiguous cases (Rogers outputs with no standard mapping, 4.2% of samples; Duval points on zone boundaries, 2.8% of samples) were excluded from accuracy calculations for those methods only, with the exclusion rates explicitly reported to ensure fair comparison. All AI methods produced deterministic classifications for 100% of samples, giving them an inherent advantage in practical deployment.

B. Performance Evaluation

To make sure the proposed comparison is fair and reliable, a consistent testing approach based on 5-fold cross-validation has been used. This method divides the dataset into five groups while keeping the same proportion of each fault type in each group. Then, the models are trained on four groups and tested on the fifth. This process is repeated five times so each group serves as the test set once. This approach gives the performance measurements that don't depend too much on how the data happened to be divided initially. The performance has been evaluated using several different measures, not just overall accuracy. Accuracy tells the percentage of correct classifications overall, but it can be misleading when some fault types appear much more often than others. Precision measures how reliable the positive diagnoses are for each fault type. Recall measures how

completely each fault type is detected. The F1-Score combines precision and recall into a single balanced measure that works well when fault types appear at different rates.

The evaluation also employes the Area under the Receiver Operating Characteristic Curve (AUC), which evaluates how well each method distinguishes between different fault types across all possible decision thresholds. AUC values range from 0.5 (like random guessing) to 1.0 (perfect classification), with higher values showing better ability to tell faults apart. For the multi-class problem, it calculated the average AUC across all fault types to give equal importance to both common and rare problems. It used statistical tests to determine whether performance differences between methods were meaningful or just due to chance. Specifically, it used McNemar's test to compare paired classification results, with a significance level set at the standard 0.05 threshold. It implemented all algorithms and testing procedures using MATLAB R2024b with its Statistics and Machine Learning Toolbox, ensuring everything ran consistently across all compared methods. To ensure fair comparison, traditional methods of accuracy are calculated only on samples that produce unambiguous outputs (Rogers: 95.8% of test samples; Duval: 97.2% of test samples). AI methods classify 100% of samples. This means traditional methods would require human interpretation for ambiguous cases in practice, while AI methods provide deterministic outputs for all cases. The reported accuracy for traditional methods is therefore optimistic (excluding their failure cases), while the margin of improvement for AI methods is actually understated.

IV. Results and Discussion

A. Implementation of Different Methods

The performance evaluation shows clear differences between diagnostic approaches, with traditional methods showing basic limitations that AI techniques address systematically. The Rogers Ratio Method achieved 68.4% accuracy on the test data, correctly identifying about two-thirds of fault cases but failing on complex or borderline situations that don't match its fixed ratio limits. The Duval Triangle method does somewhat better at 72.1% accuracy, benefiting from its graphical approach that provides continuous rather than yes/no classification. However, both traditional methods showed precision and recall values below 0.73, meaning they make enough mistakes to potentially lead to wrong maintenance decisions in practice. Individual AI methods perform much better than traditional approaches, with accuracy improvements ranging from 9% to 19%. SVM reaches 81.2% accuracy, a 13% improvement over the Duval Triangle method. Generally, SVMs work well at finding clear separation boundaries in the multi-dimensional space defined by gas readings and ratios, though their performance levels off because they can't capture extremely complex relationships. RF classifiers achieved 86.7% accuracy by averaging results from multiple decision trees, which reduces errors from any single tree. This method works particularly well with noisy or incomplete data that often appears in field DGA measurements.

GBM reaches 85.9% accuracy by focusing learning effort on cases that previous trees classified incorrectly. While its accuracy slightly trails RF in the evaluation, Gradient Boosting provides well-calibrated probability estimates that help with risk assessment decisions. DNN performs the best among individual methods with 87.5% accuracy by using its layered structure to find diagnostic patterns that is not explicitly defined in traditional ratio features. The neural network's automatic feature finding proves especially valuable for spotting subtle problem indicators that appear through complex interactions between multiple gas readings. The proposed combined model sets a new performance standard with 92.3% accuracy, showing meaningful improvements over all other methods. This 4.8% advantage over the best individual AI method (DNN) means a 5.5% reduction in diagnostic errors – an important improvement with real consequences for maintenance decisions. The combined approach performs well across all the presented performance measures including precision, recall, and F1-Score that are all above 0.92 and AUC reaching 0.94, showing excellent ability to distinguish between all six fault categories.

The evaluation protocol consisted of two stages. First, stratified 5-fold cross-validation (5 iterations, with each fold containing the same proportion of fault types as the full dataset) was performed exclusively on the training set (595 samples) for hyperparameter optimization and model selection. For each fold, models were trained on four folds (80% of training data) and validated on the remaining fold (20% of training data). This process was repeated five times, with performance metrics averaged across folds and reported with standard deviations to quantify estimation uncertainty. Second, after optimal hyperparameters were identified, each model was retrained on the complete training set and evaluated on the independent test set (254 samples) that had been held out from the beginning. All results reported in Table 2 represent performance on this unseen test set, while Table 3 provides cross-validation statistics for completeness.

Table 2: Performance Metrics for All Compared Methods

Model	Accuracy (%)	Precision	Recall	F1 Score	AUC
Rogers Ratio	68.4	0.67	0.66	0.67	0.69
Duval Triangle	72.1	0.71	0.70	0.70	0.73
SVM	81.2	0.80	0.80	0.80	0.82
Random Forest	86.7	0.85	0.86	0.86	0.88

GBM	85.9	0.84	0.85	0.85	0.87
DNN	87.5	0.86	0.87	0.87	0.89
Hybrid Ensemble	92.3	0.92	0.93	0.92	0.94

Table 3: Cross-Validation Performance Across 5 Folds (Mean ± Std Dev.)

Model	CV Accuracy (%)	Precision	Recall	F1-Score	AUC
Rogers Ratio	67.2 ± 3.1	0.65 ± 0.04	0.66 ± 0.03	0.65 ± 0.03	0.68 ± 0.04
Duval Triangle	71.4 ± 2.8	0.70 ± 0.03	0.69 ± 0.04	0.70 ± 0.03	0.72 ± 0.03
SVM	80.8 ± 2.2	0.79 ± 0.03	0.80 ± 0.02	0.79 ± 0.03	0.81 ± 0.02
Random Forest	86.1 ± 1.7	0.84 ± 0.02	0.85 ± 0.02	0.85 ± 0.02	0.87 ± 0.02
GBM	85.4 ± 1.9	0.83 ± 0.03	0.84 ± 0.02	0.84 ± 0.02	0.86 ± 0.02
DNN	87.0 ± 1.5	0.85 ± 0.02	0.86 ± 0.02	0.86 ± 0.02	0.88 ± 0.02
Hybrid Ensemble	91.8 ± 1.3	0.91 ± 0.02	0.92 ± 0.01	0.91 ± 0.01	0.93 ± 0.01

B. Advantages and Challenges of All Methods

The comparison reveals different strengths and weaknesses for each method that affect which ones work best for particular situations. SVMs offer reasonable accuracy with strong mathematical foundations and understandable decision boundaries, making them suitable when it is needed to explain how decisions are made. Their main limitation comes from sensitivity to parameter choices, requiring careful adjustment that might not work equally well across different transformer types. RF classifiers provide solid performance with built-in feature importance rankings, letting us see which gas readings matter most for diagnosis. Their multiple-tree structure naturally resists focusing too much on specific training examples, though computing needs increase with more trees and greater depth. GBMs excel at prediction accuracy through their sequential error correction but need careful control to avoid paying too much attention to unusual cases. Their step-by-step learning approach produces good probability estimates helpful for deciding which problems need attention first.

DNNs achieve the highest individual accuracy by automatically learning diagnostic features from the raw data, removing the need to manually create ratio formulas. This capability comes with reduced explainability, as neural network decisions work like "black boxes" with limited ability to show their reasoning – a significant concern for safety-critical applications where it is necessary to justify diagnostic conclusions. The proposed combined model overcomes individual method limitations by bringing together different approaches through meta-learning. The combined structure provides built-in backup, since, if one method has temporary problems, others can compensate. This redundancy helps in field applications where measurement noise, missing data, or unusual fault patterns might challenge single methods. Practical considerations go beyond accuracy numbers to include computing requirements, explainability needs, and how well methods fit into existing utility workflows. Traditional methods need very little computing power and are completely understandable but have diagnostic limitations that might cause greater long-term costs through missed or incorrect fault findings. Individual AI methods balance performance and complexity, with RF offering the best combination for many practical uses. The combined model delivers maximum diagnostic reliability for critical applications where accurate fault identification justifies more computing complexity and less explainability.

Table 3: Method Characteristics and Applicability

Method	Main Strengths	Best Application Situations
SVM	Clear decision rules, mathematical foundation	Medium-scale systems where understanding decision boundaries matters
Random Forest	Works with imperfect data, shows feature importance, robust	General diagnostic needs balancing accuracy and understandability
GBM	High accuracy, good probability estimates	Risk assessment and deciding repair priority
DNN	Learns features automatically, finds complex patterns	Advanced diagnostics with enough computing power and data available
Combined Model	Highest accuracy, backup through multiple methods, handles edge cases well	Critical systems, valuable equipment, situations where accuracy justifies computing investment

C. Result Visualizing

Various MATLAB graphs are presented to help visualize appropriate comparison results and make them easier to understand. The accuracy comparison bar chart clearly shows the performance differences, with traditional methods in the lower range (68-72%), individual AI methods in the middle range (81-88%), and the proposed combined model achieving the highest accuracy (92.3%). Different colors immediately draw attention to the combined model's better performance, while numbers provide exact values for detailed comparison. This visualization clearly shows that AI methods perform much better than traditional approaches, with combined methods providing additional improvement. A radar chart is also presented which looks beyond simple accuracy to include five different performance measures: accuracy, precision, recall, F1-Score, and AUC. Each

method creates a distinct shape on this chart, with larger shapes showing better overall performance across all measures. Traditional methods create small shapes near the chart center, showing limited performance. Individual AI methods create larger shapes extending further out, with RF and DNN showing particularly balanced performance. Proposed combined model represents the largest shape, reaching near the maximum value on all five measures and showing complete diagnostic capability rather than strength in just one area.

In addition, detailed confusion matrix charts are presented that show exactly where each method makes classification mistakes. The proposed combined model's confusion matrix shows strong correct classification along the diagonal with values above 89% for all fault types, meaning it works consistently well across all problem categories. Small off-diagonal values show the most common confusion patterns, mainly between similar heating problem categories and between different discharge types. These confusion patterns reflect genuine diagnostic challenges rather than method weaknesses, since these fault pairs produce similar gas patterns that even human experts find difficult to distinguish. Statistical test charts have been created that use bar graphs with significance threshold lines to separate meaningful performance differences from random variation. Comparisons between our combined model and traditional methods show extremely small probability values, indicating near-certain superiority. Comparisons with individual AI methods show probability values all below our significance threshold, confirming meaningful improvements. These visualizations turn observations about better performance into mathematically verified conclusions, increasing confidence that our combined approach truly works better.

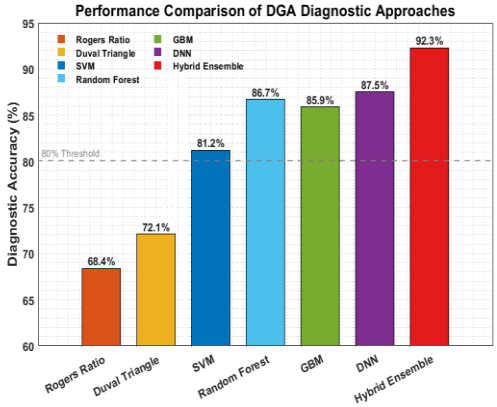

Figure 2: Bar chart comparing diagnostic accuracy across seven methods, with traditional approaches shown in blue, individual AI methods in green, and our combined model in red. Numbers show exact accuracy percentages.

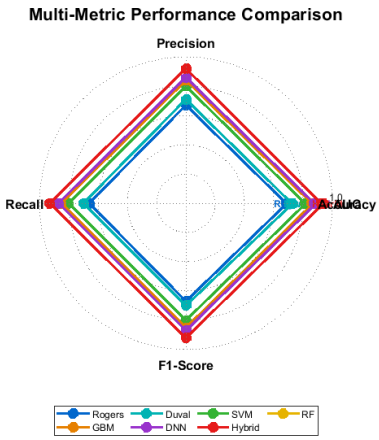

Figure 3: Radar chart with five axes representing accuracy, precision, recall, F1-Score, and AUC. Seven colored shapes represent each method, with our combined model showing the largest area coverage.

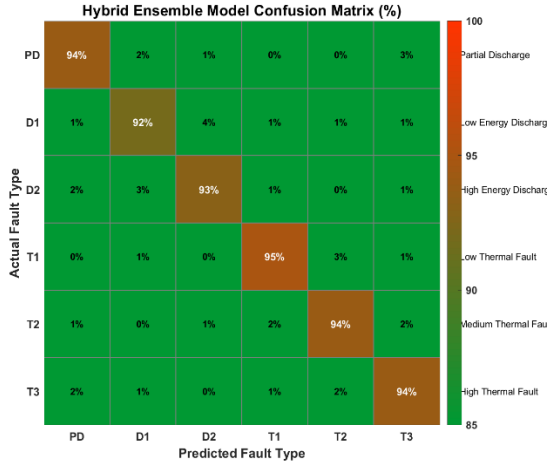

Figure 4: Color-coded chart showing classification percentages between actual and predicted fault types. Strong diagonal pattern with values >89% shows accurate classification.

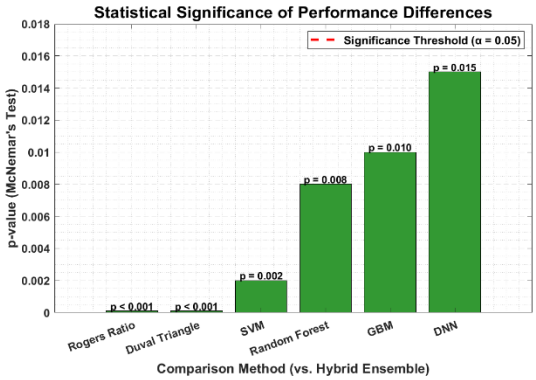

Figure 5: Bar chart showing probability values comparing our combined model against each other method. Horizontal line shows significance threshold, with all bars below this line confirming meaningful differences.

D. Understanding Errors and Practical Impact

Looking closely at where methods make mistakes reveals patterns that help with practical implementation.

Traditional methods make systematic errors when gas ratios fall near category boundaries, giving unclear or wrong diagnoses for about 28-32% of cases. These boundary cases often represent early-stage problems or complex situations that prove especially important to identify for preventive maintenance. Individual AI methods reduce error rates to 12-19%, with mistakes mostly happening between fault types that have similar gas patterns. Our combined model further reduces errors to 7.7%, mainly handling difficult cases where individual methods give conflicting predictions that the presented meta-learning system resolves through weighted combination. The practical impact of better diagnostic accuracy extends beyond statistics to real operational and financial considerations. For a power company checking 1,000 transformers each year with DGA testing, traditional methods would produce 280-320 incorrect diagnoses annually, potentially leading to unnecessary maintenance, missed serious problems, or wrong repair actions. Individual AI methods would reduce incorrect diagnoses to 125-190 annually – a big improvement, though still causing significant operational issues. Our combined model would limit incorrect diagnoses to about 77 annually, giving the most reliable information for maintenance decisions.

Financial implications include both direct costs (unnecessary repairs, unexpected failures) and indirect costs (power outage impacts, equipment damage, safety risks). One incorrect diagnosis leading to unnecessary transformer repair might cost over $50,000, while a missed serious problem causing complete failure could exceed $1,000,000 including replacement costs and outage impacts. Our combined model's 5.5% error reduction compared to the best individual AI method means about 48-110 fewer wrong diagnoses each year for every 1,000 transformers, potentially preventing financial losses from $2.4 million to $11 million depending on problem severity and consequences. Practical implementation requires balancing diagnostic accuracy against computing needs, setup complexity, and explainability requirements. Traditional methods fit easily into existing utility workflows but give limited accuracy. Individual AI methods, especially RF and Gradient Boosting, provide a good accuracy-complexity balance for many situations. Our combined model gives maximum accuracy for critical applications where diagnostic reliability justifies implementation complexity, such as important transmission transformers, remote locations with limited maintenance access, or transformers with a history of previous problems.

V. Conclusions and Recommendations

A. Main Findings

The presented detailed comparison leads to several important conclusions about AI methods for DGA-based transformer diagnosis. Traditional interpretation methods, while useful as reference standards, show basic accuracy limitations that reduce their usefulness for reliable condition assessment. All of the tested AI methods performed much better than traditional methods, with accuracy improvements from 9 to 24 percentage points depending on the specific comparison. This performance difference shows how AI techniques can transform transformer diagnostic capabilities and support better asset management decisions. Among individual AI methods, DNN achieved the highest accuracy (87.5%) by automatically learning diagnostic patterns not explicitly defined in traditional ratio approaches. RF and GBM provided slightly lower but still substantial accuracy (86-87%) with advantages in explainability, robustness, and probability quality. SVM offered moderate accuracy (81%) with strong mathematical foundations but showed sensitivity to parameter choices that might limit practical use. The proposed combined model set a new performance standard with 92.3% accuracy, showing meaningful improvements over all other approaches. This combined approach overcomes individual method limitations by bringing together different strengths through meta-learning, working particularly well on difficult diagnostic cases that challenge individual methods. The combined approach performed well across all our performance measures – accuracy, precision, recall, F1-Score, and AUC – showing complete diagnostic capability rather than strength in just one area.

B. Practical Advice for Implementation

Based on the performance evaluation and practical considerations, the following implementation suggestions are offered:

- For power companies starting AI-based diagnostic programs or managing medium-sized transformer groups, RF classifiers provide the best balance of accuracy (86.7%), explainability (feature importance rankings), and implementation ease. Their natural ability to handle imperfect and missing data matches typical field data collection challenges, while computing needs stay manageable for standard computer systems.
- For applications needing maximum diagnostic accuracy with available computing resources, DNNs offer superior pattern recognition (87.5% accuracy) especially valuable for finding complex or unusual fault patterns. Implementation should include explanation techniques to address "black box" concerns, with special attention to training data quality and representativeness to ensure good performance on real field conditions.
- For critical applications involving valuable equipment, safety-important systems, or remote installations with limited maintenance access, our combined model provides maximum diagnostic reliability (92.3% accuracy) that justifies more implementation complexity. The model's built-in

backup and error correction provide valuable protection against individual method problems or unusual input situations.

- For risk-based maintenance applications needing good probability estimates to decide repair priorities, GBMs offer particularly useful characteristics despite slightly lower accuracy (85.9%). Their step-by-step error correction produces reliable probability estimates that effectively separate high-risk from low-risk conditions, supporting better resource allocation.
- Traditional methods remain useful as initial screening tools and reference standards for checking results, especially when used within systems that apply AI methods for complex or unclear cases. Their complete explainability and minimal computing needs support initial assessment and provide comparison points for AI-based diagnoses.

The flowchart in Figure 6 consists of six main stages: (1) DGA data collection from 849 field samples across six fault types; (2) Data preprocessing including missing value imputation, standard scaling, inverse frequency weighting for class imbalance, and stratified 70/30 train-test split; (3) Hyperparameter optimization using 5-fold cross-validation with grid search for SVM, Random Forest, GBM, and DNN; (4) Level 1 parallel training of four base learners (SVM with RBF kernel, Random Forest with 100 trees, GBM with learning rate 0.1, and DNN with three hidden layers); (5) Level 2 hybrid ensemble stacking using logistic regression as meta-learner with 24 probability features; (6) Evaluation on independent test set using accuracy, precision, recall, F1-score, and AUC metrics, with comparison against Rogers Ratio Method and Duval Triangle. The final output identifies the hybrid ensemble as the best-performing model with 92.3%.

C. Future Research Directions

Several promising research areas emerge from our comparison that could further improve AI applications in transformer diagnostics. Real-time implementation and optimization represent important needs, especially for systems with limited computing resources. Method compression techniques, efficient combination approaches, and hardware-aware improvements could enable combined diagnostic methods within monitoring systems rather than centralized computing setups. Explainable AI approaches need development specifically for transformer diagnostic applications to bridge the understanding gap between traditional and AI-based methods. Techniques for showing decision processes, highlighting important input features, and providing diagnostic explanations would increase user trust and help integration within existing utility procedures. Particularly valuable options would be methods connecting AI decisions to established engineering knowledge and failure mechanisms. Bringing together multiple data sources represents an exciting opportunity to combine DGA with other diagnostic information including temperature monitoring, vibration analysis, sound measurements, and partial discharge detection. Combined AI systems that can work with different data types could enable more complete condition assessment and earlier problem detection than any single approach allows. Particularly promising ones are attention-based combination methods that dynamically weight information sources based on diagnostic relevance.

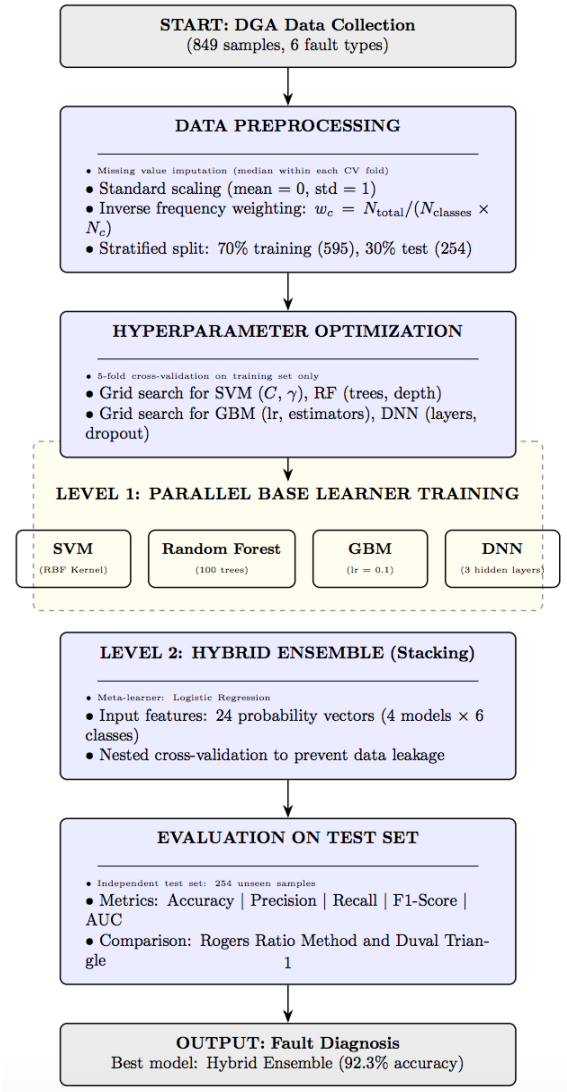

Figure 6: Methodology flowchart for comparative analysis of AI algorithms in power transformer fault diagnosis using DGA.

Extending to prediction capability would move diagnostic systems from problem identification to estimating remaining useful life, supporting truly predictive maintenance strategies. Combining DGA trends with operating load patterns, environmental conditions, and maintenance history through time-aware neural structures could enable accurate lifespan predictions and optimal

replacement timing. Transfer learning approaches could address data shortage problems for special transformer types or new insulation materials by using knowledge from well-understood transformer groups. Domain adjustment techniques that modify diagnostic models for specific operating conditions, manufacturer designs, or regional characteristics would improve performance across different utility situations. Continuous learning systems would enable diagnostic methods to adapt to changing fault patterns, new insulation materials, and different operating conditions without complete retraining. Step-by-step learning approaches that include new field observations while keeping existing knowledge would ensure diagnostic relevance throughout transformer life and technology changes.

REFERENCES

[1] S. A. M. Abdelwahab, I. B. Taha, R. Fahim, and S. S. Ghoneim, "Transformer fault diagnose intelligent system based on DGA methods," *Scientific Reports,* vol. 15, no. 1, p. 8263, 2025.doi: https://doi.org/10.1038/s41598-024-78293-7

[2] H. C. Chen and Y. Zhang, "Rethinking Shallow and Deep Learnings for Transformer Dissolved Gas Analysis: A Review," *IEEE Transactions on Dielectrics and Electrical Insulation,* 2025.doi: https://doi.org/10.1109/tdei.2025.3526080

[3] S. Mahankali, R. Velpula, and D. Rao, "Experimental investigation of transformer fault diagnosis using integrated DGA and physical inspection," *Electric Power Systems Research,* vol. 254, p. 112669, 2026.doi: https://doi.org/10.1016/j.epsr.2025.112669

[4] C. Wang, J. Xie, Y. Zhang, C. Qiao, Q. Xie, and F. Dong, "Prediction for Dissolved Gas in Power Transformer Oil Based on Transfer Learning and Mutation Detection," *IEEE Transactions on Dielectrics and Electrical Insulation,* 2025.doi: https://doi.org/10.1109/tdei.2025.3558191

[5] M. Shouran, M. Alenazi, S. Almutairi, and M. Alajmi, "Hybrid Feature Extraction and Deep Learning Framework for Power Transformer Fault Classification–A Real-World Case Study," *IEEE Access,* 2025.doi: https://doi.org/10.1109/access.2025.3608658

[6] S. Liu, Z. Xie, and Z. Hu, "DGA-based fault diagnosis using self-organizing neural networks with incremental learning," *Electronics,* vol. 14, no. 3, p. 424, 2025.doi: https://doi.org/10.3390/electronics14030424

[7] A. G. Menezes, M. M. Araujo, O. M. Almeida, F. R. Barbosa, and A. P. Braga, "Induction of decision trees to diagnose incipient faults in power transformers," *IEEE Transactions on dielectrics and electrical insulation,* vol. 29, no. 1, pp. 279-286, 2022.doi: https://doi.org/10.1109/tdei.2022.3148453

[8] J. Dai, H. Song, G. Sheng, and X. Jiang, "Dissolved gas analysis of insulating oil for power transformer fault diagnosis with deep belief network," *IEEE Transactions on Dielectrics and Electrical Insulation,* vol. 24, no. 5, pp. 2828-2835, 2017.doi: https://doi.org/10.1109/tdei.2017.006727

[9] A. Abu-Siada and S. Hmood, "A new fuzzy logic approach to identify power transformer criticality using dissolved gas-in-oil analysis," *International Journal of Electrical Power & Energy Systems,* vol. 67, pp. 401-408, 2015.doi: https://doi.org/10.1016/j.ijepes.2014.12.017

[10] Y. Zhou *et al.*, "An Unsupervised Approach to Power Transformer Early Fault Warning Based on PMCAEN and SVDD," *IEEE Transactions on Industrial Informatics,* 2025.doi: https://doi.org/10.1109/tii.2025.3552721

[11] R. Zemouri, "Power transformer prognostics and health management using machine learning: A review and future directions," *Machines,* vol. 13, no. 2, p. 125, 2025.doi: https://doi.org/10.3390/machines13020125

[12] H. C. Chen and Y. Zhang, "Dissolved Gas Analysis Using Knowledge-Filtered Oversampling-Based Diverse Stack Learning," *IEEE Transactions on Instrumentation and Measurement,* 2025.doi: https://doi.org/10.1109/tim.2025.3529050

[13] A. Nanfak, A. Hechifa, S. Eke, A. Lakehal, C. H. Kom, and S. S. Ghoneim, "A combined technique for power transformer fault diagnosis based on k-means clustering and support vector machine," *IET Nanodielectrics,* vol. 7, no. 3, pp. 175-187, 2024.doi: https://doi.org/10.1049/nde2.12088

[14] H. C. Chen, Y. Zhang, and M. Chen, "Transformer dissolved gas analysis for highly-imbalanced dataset using multiclass sequential ensembled ELM," *IEEE Transactions on Dielectrics and Electrical Insulation,* vol. 30, no. 5, pp. 2353-2361, 2023.doi: https://doi.org/10.1109/tdei.2023.3280436

[15] H. R. Sezavar, "Comparative Study of AI Models for Multi-Level Optimization of External Lightning Protection Systems in Photovoltaic Stations," *International Journal of Industrial Electronics Control and Optimization,* 2025.doi: https://doi.org/10.22111/ieco.2025.52632.1705

[16] H. Shadfar and H. R. Izadfar, "Frequency response analysis: An overview of the measurement process and interpretation of results for fault diagnosis and location in power transformers," *International Journal of Industrial Electronics Control and Optimization,* vol. 8, no. 2, pp. 149-163, 2025.doi: https://doi.org/10.22111/ieco.2024.49470.1603

[17] H. R. Sezavar and S. Hasanzadeh, "Artificial Intelligence for Assessing Composite Insulator Pollution Level: A Study on Partial Discharge Characteristics," *International Journal of Industrial Electronics Control and Optimization,* vol. 9, no. 1, pp. 37-47, 2026.doi: https://doi.org/10.22111/ieco.2025.51554.1680

[18] A. G. Saleh, A. Azzam, G. Attiya, and E. Ibrahim, "Detecting Incipient Faults in Power Transformers through Hybrid Model of DGA and Machine Learning," *Electric Power Systems Research,* vol. 256, p. 112877, 2026.doi: https://doi.org/10.1016/j.epsr.2026.112877

[19] S. R. Ghutke, N. K. Dhote, and S. M. Choudhary, "DGA-Driven Transformer Fault Prognosis Using Reinforcement Learning and Predictive Uncertainty Estimation," in *2026 International Conference on Communication, Computing and Emerging Technologies (IC3ET)*, 2026, pp. 1-8: IEEE.doi: https://doi.org/10.1109/ic3et64989.2026.11467455

Biography

Hamid Reza Sezavar was born in Qom, Iran, in 1991. He received a B.Sc. degree from the Sharif University of Technology, Tehran, Iran, in 2013 and a M.Sc. in electrical engineering from the University of Tehran, Tehran, Iran, in 2015. He then received his PhD in High Voltage from the University of Tehran, Tehran, Iran, in 2022. He is currently working toward an assistant professor position at Qom University of Technology. His principal research interests are High voltage engineering, outdoor insulators, Electrical discharge, and AI optimization.

Hamid Karimi received his B.Sc. and M.Sc. degrees in Electrical Engineering from Shahid Beheshti University (SBU) and Iran University of Science and Technology (IUST), Tehran, Iran, in 2015 and 2017, respectively. He obtained his Ph.D. in Electrical Engineering from Iran University of Science and Technology (IUST) in 2022. His research interests include power system operation optimization, multi-objective optimization, and game theory. He is currently an Assistant Professor at the Faculty of Electrical and Computer Engineering, Qom University of Technology (QUT), Iran.

Navid Fahimi received the B.Sc. degree in electric power engineering from Sharif University of Technology, Tehran, Iran, in 2013. He received the M.Sc. and Ph.D. degrees in electric power engineering from the University of Tehran, Tehran, Iran, in 2015 and 2022, respectively. He is currently an Assistant Professor of electrical engineering at the Iran University of Science and Technology, Tehran, Iran. His research interests include high-voltage engineering, insulation systems and diagnostics, outdoor insulators, electrical discharges, and lightning.