

Article

Generalization-Enhanced State Assessment of Railway Power Transformers Using Feature-Guided Stacking Learning

Yuanfang Huang ¹, Zhanhong Huang ^{2,*}  and Junbin Chen ³ 

¹ School of Rail Transit, Guangdong Communication Polytechnic, Guangzhou 510630, China; hyfang003@163.com

² School of Electric Power, South China University of Technology, Guangzhou 510641, China

³ School of Engineering, College of Engineering, Shantou University, Shantou 515000, China; junbinchen@stu.edu.cn

* Correspondence: 202211082838@mail.scut.edu.cn

Abstract

Reliable state assessment of railway traction power transformers is challenged by heterogeneous operating environments, measurement disturbances, coupled gas-generation mechanisms, and uneven fault-sample distributions. Conventional dissolved gas analysis (DGA) ratio rules and single-model classifiers often show insufficient generalization when rare faults and boundary-ambiguous operating states are encountered. To address this issue, this paper proposes a feature-guided stacking framework for state assessment of oil-immersed railway power transformers. First, a DGA-oriented fusion-feature representation is established by combining raw gas concentrations, gas-ratio descriptors, and an aggregated dissolved-gas analysis factor. Second, DBSCAN-assisted sample structuring is introduced to identify density patterns, sparse rare fault regions, and boundary samples, thereby improving the organization of imbalanced monitoring records. Third, a monitoring-feature-embedded stacking model is developed in which heterogeneous base learners are adaptively weighted according to feature-reliability information and integrated through a cross-validated meta-learner. This synthetic-data-based validation provides a controlled and reproducible proof-of-concept. Therefore, the reported results should be interpreted as evidence of methodological feasibility. Under the default synthetic setting, the proposed feature-guided stacking (FE-stacking) method achieves an accuracy of 99.70% and a macro-F1 of 99.55%. Under the severe minority-retention setting in which only 25% of low-energy discharge (LD) and low-temperature overheating (LT) training samples are preserved, it obtains an accuracy of 99.62%, a macro-F1 of 99.40%, and an LT recall of 96.61%, slightly surpassing random forest (RF) and outperforming Original Stacking in rare fault robustness. These results indicate that feature-guided ensemble learning can improve the generalization stability of DGA-based transformer state assessment under imbalanced and boundary-ambiguous conditions. From a practical perspective, the proposed framework can serve as a decision-support module for transformer condition screening, maintenance prioritization, and alarm verification in railway traction power-supply systems.



Academic Editor: Leijiao Ge

Received: 4 June 2026

Revised: 8 July 2026

Accepted: 15 July 2026

Published: 20 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: transformer state assessment; dissolved gas analysis; railway traction power transformer; generalization enhancement; feature-guided stacking; sample imbalance

1. Introduction

Railway transportation systems are developing toward denser interconnection, higher operating continuity, and stricter safety requirements. Under this trend, traction power

transformers are exposed to diverse environmental disturbances and operating stresses, including regional weather variation, installation heterogeneity, communication instability, load-dependent thermal effects, and transformer-to-transformer differences. As a key component of the traction power supply system, the oil-immersed traction transformer directly affects electrical safety and supply continuity. Incipient abnormalities, such as weak discharge or mild overheating, may not immediately trigger protection actions, but they can gradually degrade insulation and increase the risk of service interruption. Therefore, accurate transformer state assessment is essential for condition-based maintenance, asset-management decision making, and intelligent railway power-system operation [1]. Existing studies have investigated transformer maintenance modeling, asset criticality assessment, communication-network equipment state prediction, winding deformation monitoring, and transformer vibration/noise analysis from different perspectives [2,3]. Recent work on transformer insulation condition assessment has further shown that transformer monitoring is moving toward multi-source sensing, advanced feature representation, and intelligent decision support [4,5].

Dissolved gas analysis (DGA) is one of the most widely used techniques for oil-immersed transformer monitoring because it reflects the decomposition behavior of insulating oil and solid insulation under electrical and thermal faults [6–8]. Conventional DGA interpretation methods, including ratio-based rules, threshold-oriented criteria, fuzzy logic, ANFIS models, and standard gas interpretation guidelines, have been widely used because of their simplicity and engineering interpretability [9]. However, these methods usually rely on fixed diagnostic boundaries and may become unreliable when gas signatures are affected by mixed fault mechanisms, measurement disturbances, incomplete records, or transformer-specific operating histories [10]. Recent review and multi-method studies have confirmed that DGA-based transformer diagnosis is increasingly moving from conventional rule-based interpretation toward machine-learning-assisted decision models, but data quality, class imbalance, and model generalization remain important challenges [11].

To improve nonlinear state recognition from DGA measurements, machine-learning-based methods have been introduced into transformer fault diagnosis. Existing studies have investigated deep neural networks, stacked sparse auto-encoders, broad learning systems, ANN-based diagnostic models [12], KPCA-WPA-SVM, complementary SaE-ELM with arctangent transform, and wavelet-network-based diagnosis [13,14]. More recent studies have further introduced feature extraction, feature selection, optimized ensemble learning, hybrid kernel learning, Adaboost, and optimized machine-learning algorithms to improve the representation and classification capability of DGA features [15]. Although these methods improve feature-learning and diagnostic accuracy compared with traditional expert rules [16], many single classifiers and general data-driven models remain sensitive to sample distribution, class imbalance, and boundary overlap. This limitation is particularly important for railway power transformers because minority states, such as low-energy discharge and low-temperature overheating, are often less represented than normal or severe fault conditions [17,18]. When these minority samples are sparse or close to neighboring classes in the feature space, a single model may generalize poorly to unseen operating scenarios [19,20]. Thus, under the framework of integrated feature description, the ability to evaluate the distinguishability of the operating status of railway transformers remains a challenging problem. Furthermore, it is necessary to enhance the robustness of the performance of deep learning algorithms under conditions of limited samples and sparsity.

Ensemble learning provides a promising way to enhance diagnostic robustness by combining heterogeneous learners with complementary decision patterns. Stacking learning is especially suitable for this purpose because it constructs a second-level meta-learner

from the out-of-fold predictions of multiple base learners [21]. Feature-engineering and ensemble-learning methods can improve nonlinear state recognition by constructing high-dimensional gas-ratio descriptors and combining complementary classifiers. Imbalance-learning methods, such as SMOTE-based sampling, cost-sensitive learning, and active learning, have also been investigated to alleviate the adverse effect of sparse fault samples. Nevertheless, conventional stacking still has two limitations in imbalanced DGA-based assessment. First, the meta-learner may inherit the bias of base learners trained on majority-dominated samples. Second, the conventional fusion strategy usually treats the reliability of each base learner as fixed, whereas different learners may perform better in different feature regions. In addition, density-based sample-structuring methods such as density-based spatial clustering of applications with noise (DBSCAN) have been shown to be useful for identifying irregular data structures and clustering patterns in power-system applications [22], but their potential for revealing sparse rare fault regions and boundary-ambiguous DGA samples has not been sufficiently explored in transformer state assessment. A fixed or globally estimated fusion strategy may cause the meta-learner to inherit majority-class bias or underrepresent locally ambiguous DGA samples. Although imbalance-handling techniques have been used in transformer fault diagnosis, many of them primarily operate at the data-distribution level by modifying class frequencies or sample weights. These methods are useful for improving minority-class exposure, but they do not necessarily reveal whether a minority sample is located in a dense fault region, a sparse operating region, or a boundary-adjacent transition area. Moreover, they usually do not adapt the contribution of different base learners according to the local DGA feature structure. This leaves a research gap for integrating sample-structure awareness with feature-dependent ensemble fusion.

Different from these existing adaptive ensemble-learning approaches, this paper develops a feature-guided stacking framework for railway traction transformer state assessment. The comparison of each algorithm is shown in Table 1. The proposed method integrates DGA-oriented fusion-feature representation, DBSCAN-assisted sample structuring, and monitoring-feature-embedded adaptive weighting into a unified ensemble model. Correspondingly, the working hypothesis of this study is that coupling DGA-specific feature reliability with density-structure information can improve the robustness of ensemble fusion under imbalanced DGA conditions. Therefore, the novelty of this study does not lie merely in applying a stacking classifier, but in constructing an adaptive fusion mechanism that explicitly connects base-learner reliability with DGA feature reliability, sample-density information, rare fault regions, and boundary ambiguity. This design is expected to improve the generalization stability of transformer state assessment under imbalanced and heterogeneous monitoring conditions.

Motivated by these gaps, this paper proposes a feature-guided stacking framework for adaptive state assessment of railway power transformers. The proposed method integrates DGA-oriented fusion-feature characterization, DBSCAN-assisted sample structuring, and monitoring-feature-embedded adaptive weighting within a stacking architecture. Different from conventional DGA interpretation methods, single machine-learning classifiers, and general ensemble-learning approaches, the proposed framework aims to improve the representation of minority and boundary-ambiguous samples and to adaptively adjust the contribution of heterogeneous base learners according to DGA-specific feature-reliability information. In this way, the ensemble model is better aligned with the structure of DGA monitoring data and becomes more robust under imbalanced synthetic test conditions.

Table 1. Comparison of existing DGA-based transformer state assessment methods and the proposed framework.

Method Category	Main Idea	Main Limitations	Difference of This Paper
Conventional DGA interpretation rules	Use predefined gas ratios, thresholds, or graphical regions to identify transformer fault types	Fixed decision boundaries; limited adaptability to mixed faults, measurement noise, and transformer-specific operating conditions	Uses data-driven feature representation while retaining DGA-oriented gas-ratio information
Single machine-learning classifiers	Learn nonlinear mapping from DGA features to transformer states using SVM, RF, neural networks, or gradient boosting models	Sensitive to sample distribution, class imbalance, sparse minority faults, and boundary-overlapping samples	Introduces sample-structure awareness and minority-region characterization before supervised learning
Deep-learning and hybrid feature-extraction methods	Improve diagnostic accuracy through more sophisticated feature extraction and nonlinear representation learning	Usually require sufficient training samples; interpretability and robustness under imbalanced monitoring data remain challenging	Focuses on generalization stability under imbalanced and boundary-ambiguous DGA conditions
Imbalance-learning methods	Rebalance minority samples using oversampling, sample weighting, cost-sensitive learning, or active learning	Mainly address class-frequency imbalance; local density structure, boundary ambiguity, and feature-dependent learner reliability are usually not jointly modeled	Integrates DBSCAN-derived sample structure and DGA feature reliability into adaptive ensemble fusion
Conventional ensemble-learning methods	Combine multiple learners through bagging, boosting, RF, or XGBoost-based fusion to improve prediction stability	Learner weights are mainly determined by global training or validation performance, with limited use of DGA-specific monitoring structure	Embeds DGA feature reliability and sample-density information into adaptive ensemble fusion
Conventional stacking models	Integrate heterogeneous base learners through cross-validated meta-learning	The meta-learner may inherit bias from majority-dominated base learners; learner reliability is not feature-dependent	Develops monitoring-feature-embedded adaptive weighting within the stacking architecture
Proposed FE-Stacking framework	Combines DGA fusion features, DBSCAN-assisted sample structuring, and feature-guided adaptive stacking	Current validation is based on synthetic data and requires further field-data verification	Provides a reproducible proof-of-concept for improving generalization under imbalanced and boundary-ambiguous DGA conditions

The main contributions of this study are summarized as follows.

1. A generalization-oriented DGA feature representation is constructed by combining raw gas concentrations, gas-ratio descriptors, and a statistically estimated DGAF descriptor. This representation is designed to improve the separability and strengthens the state sensitivity of transformer operating states under heterogeneous gas-generation patterns
2. A DBSCAN-assisted sample structuring mechanism is introduced to characterize density patterns, sparse rare fault regions, and boundary-ambiguous samples. This mechanism supports sample reorganization and addresses the minority-representation issue on subsequent supervised learning.
3. A feature-guided stacking framework is proposed to overcome the fixed-fusion limitation of conventional stacking. The reliability of heterogeneous base learners is adaptively adjusted according to monitoring features and validation performance,

thereby improving the generalization stability of transformer state assessment under severe minority-sample compression.

The remainder of this paper is organized as follows. Section 2 introduces the DGA-based state characterization and DBSCAN-assisted sample structuring strategy. Section 3 presents the feature-guided stacking framework and its adaptive learning procedure. Section 4 reports simulation-based validation on a synthetic DGA dataset. Section 5 concludes the paper and discusses future work.

2. Related Foundations

2.1. DGA-Based Characterization of Oil-Immersed Transformer States

During transformer operation, electrical, thermal, and insulation-related stresses promote the decomposition of insulating oil and solid insulation, thereby generating characteristic dissolved gases. Following standard DGA interpretation principles and previous transformer diagnosis studies, key gases are considered in this study: H_2 , CH_4 , C_2H_6 , C_2H_4 , C_2H_2 , and CO . These gases are selected because they are strongly associated with the main electrical and thermal degradation mechanisms of oil-immersed transformers. H_2 is sensitive to discharge-related phenomena, especially partial discharge and low-energy discharge. CH_4 is related to low-temperature oil decomposition and may also appear under discharge-related oil degradation. C_2H_6 is mainly associated with low-temperature overheating, whereas C_2H_4 is more indicative of higher-temperature overheating. C_2H_2 is a key gas for arcing or high-energy discharge. CO is mainly generated from the thermal degradation of cellulose insulation and is therefore useful for identifying overheating involving solid insulation [7]

$$DGAF_i = \sum_{g \in G} \omega_g \check{c}_{i,g} + \sum_{r \in R} \rho_r \check{r}_{i,r} \quad (1)$$

Let $c_i = [c_{H_2,i}, c_{CH_4,i}, c_{C_2H_6,i}, c_{C_2H_4,i}, c_{C_2H_2,i}, c_{CO,i}]$ denote the raw gas vector of sample i . To improve state characterization, the raw gas values are combined with ratio-oriented descriptors and an aggregated dissolved-gas analysis factor (DGAF) [23]. The DGAF descriptor is constructed in this study as an aggregated training-set descriptor. Its weights are statistically estimated from the state-discriminative capability of the candidate DGA features. In this study, the DGAF weights are not manually assigned by experts and are not optimized as additional free parameters of the final classifier. Instead, they are statistically estimated from the training samples according to the state-discriminative capability of each gas component and gas-ratio descriptor. Specifically, all gas concentrations and ratio descriptors are first normalized using the statistics of the training set. For each candidate feature $f \in G \cup R$, a Fisher-type separability score is calculated by comparing its between-state dispersion with its within-state dispersion, i.e.

$$s_f = \frac{\sum_c n_c (\mu_{c,f} - \mu_f)^2}{\sum_c \sum_{i \in D_c} (z_{i,f} - \mu_{c,f})^2 + \varepsilon} \quad (2)$$

where $\mu_{c,f}$ is the mean value of feature f in state c , μ_f is the global mean value of feature f in the training set, n_c is the number of training samples in state c , D_c denotes the training samples belonging to state c , $z_{i,f}$ is the normalized value of feature f for sample i , and ε is a small positive constant used to avoid division by zero. The DGAF weights are then obtained by normalizing the separability scores as $w_f = s_f / \sum_{l \in G \cup R} s_l$, where w_f

corresponds to ω_g when f is a gas-concentration feature and to ρ_r when f is a gas-ratio feature. Therefore, the DGAF used in this study can be expressed as follows:

$$DGAF_i = \sum_{g \in G} \omega_g c_{g,i} + \sum_{r \in R} \rho_r r_{i,r} \quad (3)$$

In each training split, the weights are estimated only from the training subset and then fixed when constructing validation and test features [24]. Therefore, the DGAF is a statistically estimated training-set descriptor rather than an expert-defined index or a classifier-optimized parameter. This design avoids introducing subjective empirical coefficients and prevents the DGAF construction from using information from the test set. In this framework, the DGAF is not used to replace conventional DGA ratios. Instead, it is combined with gas proportions and derived state features to increase the descriptive power of the monitoring data.

Based on this representation, the transformer state discrimination process shown in Figure 1 converts raw dissolved-gas observations into state-relevant descriptors. This feature construction stage provides a structured input space for sample structuring and ensemble learning, making the subsequent learning process more consistent with the fault-evolution characteristics of oil-immersed railway transformers.

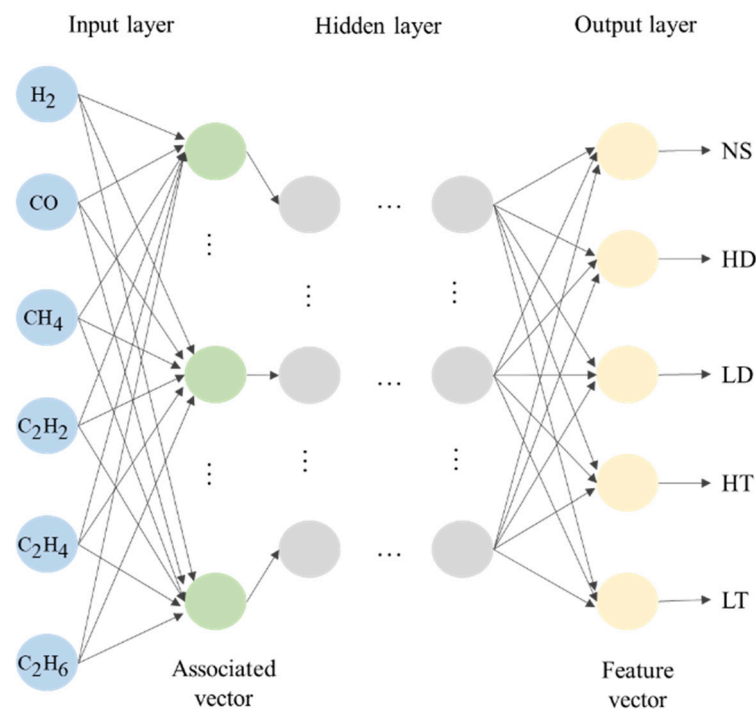


Figure 1. Transformer state classification network based on dissolved gas monitoring. Note: Arrows indicate the direction of data flow; colors distinguish raw gas inputs, intermediate feature representations, and transformer-state outputs.

2.2. DBSCAN-Assisted Structuring of Transformer State Samples

Transformer state samples collected from monitoring systems are rarely distributed uniformly. Normal-state samples usually dominate, whereas some abnormal states occur infrequently and may overlap with adjacent operating states. This imbalance leads to two difficulties for data-driven diagnosis: minority faults are insufficiently represented during training, and ambiguous samples may be overwhelmed by majority-class patterns. Direct supervised learning on the raw sample set may therefore result in unstable decision boundaries and weak generalization to rare operating conditions.

To address this issue, this study introduces density-based spatial clustering of applications with noise (DBSCAN) into the sample structuring stage. DBSCAN is suitable for DGA monitoring data because it can identify dense structures and isolate irregular or sparse observations without requiring a predefined number of clusters. In this study, the DBSCAN parameters are not fixed arbitrarily. The neighborhood radius ε and the minimum number of samples MinPts are selected based on the joint consideration of k-distance variation, clustering compactness, noise proportion, and downstream classification stability. Specifically, the k-distance curve is first used to determine a reasonable candidate range of ε , while MinPts is varied within a small interval related to the feature dimension and sample density. For each candidate parameter pair, the silhouette coefficient and the proportion of samples identified as sparse or boundary points are evaluated. The final parameter setting is selected when the clustering structure remains stable and the subsequent state-assessment performance does not show obvious degradation. To evaluate the compactness and separation of the clustering structure, the silhouette coefficient is used as follows (4). In this paper, DBSCAN is not used as the final classifier. Instead, its role is to reveal density structure, rare fault regions, boundary samples, and outliers, which are then used to support sample rebalancing and feature-guided learning. The sensitivity of the proposed framework to DBSCAN parameters and the corresponding clustering visualization results are further presented in Section 4.

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (4)$$

where a_i is the average distance between sample i and other samples in the same cluster, and b_i is the minimum average distance between sample i and samples in another cluster. A larger silhouette coefficient indicates better intra-cluster compactness and inter-cluster separation.

2.3. Construction and Rebalancing of State Feature Samples

On the basis of DGA-oriented feature representation and DBSCAN-assisted structuring, a cluster-aware rebalancing strategy is developed to improve the effective representation of minority and boundary-ambiguous transformer states. Let $D = \{(z_i, y_i)\}_{i=1}^N$ denote the constructed training set, where z_i is the DGA feature vector of sample i , $y_i \in \{1, \dots, C\}$ is the corresponding state label, N is the number of training samples, and C is the number of transformer states. For class c , the number of samples is denoted by n_c . To reduce the dominance of majority states, a class-level balancing factor is first defined as $\alpha_c = (N/(Cn_c))^\gamma$, where $\gamma \in [0, 1]$ controls the compensation strength for minority classes. A larger value of α_c is assigned to under-represented states such as low-energy discharge and low-temperature overheating. Then, the DBSCAN clustering result is used to further characterize the local sample structure. For each sample i , a structure-aware factor is defined as $\beta_i = 1 + \lambda_b I(i \in B) + \lambda_s I(i \in S)$, where B denotes the set of boundary-ambiguous samples, S denotes the set of sparse or low-density samples identified by DBSCAN, and $I(\cdot)$ is the indicator function. The coefficients λ_b and λ_s determine the additional emphasis placed on boundary and sparse samples, respectively. The final rebalancing weight of sample i is therefore calculated as $q_i = \alpha_{y_i} \beta_i / \sum_{j=1}^N \alpha_{y_j} \beta_j$. During model training, this weight is used as the sample weight in the supervised learning objective, so that the empirical loss is calculated as $L = -\sum_{i=1}^N q_i \sum_{c=1}^C I(y_i = c) \log p_{i,c}$, where $p_{i,c}$ is the predicted probability that sample i belongs to state c . Accordingly, minority-state samples and boundary-ambiguous samples contribute more effectively to the learning process, while majority-state samples are not completely discarded. This strategy differs from simple random oversampling because it considers not only class frequency but also the local density and boundary structure of the DGA feature space. In this way, the rebalanced

sample set better preserves representative majority-state information while strengthening the influence of sparse rare fault regions in subsequent feature-guided stacking learning.

3. Feature-Guided Stacking Framework for Transformer State Assessment

3.1. Stacking Learning and Its Limitation in Transformer State Assessment

Stacking learning is a hierarchical ensemble framework in which multiple base learners first generate preliminary predictions and a meta-learner subsequently integrates these outputs into the final decision. Different base models often capture different aspects of the feature distribution. Therefore, stacking can improve prediction stability when the task contains heterogeneous, nonlinear, or partially overlapping patterns. This characteristic is valuable for transformer state assessment because dissolved-gas signatures vary across fault mechanisms and operating conditions.

However, the effectiveness of conventional stacking depends strongly on sample composition and fusion strategy. If the training samples are severely imbalanced, base learners may become biased toward dominant states, and the meta-learner may inherit this bias. In addition, conventional stacking often combines base-learner outputs using a fixed integration mechanism, which ignores the fact that different learners may be reliable in different feature regions. Therefore, an adaptive stacking architecture should explicitly account for sample structure and feature-dependent learner reliability.

Although adaptive stacking and adaptive ensemble-learning methods have been studied to improve the flexibility of model fusion, most existing approaches adjust the contribution of base learners mainly according to global validation accuracy, prediction confidence, or sample-level classification difficulty. Such strategies are effective for general classification tasks, but they do not explicitly consider the physical and structural characteristics of DGA-based transformer monitoring data. In particular, the reliability of a base learner may vary with gas-generation patterns, class-dependent fault signatures, local sample density, and boundary ambiguity. Therefore, simply assigning adaptive weights according to overall validation performance may still cause the meta-learner to inherit majority-class bias or to underrepresent sparse rare fault regions. The advancement of the proposed framework lies in embedding DGA-specific monitoring features and DBSCAN-derived sample-structure information into the adaptive stacking process. Thus, base-learner fusion is guided not only by prediction performance, but also by the local feature reliability and structural position of each sample in the DGA feature space.

3.2. K-Fold Cross-Validation and Meta-Sample Construction

K-fold cross-validation is used in this study as an implementation procedure for constructing leakage-free meta-samples rather than as an independent methodological novelty. In the stacking framework, if the meta-learner is trained directly on the predictions of base learners evaluated on the same samples used for base-model training, the resulting meta-features may be overly optimistic and may increase the risk of overfitting. To avoid this problem, the training subset is divided into K mutually exclusive folds. In each round, $K - 1$ folds are used to train each base learner, and the remaining fold is used to generate out-of-fold predictions. After all folds are traversed, the out-of-fold predictions of all training samples are assembled to form the meta-sample set for training the second-level learner. In this way, the meta-learner observes the validation behavior of the base learners rather than their in-sample fitting results, which improves the reliability of the stacking process under imbalanced DGA data.

In the experiments, $K = 5$ is adopted. This value is selected as a trade-off between meta-sample reliability, minority-class representation, and computational cost. A smaller value of K , such as $K = 3$, increases the size of each validation fold but reduces the

number of out-of-fold evaluations, which may make the estimated base-learner behavior less stable. A larger value of K , such as $K = 10$, allows each base learner to be trained with more samples in each round, but it also increases computational burden and may lead to very small validation folds for sparse minority states such as low-energy discharge and low-temperature overheating. The five-fold setting keeps 80% of the training samples for base-learner fitting in each round while retaining sufficiently large validation folds for constructing stable meta-features. Therefore, the selected value of K is appropriate for the present imbalanced DGA assessment task. The main methodological contribution of this paper lies not in the use of K -fold cross-validation itself, but in the subsequent monitoring-feature-embedded adaptive weighting and feature-guided stacking mechanism built upon these reliable out-of-fold meta-samples.

3.3. Monitoring-Feature-Embedded Adaptive Learning

The key idea of the proposed framework is to embed monitoring features into the stacking architecture, so that the ensemble does not rely only on fixed base-learner outputs. Instead, the model uses feature-reliability information to determine how strongly each base learner should contribute to the final decision. The overall architecture is shown in Figure 2.

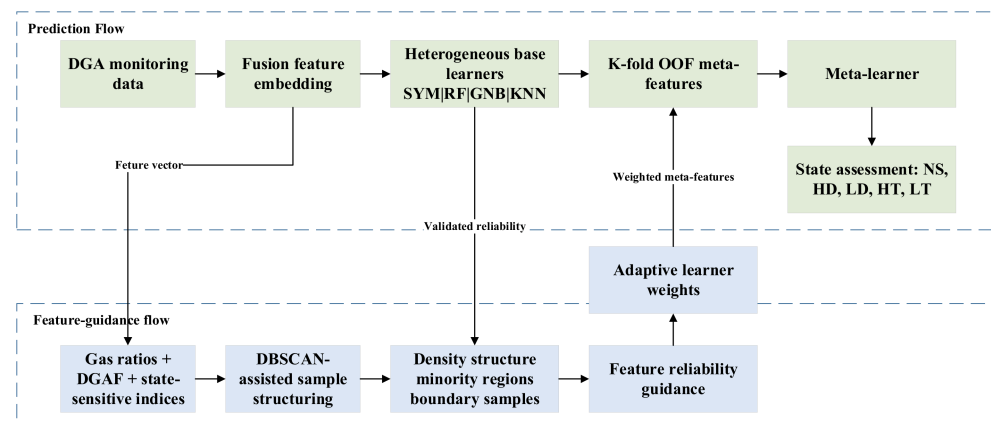


Figure 2. Architecture of the proposed feature-guided stacking framework for DGA-based railway transformer state assessment.

The adaptive learning process consists of four main steps. First, a feature-enhanced training objective is adopted for the meta-learner [21]. The objective combines supervised classification error and a feature-structure regularization term, as shown in Equation (5):

$$L_{\text{meta}} = -\frac{1}{N} \sum_i \sum_c y_i^{(c)} \log(\hat{p}_{i,c}) + \lambda \|g_i - \mu_{(y_i)}\|_2^2 \quad (5)$$

where N is the number of training samples, $y_i^{(c)}$ is the one-hot label of sample i for class c , $\hat{p}_{i,c}$ is the predicted probability of the meta-learner, g_i is the monitoring-feature vector, $\mu_{(y_i)}$ is the feature centroid of the corresponding state, and λ is a regularization coefficient. This term encourages the meta-learning process to remain consistent with the underlying state-feature structure.

Second, the fold-wise validation performance of each base learner is evaluated during K -fold training [25]. The validation error of base learner m in fold k is expressed by Equation (6):

$$e_{m,k} = \frac{1}{|V_k|} \sum_{i \in V_k} (h_m(x_i) \neq y_i) \quad (6)$$

where V_k denotes the validation subset in fold k , h_m is the prediction function of base learner m , and $I(\cdot)$ is the indicator function. The average validation performance provides a global reliability reference for each base learner.

Feature-dependent adaptive weights are calculated for the base learners by combining global validation reliability and local DGA feature-structure information. Different from conventional adaptive stacking methods that mainly rely on learner-level validation accuracy or prediction confidence, the proposed weighting mechanism evaluates whether a base learner is reliable for the specific monitoring pattern of a given sample. For sample i and base learner m , the feature-reliability score is constructed from three components: the fold-wise validation performance of learner m , the class-wise validation reliability for the state region predicted by learner m , and the local structure information obtained from the DGA feature space. The local structure information includes the distance between the sample and the corresponding state-feature centroid, as well as whether the sample belongs to a sparse or boundary-ambiguous region identified by DBSCAN. Therefore, the adaptive weight is not a fixed learner-level coefficient, but a sample-dependent fusion coefficient. For sample i , the weight assigned to base learner m is formulated as Equation (5), where η controls the sharpness of the weighting distribution and $\phi_{i,m}$ denotes the monitoring-feature-embedded reliability score. A larger $\phi_{i,m}$ indicates that learner m shows stronger validation reliability and better consistency with the local DGA feature structure of sample i . In this way, learners that are more reliable for minority, sparse, or boundary-ambiguous samples can receive larger contributions during meta-feature construction, which distinguishes the proposed FE-Stacking from conventional stacking and general adaptive ensemble fusion. Specifically, $\phi_{i,m}$ can be represented as a composite reliability score integrating validation reliability, class-wise recall, centroid consistency, and DBSCAN-based structure indicators, so that the adaptive fusion process is jointly driven by model performance and DGA-specific sample structure.

To make the monitoring-feature-embedded adaptive learning process reproducible, the variables used in the adaptive fusion module are explicitly defined as follows. For sample i , the DGA feature vector is denoted as z_i , which contains normalized gas concentrations, gas-ratio descriptors, the DGAF descriptor, and DBSCAN-derived structure indicators. The base-learner set is denoted as $\mathcal{M} = \{1, \dots, M\}$, and the probability output of base learner m for sample i is denoted as $p_{i,m} = [p_{i,m}^1, \dots, p_{i,m}^C]$, where C is the number of transformer states. The reliability score of learner m for sample i is constructed from three components: the average fold-wise validation accuracy A_m , the class-wise validation recall $R_{m,\hat{y}_{i,m}}$ corresponding to the predicted class $\hat{y}_{i,m} = \arg\max_c p_{i,m}^c$, and the local feature-structure consistency H_i . The local consistency term is calculated according to the distance between sample i and the centroid of its predicted state in the normalized DGA feature space, together with the DBSCAN-derived sparse and boundary indicators. Specifically, $H_i = \exp(-d_i^2/\sigma^2)(1 + \lambda_s I(i \in S) + \lambda_b I(i \in B))$, where d_i is the Euclidean distance to the corresponding state centroid, σ is the average centroid distance of the training samples, S and B denote the sparse-sample set and boundary-sample set identified by DBSCAN, respectively, and $I(\cdot)$ is the indicator function. The composite monitoring-feature-embedded reliability score is then calculated as $\phi_{i,m} = \alpha A_m + \beta R_{m,\hat{y}_{i,m}} + \delta H_i$, where α , β , and δ are nonnegative coefficients satisfying $\alpha + \beta + \delta = 1$.

$$\alpha_m(x_i) = \frac{\exp(\eta q_m(x_i))}{\sum_l \exp(\eta q_l(x_i))} \quad (7)$$

where $\alpha_m(x_i)$ is the adaptive weight, η controls the sharpness of the weighting distribution, and $q_m(x_i)$ is a feature-reliability score derived from monitoring features and

validation performance [26]. Thus, a base learner that is more reliable for a specific feature pattern obtains a larger contribution in the ensemble decision. The following objective function and adaptive weighting formulation are constructed in this study to embed monitoring-feature information into the stacking process. The classification-loss term follows the standard supervised learning setting, whereas the feature-structure regularization and adaptive reliability score are introduced to reflect the local DGA feature structure and DBSCAN-derived sample indicators.

Based on the above reliability score, the adaptive weight assigned to base learner m for sample i is obtained through softmax normalization, $a_{i,m} = \exp(\eta\phi_{i,m}) / \sum_{l=1}^M \exp(\eta\phi_{i,l})$, where η controls the sharpness of the weighting distribution. A larger η makes the fusion more dependent on the learner with the highest reliability score, whereas a smaller η produces a smoother combination of base learners. The weighted meta-feature vector is then constructed as $u_i = [a_{i,1}p_{i,1}, a_{i,2}p_{i,2}, \dots, a_{i,M}p_{i,M}, z_i^s]$, where z_i^s denotes the structure-related indicators used in the implementation, such as the DBSCAN noise indicator, sparse-sample indicator, and boundary-sample indicator. The meta-learner is trained on $\{(u_i, y_i)\}_{i=1}^N$ with the rebalanced sample weights defined in Section 2.3. During testing, the trained base learners first generate probability vectors for a new sample; then the reliability score and adaptive weights are calculated using the training-set centroids, fixed DGAF weights, and fixed DBSCAN parameter settings; finally, the trained meta-learner outputs the transformer state. All normalization parameters, DGAF weights, state centroids, DBSCAN parameters, and reliability coefficients are estimated only from the training data and fixed during validation and testing to avoid information leakage.

$$z_i = [\alpha_1(x_i)p_1(x_i), \alpha_2(x_i)p_2(x_i), \dots, \alpha_M(x_i)p_M(x_i), g_i] \quad (8)$$

where $p_M(x_i)$ is the prediction vector generated by base learner m and M is the number of base learners. The meta-learner is trained on z_i to produce the final transformer state assessment. Compared with Original Stacking, the proposed FE-Stacking not only preserves the diversity of heterogeneous base learners, but also links ensemble fusion to DGA feature structure and sample-density information. This design improves the generalization stability of the model under imbalanced and boundary-ambiguous monitoring conditions.

3.4. Algorithmic Procedure

The feature-guided stacking learning procedure for DGA-based transformer state assessment is presented as follows. The procedure is described in a stepwise format, including the input, output, and sequential operations.

Input: training dataset $D = \{(x_i, y_i)\}_{i=1}^N$, where x_i denotes the raw DGA measurements and y_i denotes the transformer state label; candidate gas-ratio feature set R ; DBSCAN parameters ε and MinPts; number of folds K ; base-learner set $\mathcal{M} = \{1, \dots, M\}$; adaptive-weighting parameter η ; rebalancing parameters γ , λ_s , and λ_b .

Output: trained FE-Stacking model for transformer state assessment.

Step 1. Normalize raw gas concentrations and diagnostic gas-ratio descriptors using the statistics of the training set. Estimate the DGAF weights from the training data according to the feature separability scores described in Section 2.1, and construct the DGA feature vector z_i for each sample.

Step 2. Apply DBSCAN to the normalized DGA feature space using the selected ε and MinPts. Identify dense regions, sparse samples, boundary-ambiguous samples, and noise samples. Construct the structure indicators used for subsequent rebalancing and adaptive fusion.

Step 3. Calculate the class-level balancing factor and DBSCAN-based structure-aware factor for each training sample according to Section 2.3. Obtain the final sample weight q_i ,

which increases the effective contribution of minority, sparse, and boundary-ambiguous samples during supervised learning.

Step 4. Divide the training set into K folds. For each base learner $m \in \mathcal{M}$ and each fold k , train the base learner on $K - 1$ folds and generate out-of-fold probability predictions on the remaining fold. Assemble the out-of-fold prediction vectors $p_{i,m}$ for all training samples.

Step 5. Calculate the monitoring-feature-embedded reliability score $\phi_{i,m}$ for each sample and each base learner by combining fold-wise validation accuracy, class-wise validation recall, centroid-based feature consistency, and DBSCAN-derived sparse/boundary indicators.

Step 6. Normalize the reliability scores through softmax to obtain the sample-dependent base-learner weight $a_{i,m}$. Construct the weighted meta-feature vector $u_i = [a_{i,1}p_{i,1}, a_{i,2}p_{i,2}, \dots, a_{i,M}p_{i,M}, z_i^s]$, where z_i^s denotes the structure-related indicators.

Step 7. Train the meta-learner using the weighted meta-feature set $\{(u_i, y_i)\}_{i=1}^N$ and the rebalanced sample weights q_i . The final state assessment model consists of the trained base learners, the adaptive weighting module, and the trained meta-learner.

Step 8. For a new sample, normalize its DGA features using training-set statistics, construct its DGAF descriptor, obtain the probability outputs of all trained base learners, calculate the monitoring-feature-embedded reliability scores and adaptive weights using fixed training-stage parameters, construct the weighted meta-feature vector, and output the transformer state through the trained meta-learner.

The computational complexity of the proposed framework mainly consists of DGA feature construction, DBSCAN-assisted sample structuring, K-fold base-learner training, adaptive meta-feature construction, and meta-learner training. Let N be the number of training samples, d be the number of DGA features, C be the number of transformer states, M be the number of base learners, and K be the number of folds. The DGAF and feature-construction stage requires $O(Nd)$ operations. The complexity of DBSCAN is generally $O(N \log N)$ when an efficient neighborhood-search structure is used, and may increase to $O(N^2)$ in the worst case. The dominant computational cost comes from K-fold base-learner training, which can be expressed as $O\left(K \sum_{m=1}^M T_m(N(K-1)/K, d)\right)$, where $T_m(\cdot)$ denotes the training complexity of base learner m . The construction of reliability scores and adaptive weights requires $O(NMC)$ operations because probability outputs from all base learners are processed for all samples and classes. The meta-learner is trained on weighted meta-features with dimension approximately $MC + d_s$, where d_s is the number of structure-related indicators. Therefore, its cost can be written as $O(T_{\text{meta}}(N, MC + d_s))$. Overall, the computational burden of the proposed framework is mainly determined by the repeated training of heterogeneous base learners in the K-fold stacking procedure, while the DGAF calculation, DBSCAN-based indicators, and adaptive weighting introduce only moderate additional overhead.

The memory requirement mainly includes storage of the DGA feature matrix, DBSCAN-related structure indicators, out-of-fold base-learner probability outputs, and weighted meta-features. The feature matrix requires $O(Nd)$ memory, while the out-of-fold prediction matrix requires $O(NMC)$ memory, where N is the number of samples, d is the number of DGA features, M is the number of base learners, and C is the number of transformer states. The additional structure indicators introduced by DBSCAN require only $O(Nd_s)$ memory and therefore contribute limited additional storage overhead. If DBSCAN is implemented using a full pairwise distance matrix, the memory requirement may increase to $O(N^2)$; otherwise, neighborhood-search implementations can reduce the practical memory burden. In the present implementation, the main memory cost comes from storing the out-of-fold probability outputs for meta-learning, rather than from the DGAF construction or the adaptive weighting module. The offline training time is mainly

determined by the repeated K-fold training of heterogeneous base learners, whereas the DBSCAN-assisted structuring and adaptive-weight calculation introduce relatively small additional computational overhead. During online testing, only one forward prediction of each trained base learner, adaptive-weight calculation, and meta-learner inference are required; therefore, the per-sample testing time remains suitable for offline assessment and decision-support applications.

4. Simulation-Based Validation and Discussion

4.1. Dataset Description and Feature Construction

To verify the effectiveness of the proposed FE-Stacking method under controlled and reproducible conditions, a synthetic dissolved-gas-analysis dataset was constructed for railway power transformer state assessment. The synthetic samples were generated according to physically consistent gas-distribution patterns associated with five transformer states, namely, normal state (NS), high-energy discharge (HD), low-energy discharge (LD), high-temperature overheating (HT), and low-temperature overheating (LT).

Each sample contains six raw gas variables, namely, H_2 , CH_4 , C_2H_6 , C_2H_4 , C_2H_2 , and CO. These gases were selected according to standard DGA diagnostic principles because they represent the main combustible and insulation-decomposition gases associated with discharge-related and overheating-related transformer states. Based on these raw variables, three diagnostic gas-ratio features and one DGAF-related aggregated descriptor are further constructed. To improve the realism of the simulation, transformer-specific offsets, multiplicative measurement noise, and a small proportion of outliers were introduced during data generation. The synthetic dataset was generated under a controlled proof-of-concept validation setting. The generation process was designed to reflect five representative transformer states, including NS, HD, LD, HT, and LT. The state-dependent gas patterns were set to be consistent with DGA diagnostic knowledge: discharge-related states contain higher H_2 and C_2H_2 levels, whereas overheating-related states contain higher C_2H_4 , C_2H_6 , and CO levels. The generated samples were then perturbed by virtual-transformer-level offsets, multiplicative noise, and a small proportion of outliers to approximate monitoring variability.

The synthetic dataset was generated at the virtual-transformer level rather than by independently sampling isolated records. Each virtual transformer carries mild domain shifts in thermal response, discharge tendency, and gas-level offsets. Therefore, the data partition is more consistent with cross-device generalization and reduces the risk of record-level information leakage. In addition, the sample distribution was intentionally imbalanced, with LD and LT treated as minority and boundary-ambiguous fault categories. This setting is aligned with the objective of evaluating generalization stability under imbalanced DGA-based state assessment.

Although the synthetic DGA dataset enables controlled and reproducible evaluation of the proposed framework under imbalanced and boundary-ambiguous conditions, validation using real transformer monitoring records is essential for further assessing its field applicability. In practical railway traction power-supply systems, long-term DGA records are usually subject to data confidentiality, incomplete maintenance annotations, inconsistent sampling intervals, and a limited number of confirmed fault events. These factors make it difficult to construct a sufficiently labeled real-world dataset covering all considered transformer states. Therefore, this study uses a virtual-transformer-level synthetic dataset as a proof-of-concept validation platform, where transformer-specific offsets, measurement disturbances, outliers, and rare fault imbalance are introduced to approximate realistic data heterogeneity. Nevertheless, the present results should be interpreted as evidence of methodological feasibility rather than direct field-deployment validation. Future work will further evaluate the proposed framework using long-term field DGA records from railway

traction transformers and compare its diagnostic performance with expert maintenance records and operational inspection results.

The values reported in Table 2 are not field-measured DGA concentrations. They are statistical summaries of the generated synthetic dataset. Specifically, for each transformer state and each gas variable, the reported concentration is the arithmetic mean of all generated samples belonging to that state after applying the state-specific gas-generation pattern, virtual-transformer offset, multiplicative measurement disturbance, and outlier injection. The unit is ppm. The sample count column reports the number of generated samples in each state. To visualize the distribution of the synthetic DGA guide features, the projected distribution is shown in Figure 3.

Table 2. Class-wise sample means and class distribution of the generated synthetic DGA dataset.

State	H ₂ (ppm)	CH ₄ (ppm)	C ₂ H ₆ (ppm)	C ₂ H ₄ (ppm)	C ₂ H ₂ (ppm)	CO (ppm)	Sample Count
NS	8.63	15.23	4.50	9.72	0.45	71.07	960
HD	644.60	690.32	81.07	393.08	190.92	253.47	380
LD	277.81	136.50	31.37	62.60	26.54	151.32	130
HT	185.84	336.76	172.61	769.07	28.59	206.53	470
LT	174.15	164.30	64.69	463.88	11.30	156.31	190

Note: The gas concentrations in this table are generated dataset statistics. For each state, the reported value is the arithmetic mean of the corresponding gas concentration over all generated samples in that class. The unit is ppm.

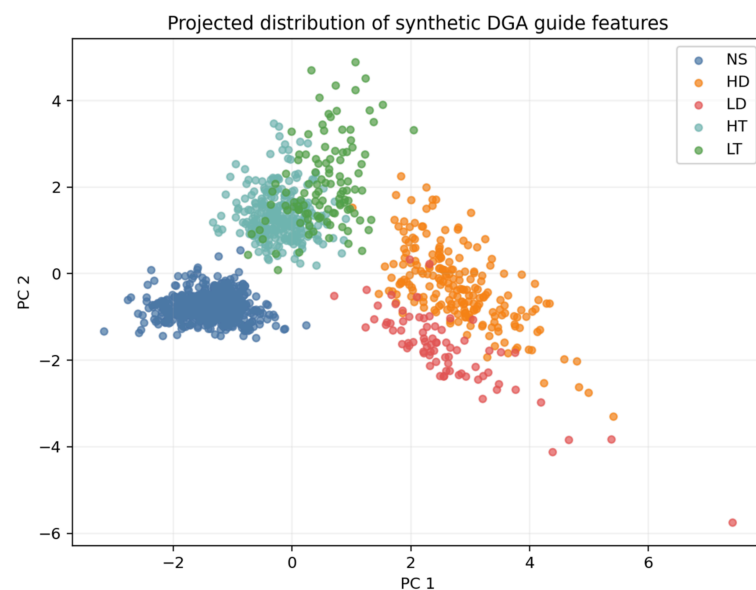


Figure 3. Projected distribution of synthetic DGA guide features.

To improve the reproducibility of the experimental protocol, the key hyperparameters and implementation settings used in this study are summarized in Table 3. The table reports the random seeds, transformer state labels, raw gas variables, derived DGA features, DBSCAN parameter settings, base learners, evaluation metrics, and minority-retention sensitivity settings. All feature preprocessing, DBSCAN-assisted structuring, and model evaluation procedures were conducted under the same random seeds and data partitions. The items summarized in Table 3 were selected according to the principles of reproducibility, task relevance, fair comparison, and consistency with the research questions. The selected base learners cover different learning mechanisms, including margin-based learning, tree-based ensemble learning, probabilistic classification, and instance-based classification. The ensemble variants are used to isolate the effects of DBSCAN-assisted sample structuring

and feature-guided adaptive weighting. Cross-validation, random seeds, and evaluation metrics are included to ensure reproducible and fair model evaluation. The minority-retention sensitivity setting is included to examine whether the proposed framework remains robust when LD and LT minority samples are progressively compressed.

Table 3. Experimental configuration and item-selection rationale for reproducible evaluation.

Item	Setting
Random seeds	2026, 2042, 2058
Transformer states	NS, HD, LD, HT, LT
Raw gas variables	H ₂ , CH ₄ , C ₂ H ₆ , C ₂ H ₄ , C ₂ H ₂ , CO
Derived features	Gas ratios, DGAF-related descriptor, thermal index, discharge index, logarithmic gas features
DBSCAN epsilon	Representative value 1.760; sensitivity range {1.408, 1.584, 1.760, 1.936, 2.112}
DBSCAN MinPts	Representative value 10; sensitivity range {8, 10, 12}
DBSCAN evaluation note	Silhouette coefficient computed after excluding noise points; sparse/boundary proportion reported as noncore samples plus noise
Base learners	SVM, RF, GNB, KNN
Evaluation metrics	Accuracy, macro-F1, balanced accuracy, class-wise recall
Minority-retention sensitivity setting	LD and LT training samples retained at 100%, 70%, 40%, and 25%

The same settings were used across the compared methods unless otherwise specified, so that the performance differences mainly reflect the proposed feature-guided stacking and sample-structuring mechanisms rather than inconsistent experimental configurations.

$$STD = \sqrt{(\sum_{r=1}^R (m_r - mean)^2 / (R - 1))} \quad (9)$$

4.2. Experimental Protocol and Evaluation Metrics

The synthetic dataset was partitioned according to virtual transformer identities into training, validation, and test subsets, thereby reducing information leakage caused by record-level random splitting. All numerical features were standardized using training-set statistics only. The FE-Stacking framework employed four first-level learners, namely, a linear support vector machine (SVM), a random forest (RF), a Gaussian naive Bayes classifier (GNB), and a k-nearest neighbor classifier (KNN). Five-fold cross-validation was adopted to generate out-of-fold meta-features for the second-level learner. In the improved framework, DBSCAN-assisted sample handling was used to enhance rare fault representation, while feature-guided adaptive weighting dynamically adjusted the contribution of different base learners according to ratio-oriented fault descriptors.

The performance of different methods was evaluated using accuracy, macro-F1, balanced accuracy, and class-wise recall. Because the synthetic dataset was deliberately imbalanced, macro-F1 and balanced accuracy were treated as primary indicators in addition to overall accuracy. Besides the default synthetic test setting, an additional sensitivity experiment progressively reduced the retained LD and LT training samples to 70%, 40%, and 25%, respectively. This harder setting was designed to examine whether the proposed method can maintain robust rare fault recognition when class imbalance becomes severe.

A parameter-stability analysis was conducted for the DBSCAN-assisted structuring module on the training-side DGA feature space. DBSCAN was evaluated as an auxiliary sample-structuring tool rather than as the final classifier. The default neighborhood radius was estimated from the 10-NN distance curve, which yielded an elbow near epsilon = 1.760. A narrow perturbation range of epsilon in {1.408, 1.584, 1.760, 1.936, 2.112} and MinPts in {8, 10, 12} was then examined. As shown in Table 4, the results show that excessively small epsilon values increase cluster fragmentation and the proportion of points labeled as noise, whereas overly large epsilon values tend to merge adjacent state regions and reduce

the amount of sparse/boundary samples highlighted by DBSCAN. In contrast, within the moderate parameter range around $\epsilon = 1.760$ and $\text{MinPts} = 10$, the number of clusters, noise proportion, and silhouette coefficient remain comparatively stable, indicating that the sample-structuring behavior is not overly sensitive to small parameter perturbations.

Table 4. Parameter-stability analysis for the DBSCAN-assisted sample-structuring module.

Epsilon	MinPts	Number of Clusters	Noise Proportion (%)	Silhouette Coefficient	Sparse/Boundary Proportion (%)	LD/LT Coverage
1.408	8	3	4.96	0.578	13.28	LD: 12.1%; LT: 26.3%
1.584	8	3	2.88	0.576	7.65	LD: 7.7%; LT: 14.3%
1.760	8	3	1.95	0.572	5.10	LD: 6.6%; LT: 6.0%
1.936	8	3	1.41	0.570	3.82	LD: 6.6%; LT: 4.5%
2.112	8	2	1.07	0.595	2.48	LD: 5.5%; LT: 2.3%
1.408	10	3	5.37	0.580	15.69	LD: 14.3%; LT: 31.6%
1.584	10	3	3.42	0.577	9.86	LD: 11.0%; LT: 19.5%
1.760	10	3	2.28	0.574	5.90	LD: 7.7%; LT: 9.0%
1.936	10	3	1.54	0.571	4.76	LD: 7.7%; LT: 6.0%
2.112	10	2	1.21	0.596	2.88	LD: 5.5%; LT: 3.8%
1.408	12	3	6.17	0.582	18.18	LD: 19.8%; LT: 36.1%
1.584	12	3	3.89	0.578	12.07	LD: 12.1%; LT: 24.8%
1.760	12	3	2.41	0.574	7.31	LD: 8.8%; LT: 14.3%
1.936	12	3	1.61	0.571	5.16	LD: 7.7%; LT: 8.3%
2.112	12	2	1.27	0.596	3.55	LD: 6.6%; LT: 4.5%

Note: For each method and each metric, the reported mean and standard deviation are calculated as the arithmetic mean and sample standard deviation. STD denotes the sample standard deviation of the test-set metric values over three repeated computational runs with random seeds 2026, 2042, and 2058. It does not denote the standard deviation of repeated gas-concentration measurements.

Under the representative setting $\epsilon = 1.760$ and $\text{MinPts} = 10$, DBSCAN identified 3 clusters with a noise proportion of 2.28%, a silhouette coefficient of 0.574, and a sparse/boundary proportion of 5.90%. The sparse-region observations (LD: 7.7%; LT: 9.0%) further show that minority states are more frequently located in low-density or boundary-adjacent regions than dominant samples, which is consistent with the intended role of DBSCAN in this framework. Figure 4 visualizes the same DGA samples in a two-dimensional PCA space. Figure 4a displays the ground-truth transformer states, while Figure 4b shows the corresponding DBSCAN clusters, with noncore sparse/boundary samples and noise points explicitly marked. The visualization confirms that DBSCAN mainly serves to expose dense operating regions, sparse minority patterns, and ambiguous transition areas, thereby providing structural guidance for later supervised learning instead of replacing the classifier itself.

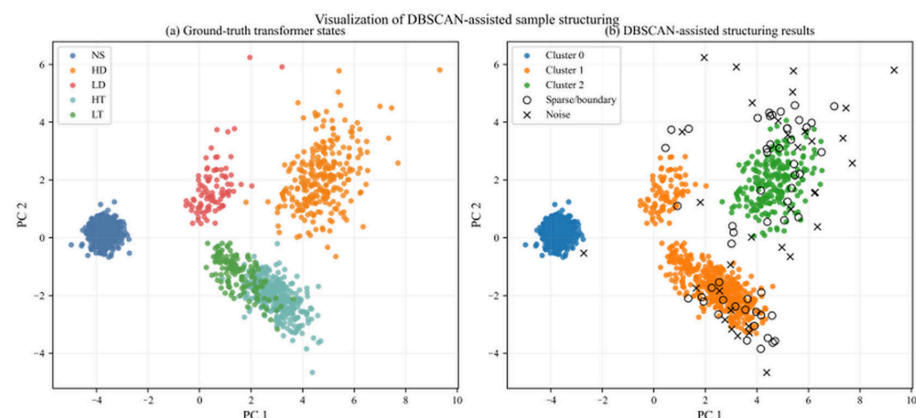


Figure 4. Visualization of DBSCAN-assisted sample structuring. (a) Two-dimensional projection of DGA samples labeled by transformer states. (b) DBSCAN clustering results with sparse and noise samples marked.

4.3. Results and Discussion

4.3.1. Overall Classification Performance

Under the default synthetic setting, RF remained the strongest single method, indicating that the generated DGA features were highly compatible with tree-based nonlinear partitioning. Nevertheless, FE-Stacking achieved a competitive and balanced recognition profile across all five states. This result suggests that the proposed framework can effectively combine complementary base learners while maintaining stable generalization under mixed and imbalanced gas patterns.

The overall performance comparison and representative-split confusion matrix under the default synthetic setting are shown in Figure 5. Although RF achieved the highest average performance under the default synthetic setting, with an accuracy of 99.85% and a macro-F1 of 99.79%, this result should be interpreted together with the difficulty level of the default dataset and the objective of the proposed framework. The default synthetic setting still provides relatively sufficient rare fault samples and clear feature separability, under which a strong tree-based ensemble such as RF can already learn highly effective decision boundaries. Therefore, FE-Stacking is not intended to be claimed as universally superior to RF in all scenarios, nor is its additional complexity justified by marginal accuracy improvement under this relatively easy setting alone. Instead, the main value of FE-Stacking lies in improving robustness when the training distribution becomes more imbalanced and boundary-ambiguous. As shown in the subsequent minority-retention sensitivity analysis, the proposed feature-guided sample structuring and adaptive fusion mechanism become more beneficial when LD and LT samples are severely compressed. From a practical perspective, RF may be sufficient for routine assessment when labeled DGA samples are abundant and well separated, whereas FE-Stacking provides a more robust decision-support option for difficult diagnostic scenarios involving sparse minority faults, boundary-overlapping DGA signatures, and cross-transformer heterogeneity.

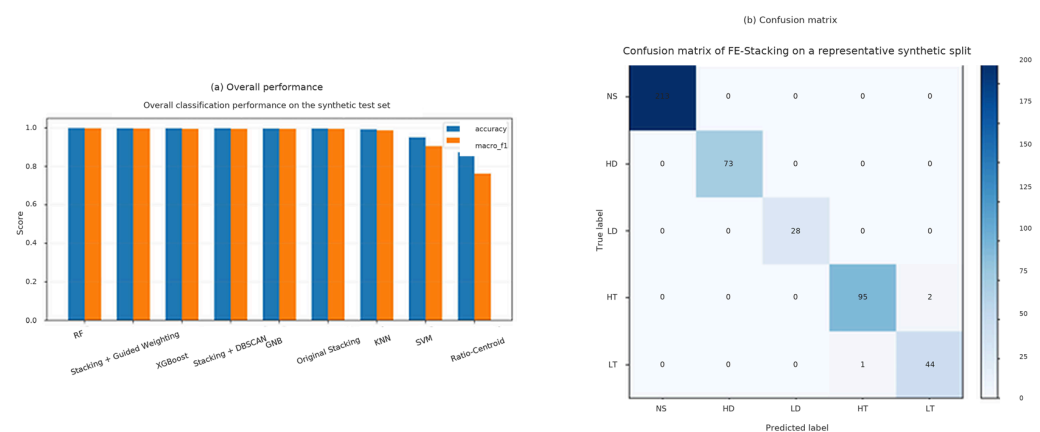


Figure 5. Overall performance comparison and confusion-matrix visualization under the default synthetic setting.

4.3.2. Ablation Study on Sample Rebalancing and Feature Embedding

To evaluate the contribution of each module, a more challenging ablation experiment was performed by retaining only 25% of the LD and LT training samples while keeping the test set unchanged. This setting better matches the intended use of FE-Stacking, namely, rare fault recognition under severe class imbalance. Table 5 shows that FE-Stacking obtained the highest accuracy (99.62%), macro-F1 (99.40%), and balanced accuracy (99.25%) among the ensemble variants under the 25% minority-sample retention setting. The overall test-set performance under the default synthetic setting is summarized in Table 6. The result

indicates that DBSCAN-assisted sample handling and feature-guided weighting become more valuable when rare fault information is strongly compressed.

Table 5. Ablation results under 25% minority-sample retention for LD and LT (mean, %).

Method	Accuracy	Macro-F1	Balanced Accuracy
Original Stacking	99.23	98.70	98.18
Stacking + DBSCAN	99.46	99.13	98.88
Stacking + Guided Weighting	99.38	99.02	98.66
FE-Stacking	99.62	99.40	99.25
RF	99.53	99.26	98.98

Table 6. Overall test-set performance under the default synthetic setting (mean $\pm$ std, %).

Method	Accuracy	Macro-F1	Balanced Accuracy
Ratio-Centroid	87.32 $\pm$ 0.70	76.23 $\pm$ 0.94	72.69 $\pm$ 1.07
SVM	95.09 $\pm$ 0.32	90.47 $\pm$ 0.73	88.00 $\pm$ 0.74
KNN	99.23 $\pm$ 0.27	98.72 $\pm$ 0.56	98.57 $\pm$ 0.43
GNB	99.63 $\pm$ 0.63	99.47 $\pm$ 0.92	99.58 $\pm$ 0.73
RF	99.85 $\pm$ 0.25	99.79 $\pm$ 0.37	99.86 $\pm$ 0.24
Original Stacking	99.63 $\pm$ 0.45	99.44 $\pm$ 0.66	99.46 $\pm$ 0.53
FE-Stacking	99.70 $\pm$ 0.33	99.55 $\pm$ 0.49	99.53 $\pm$ 0.43

The ablation comparison under the 25% minority-sample retention setting is visualized in Figure 6. Compared with Original Stacking, the full FE-Stacking framework improved macro-F1 by 0.70 percentage points under the 25% minority-retention setting. Compared with the guided-weighting-only variant, FE-Stacking also retained a measurable advantage. This indicates that explicit enhancement of rare fault representation still contributes to performance when the synthetic data are intentionally made more challenging.

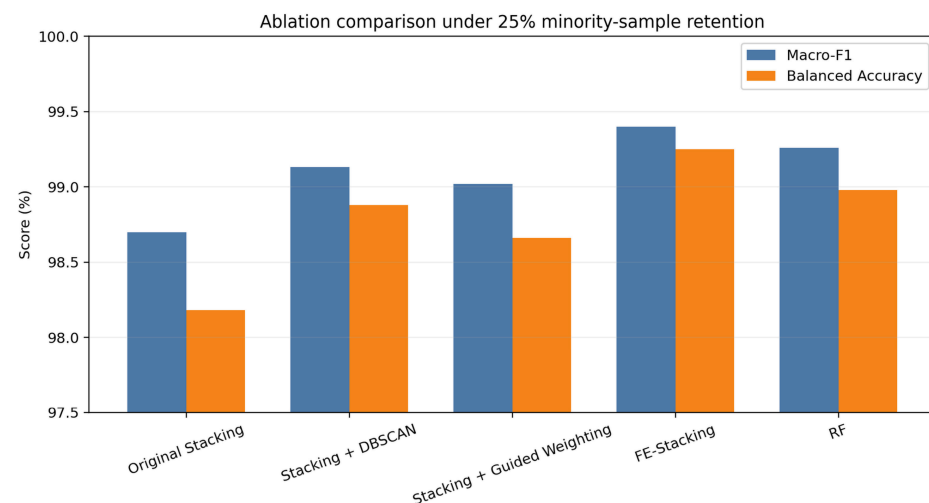


Figure 6. Ablation comparison under 25% minority-sample retention.

4.3.3. Recognition of Minority and Boundary-Ambiguous Faults

The recognition of LD and LT is particularly important, because these two classes were designed as minority and partially overlapping fault states in the synthetic dataset. Table 7 shows that, under the 25% minority-retention setting, FE-Stacking maintained a recall of 100.00% for LD and 96.61% for LT. Its LD recall was tied with several strong baselines, while its LT recall was higher than those of RF, Original Stacking, and the guided-weighting-only

variant. The rare-fault recall results for LD and LT are visualized in Figure 7. This result is meaningful because LT is the more boundary-ambiguous overheating category and is therefore more sensitive to insufficient rare fault representation. These observations indicate that the advantage of FE-Stacking does not lie in uniformly dominating RF under easier conditions. Rather, it lies in maintaining stronger rare fault robustness when the training distribution becomes more challenging. In this sense, the proposed framework better matches the objective of adaptive state assessment under imbalanced DGA data.

Table 7. Rare fault recall under 25% minority-sample retention (mean, %).

Method	Recall of LD	Recall of LT
SVM	9.44	27.33
Original Stacking	98.41	92.49
Stacking + DBSCAN	100.00	94.76
Stacking + Guided Weighting	100.00	93.28
FE-Stacking	100.00	96.61
RF	100.00	94.89

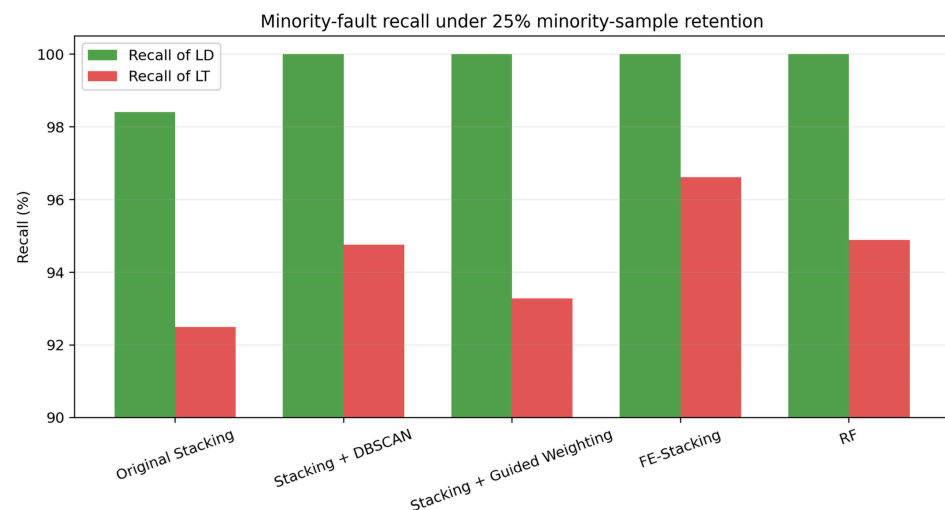


Figure 7. Rare fault recall under 25% minority-sample retention.

4.3.4. Robustness and Sensitivity Analysis

To assess how the methods behave when the test samples increasingly differ from the training distribution, we gradually increased the proportion of unseen samples from 20% to 100%. As shown in Figure 8, the macro-F1 score declined for all methods, although the decline was much less pronounced for FE-Stacking. When only a relatively small proportion of samples was unseen, FE-Stacking achieved a macro-F1 of around 0.985, while the other methods generally performed between approximately 0.91 and 0.97. As the proportion of unseen samples increased, the performance gap became more noticeable. Under the most challenging setting, FE-Stacking still remained close to 0.92, whereas Stacking + Guided Weighting and Original Stacking fell to roughly 0.88 and 0.82, respectively. The scores of RF and SVM decreased further, reaching below 0.70 and approximately 0.62. Overall, FE-Stacking lost only about 0.07 in macro-F1 across the tested range, while the other methods experienced larger declines, particularly RF and SVM. At the highest unseen-sample proportion, FE-Stacking performed more than 0.2 higher than RF and nearly 0.3 higher than SVM. This trend suggests that the feature-guided fusion mechanism helps preserve useful class-discriminative information when the test distribution becomes less familiar.

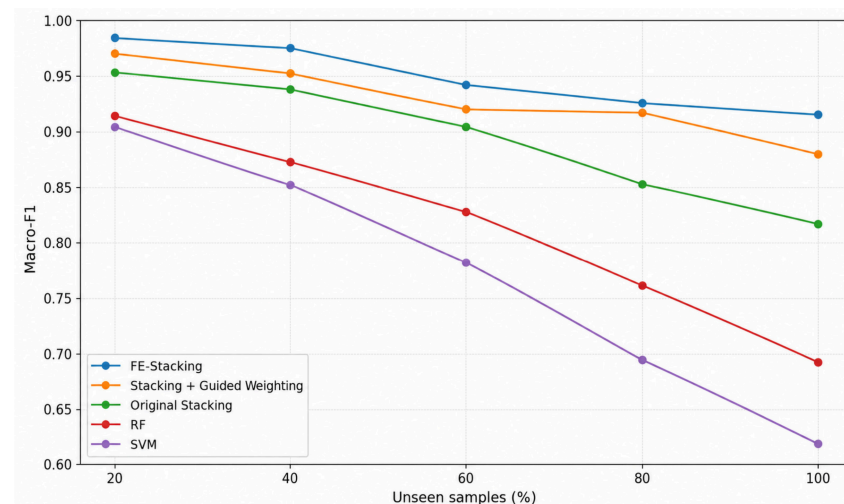


Figure 8. Sensitivity of macro-F1 to the proportion of unseen samples.

The slower degradation of FE-Stacking under severe minority-sample compression is attributed to the combined effect of DBSCAN-assisted sample structuring, cluster-aware rebalancing, and monitoring-feature-embedded adaptive weighting. DBSCAN helps identify sparse and boundary-adjacent DGA samples. The rebalancing strategy increases their effective contribution during training. The adaptive weighting module further adjusts the contribution of heterogeneous base learners according to local feature reliability and validation behavior. Therefore, FE-Stacking can better preserve minority-state information when LD and LT samples are severely reduced. Nevertheless, this interpretation is based on the controlled synthetic validation setting and should be further examined using real railway transformer DGA records.

The sensitivity trend supports a nuanced conclusion. RF yields the highest average score under the default synthetic setting, while FE-Stacking provides a more favorable trade-off between ensemble enhancement and rare fault robustness under the more demanding imbalance-oriented setting. This behavior is consistent with the proposed motivation that feature-guided ensemble learning is most useful when rare fault information is insufficient and boundary ambiguity becomes more pronounced.

4.4. Discussion

The simulation-based experiments provide several observations regarding the proposed FE-Stacking framework. Under the default synthetic validation setting, both RF and FE-Stacking achieve highly competitive overall performance. In particular, RF obtains the highest average score on the default synthetic test set, indicating that tree-based ensemble models remain strong baselines when DGA feature regions are relatively well separated and minority-state samples are still sufficiently represented. Therefore, the proposed FE-Stacking framework should not be interpreted as being universally superior to RF under all diagnostic conditions. These findings support the use of feature-guided stacking as an adaptive ensemble framework for imbalanced DGA-based railway power transformer state assessment.

From a practical perspective, the proposed framework may be useful as a decision-support module for railway transformer condition screening, alarm verification, and maintenance prioritization. In such applications, the model output should not replace expert judgment or standard maintenance procedures. Instead, it can provide an additional data-driven reference for identifying suspicious operating states, prioritizing transformers requiring inspection, and supporting maintenance decision making. The feature-guided

structure also provides a more transparent way to relate model behavior to DGA feature reliability and local sample density, which is valuable for condition-monitoring applications.

Although some critical results have been achieved, this work still has limitations. The dataset is generated at the virtual simulation. Although the generator introduces transformer-specific offsets, measurement disturbances, outliers, and class imbalance, it cannot fully reproduce sensor drift, oil-sampling uncertainty, maintenance-history uncertainty, irregular monitoring intervals, aging trajectories, and site-specific operating disturbances. In addition, the proposed model may still learn patterns that are specific to the synthetic data-generation mechanism. The virtual-transformer-level split and the use of multiple random seeds reduce record-level information leakage and improve the stability of the evaluation, but they cannot completely eliminate the possibility that the model adapts to generator-specific regularities. The practical applicability of the proposed framework should be further verified using multi-site monitoring data and long-term on-line condition-monitoring records. Therefore, the reported performance should be regarded as proof-of-concept evidence for methodological feasibility.

Future work will focus on three aspects. First, long-term field DGA records from railway traction power transformers should be collected and anonymized to validate the proposed framework under real monitoring conditions. Second, multi-site datasets should be used to evaluate cross-transformer and cross-region generalization. Third, additional monitoring variables, such as CO₂, O₂, N₂, moisture content, oil temperature, loading level, and maintenance-history information, should be incorporated to improve field applicability. These extensions will help determine whether the proposed feature-guided fusion strategy can be reliably transferred from controlled synthetic validation to practical railway transformer condition assessment.

5. Conclusions

This paper presented a feature-guided stacking framework for adaptive state assessment of railway traction power transformers. The proposed method integrates DGA-based fusion-feature characterization, DBSCAN-assisted sample structuring, and monitoring-feature-embedded adaptive weighting within a stacking architecture. The objective is to improve the representation of imbalanced fault samples and enhance the discrimination of boundary-ambiguous states.

Simulation-based experiments on a synthetic DGA dataset show that FE-Stacking remains highly competitive under the default synthetic setting and becomes more advantageous when rare fault categories are severely underrepresented. Specifically, FE-Stacking achieves an accuracy of 99.70% and a macro-F1 of 99.55% under the default setting. Under the 25% minority-retention setting, it obtains an accuracy of 99.62%, a macro-F1 of 99.40%, and a higher LT recall than RF and other ensemble variants. These results indicate that feature-guided ensemble learning can improve rare fault robustness and generalization stability in imbalanced DGA-based state assessment.

The present study should be regarded as a reproducible proof-of-concept based on virtual-transformer-level synthetic data. Future work will validate the proposed framework using real railway traction transformer DGA records, multi-site monitoring data, and long-term online condition-monitoring datasets. Further research will also examine model calibration, uncertainty quantification, and online updating strategies for practical deployment.

A limitation of this study is that the experimental validation is currently based on a virtual-simulation synthetic DGA dataset, for which design supports controlled comparison and reproducible analysis under rare fault and boundary-ambiguous conditions. Nevertheless, it cannot fully represent the complexity of real railway transformer monitoring data, such as sensor drift, maintenance-history uncertainty, irregular sampling,

incomplete fault annotations, transformer aging histories, and site-specific operating disturbances. Therefore, the present results should be interpreted as proof-of-concept evidence for methodological feasibility rather than direct evidence of operational deployment readiness. The robustness of stagey against synthetic-data bias and generator-specific patterns requires further investigation. Future work will focus on collecting and anonymizing long-term field DGA records from railway traction power transformers, validating the proposed framework on real monitoring datasets, and further assessing its robustness under practical condition-screening, maintenance-prioritization, and alarm-verification scenarios.

Author Contributions: Conceptualization, Y.H. and Z.H.; methodology, Y.H.; validation, Z.H. and J.C.; formal analysis, Y.H.; investigation, Z.H.; data curation, J.C.; writing—original draft preparation, Y.H.; writing—review and editing, Z.H. and J.C.; visualization, J.C.; supervision, Z.H.; funding acquisition, Z.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by 2025 Scientific Research Platforms and Projects of Guangdong Provincial Education Department, Grant No. 2025KTSCX290; and the 2025 Central Government Guided Local Science and Technology Development Special Fund (Shantou Innovative City Construction), Grant No. STKJ2025095.

Data Availability Statement: The synthetic data generated and analyzed in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Jahromi, A.N.; Piercy, R.; Cress, S.; Service, J.; Fan, W. An approach to power transformer asset management using health index. *IEEE Electr. Insul. Mag.* **2009**, *25*, 20–34. [\[CrossRef\]](#)
2. Abu-Siada, A.; Islam, S. A new approach to identify power transformer criticality and asset management decision based on dissolved gas-in-oil analysis. *IEEE Trans. Dielectr. Electr. Insul.* **2012**, *19*, 1007–1012. [\[CrossRef\]](#)
3. Jin, L.; Kim, D.; Chan, K.Y.; Abu-Siada, A. Deep machine learning-based asset management approach for oil-immersed power transformers using dissolved gas analysis. *IEEE Access* **2024**, *12*, 27794–27809. [\[CrossRef\]](#)
4. Dladla, V.M.N.; Thango, B. Fault classification in power transformers via dissolved gas analysis and machine learning algorithms: A systematic literature review. *Appl. Sci.* **2025**, *15*, 2395. [\[CrossRef\]](#)
5. Sun, H.C.; Huang, Y.C.; Huang, C.M. A review of dissolved gas analysis in power transformers. *Energy Procedia* **2012**, *14*, 1220–1225. [\[CrossRef\]](#)
6. Wani, S.A.; Rana, A.S.; Sohail, S.; Rahman, O.; Parveen, S.; Khan, S.A. Advances in DGA based condition monitoring of transformers: A review. *Renew. Sustain. Energy Rev.* **2021**, *149*, 111347. [\[CrossRef\]](#)
7. *IEEE Std C57.104-2019*; IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers. IEEE: New York, NY, USA, 2019. [\[CrossRef\]](#)
8. *IEC 60599:2022*; Mineral Oil-Filled Electrical Equipment in Service—Guidance on the Interpretation of Dissolved and Free Gases Analysis. International Electrotechnical Commission: Geneva, Switzerland, 2022.
9. Duval, M. A review of faults detectable by gas-in-oil analysis in transformers. *IEEE Electr. Insul. Mag.* **2002**, *18*, 8–17. [\[CrossRef\]](#)
10. Su, Q.; Mi, C.; Lai, L.L.; Austin, P. A fuzzy dissolved gas analysis method for the diagnosis of multiple incipient faults in a transformer. *IEEE Trans. Power Syst.* **2000**, *15*, 593–598. [\[CrossRef\]](#)
11. Khan, S.A.; Equbal, D.; Islam, T. A comprehensive comparative study of DGA based transformer fault diagnosis using fuzzy logic and ANFIS models. *IEEE Trans. Dielectr. Electr. Insul.* **2015**, *22*, 590–596. [\[CrossRef\]](#)
12. Ghoneim, S.S.M.; Taha, I.B.M.; Elkalashy, N.I. Integrated ANN-based proactive fault diagnostic scheme for power transformers using dissolved gas analysis. *IEEE Trans. Dielectr. Electr. Insul.* **2016**, *23*, 1838–1845. [\[CrossRef\]](#)
13. Li, S.; Wu, G.; Gao, B.; Hao, C.; Xin, D.; Yin, X. Interpretation of DGA for transformer fault diagnosis with complementary SaE-ELM and arctangent transform. *IEEE Trans. Dielectr. Electr. Insul.* **2016**, *23*, 586–595. [\[CrossRef\]](#)
14. Taha, I.B.M.; Ibrahim, S.; Mansour, D.E.A. Power transformer fault diagnosis based on DGA using a convolutional neural network with noise in measurements. *IEEE Access* **2021**, *9*, 111162–111170. [\[CrossRef\]](#)

15. Tan, X.; Guo, C.; Wang, K.; Wan, F. A novel two-stage dissolved gas analysis fault diagnosis system based semi-supervised learning. *High Volt.* **2022**, *7*, 676–691. [[CrossRef](#)]
16. Rao, S.; Zou, G.; Yang, S.; Barmada, S. A feature selection and ensemble learning based methodology for transformer fault diagnosis. *Appl. Soft Comput.* **2024**, *150*, 111072. [[CrossRef](#)]
17. Xu, G.; Zhang, M.; Chen, W.; Wang, Z. Transformer fault diagnosis utilizing feature extraction and ensemble learning model. *Information* **2024**, *15*, 561. [[CrossRef](#)]
18. Wang, L.; Chi, J.; Ding, Y.; Yao, H.; Guo, Q.; Yang, H. Transformer fault diagnosis method based on SMOTE and NGO-GBDT. *Sci. Rep.* **2024**, *14*, 7179. [[CrossRef](#)] [[PubMed](#)]
19. Zhang, L.; Sheng, G.; Hou, H.; Zhou, N.; Jiang, X. An adaptive fault diagnosis method of power transformers based on combining oversampling and cost-sensitive learning. *IET Smart Grid* **2021**, *4*, 623–635. [[CrossRef](#)]
20. Dladla, V.M.N.; Thango, B. Fault classification in power transformers using dissolved gas analysis and optimized machine learning algorithms. *Machines* **2025**, *13*, 742. [[CrossRef](#)]
21. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [[CrossRef](#)]
22. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR, USA, 2–4 August 1996; pp. 226–231.
23. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [[CrossRef](#)]
24. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [[CrossRef](#)]
25. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, QC, Canada, 20–25 August 1995; pp. 1137–1145.
26. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.