

Article | Received 28 June 2026; Revised 6 September 2026; Accepted 28 September 2026; Published 10 October 2026
<https://doi.org/10.55092/aimat20260013>

Evidence-hierarchy audit of public DGA benchmarks for trustworthy AI-assisted transformer assessment



Yebin Zou¹, Richang Xian^{1,*}, Yiming Qin¹, Zihao Chu², Tao Yu³ and Chuancai Jing³

¹ College of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China

² STEM College and Royal Melbourne Institute of Technology University, Melbourne, Australia

³ Shandong Huineng Electric Co., Ltd., Zibo, China

* Correspondence author; E-mail: xianrc@sdut.edu.cn.

Highlights:

- A DGA evidence hierarchy separates benchmark agreement from field diagnostic validity.
- Exact five-gas fingerprints show that alan-456 is not independent of the Kaggle benchmark.
- MCS audits whether positive SHAP evidence follows class-specific DGA mechanisms.
- Leakage-sensitive validation shows lower Macro-F1 under transformer-grouped splitting.
- Confidence-MCS review logic flags reportable and expert-review cases.

Abstract: Dissolved gas analysis (DGA) is widely used for condition assessment of oil-immersed power transformers, but public-data artificial intelligence (AI) studies can overstate evidential reliability when labels are heterogeneous, samples overlap across benchmarks, or random splits ignore asset and time structure. This study presents an evidence-hierarchy audit framework for trustworthy AI-assisted transformer DGA assessment. The framework separates public-label benchmark agreement, transparent weak-label reproduction, leakage-sensitive validation, and future field-confirmed validity; combines exact gas-label fingerprinting, tree-ensemble baselines, Shapley additive explanations (SHAP) attribution, and a Mechanism-Consistency Score (MCS); and uses confidence-MCS logic as an expert-review trigger. On the Kaggle five-gas benchmark, LightGBM, Random Forest, and XGBoost achieved mean Macro-F1 values of 0.8982, 0.8884, and 0.8878, while exact fingerprinting showed that all 2321 alan-456 samples overlapped with Kaggle-derived fingerprints. After overlap removal, cross-dataset Macro-F1 decreased to 0.4688–0.5015. On the Lewis weak-label data, transformer-grouped validation reduced Macro-F1 from 0.9992 to 0.8691, showing sensitivity to asset-level splitting assumptions. MCS analysis identified class-dependent mechanism consistency and low-MCS review signals under five-gas inputs. The results support public-benchmark reliability and evidence-quality auditing rather than field-confirmed diagnostic performance.

Keywords: dissolved gas analysis; power transformer; trustworthy artificial intelligence; weak labels; benchmark reliability; SHAP; Mechanism-Consistency Score



Copyright©2026 by the authors. Published by ELSP. This work is licensed under Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

1. Introduction

Power transformers are high-value critical assets in transmission and distribution networks, and their failures can seriously affect supply reliability, system safety, and economic performance [1]. Dissolved gas analysis (DGA) detects gases generated in oil under electrical stress, thermal stress, and ageing, providing a convenient and physically grounded basis for assessing transformer insulation condition [2–4]. Oil-condition guidance and DGA-based asset-assessment studies further support the use of gas evidence in transformer condition evaluation [5–7]. However, gas patterns and assessment classes do not have a strict one-to-one correspondence [8,9].

Machine learning can model nonlinear relationships among DGA gas concentrations and ratios. Public datasets support algorithm reproduction and cross-study comparison, but they also introduce methodological risks [10,11]. If label sources, device identities, temporal structure, and dataset overlap are unclear, high random-split performance may reflect reproduction of rule-derived labels, correlation among repeated samples from the same device, or overlap among public samples rather than evidence of independent field validity.

A second issue is explanation quality. Shapley additive explanations (SHAP) values and feature-importance rankings indicate which variables influence model output [12,13], but the attributions do not by themselves show whether the evidence supporting a prediction is physically plausible [14–16]. In industrial condition monitoring, experts need to know not only which features matter, but also whether the positive evidence chain supporting the final assessment class is compatible with domain mechanisms, and whether the output should be accepted, rejected, or sent for review.

Using transformer DGA as a case of trustworthy assessment for weakly labelled industrial AI, this study proposes an evidence-hierarchy assessment framework. The framework separates evidence from public-label benchmarks, transparent weak-label reproduction, and future field-confirmed validation. It combines public-benchmark overlap auditing, leakage-sensitivity validation, and Mechanism-Consistency Score (MCS)-based explanation auditing. These components complement engineering DGA practice by giving public-data conclusions a reproducible, bounded, and auditable evidential basis, while preparing the method for later field-confirmed diagnostic evaluation.

From an AI-method perspective, the proposed framework shifts the problem from isolated classification to evidence-level auditable learning. In this formulation, the classifier is not evaluated only by its predicted assessment category; each prediction is accompanied by metadata describing label provenance, validation protocol, supportive attribution evidence, mechanism consistency, and review-trigger status. This design turns post-hoc explanation into a prediction-level audit variable and makes the public benchmark itself an object of reliability assessment rather than merely a source of accuracy values.

2. Materials and methods

2.1. Weak-label protocol

In the Lewis workflow, weak labels were defined as transparent operating-assessment classes derived from gas severity and key-gas rules, rather than maintenance-confirmed fault states [17,18]. The rules

were executed in priority order to handle overlap among discharge, partial-discharge-like (PD-like), thermal-fault, abnormal, and normal patterns [1–4].

Table 1 lists the trigger conditions, threshold logic, priority, and conflict handling of the weak-label rules. These rules are not intended to provide field ground truth. Instead, they make the Lewis weak-label workflow reproducible, auditable, and replaceable, which helps distinguish rule-reproduction ability from true field diagnostic validity.

Table 1. Reproducible weak-label protocol for the Lewis seven-gas workflow. $\varepsilon = 1\text{e-}6$ was used to avoid division by zero.

Class	Trigger condition	Threshold logic	Priority	Conflict handling
Discharge	Elevated C_2H_2 or abnormal acetylene/ethylene ratio	$\text{C}_2\text{H}_2 \geq 35 \text{ ppm}$, or $\text{C}_2\text{H}_2 \geq 5 \text{ ppm}$ and $\text{C}_2\text{H}_2 / (\text{C}_2\text{H}_4 + \varepsilon) \geq 0.5$	1	Assigned first when triggered
PD-like	H_2 -dominant pattern	$\text{H}_2 \geq 100 \text{ ppm}$ and $\text{H}_2 \geq 2 \times \max(\text{CH}_4, \text{C}_2\text{H}_4, \text{C}_2\text{H}_6, \varepsilon)$	2	Assigned if Discharge is not triggered
Thermal	Hydrocarbon gases or CO elevated under thermal-fault conditions	$\text{C}_2\text{H}_4 \geq 50 \text{ ppm}$, or $\text{C}_2\text{H}_6 \geq 65 \text{ ppm}$, or $\text{CH}_4 \geq 120 \text{ ppm}$, or $\text{CO} \geq 350 \text{ ppm}$	3	Evaluated after Discharge and PD-like
Abnormal	Abnormal overall gas severity	$\text{TDCG} \geq 720 \text{ ppm}$, or $\text{CO}_2 \geq 2500 \text{ ppm}$	4	Assigned when no more specific rule is triggered
Normal	No abnormal rule triggered	None of the above conditions are met	5	Default operating class

2.2. Feature construction

The input vectors consisted of gas concentrations and DGA mechanism-derived features [1,2]. Seven-gas records contained H_2 , CH_4 , C_2H_6 , C_2H_4 , C_2H_2 , CO , and CO_2 . Five-gas records contained only H_2 , CH_4 , C_2H_6 , C_2H_4 , and C_2H_2 , and therefore could not evaluate cellulose-degradation evidence related to CO and CO_2 [5,7]. In this study, seven-gas total dissolved combustible gas (TDCG_7), total hydrocarbons (THC), gas-ratio features, and five-gas TDCG_5 were calculated using Equations (1)–(4), respectively.

$$\text{TDCG}_7 = \text{H}_2 + \text{CH}_4 + \text{C}_2\text{H}_6 + \text{C}_2\text{H}_4 + \text{C}_2\text{H}_2 + \text{CO} \quad (1)$$

$$\text{THC} = \text{CH}_4 + \text{C}_2\text{H}_6 + \text{C}_2\text{H}_4 + \text{C}_2\text{H}_2 \quad (2)$$

$$r_i(a, b) = \frac{g_{i,a}}{g_{i,b} + \varepsilon} \quad (3)$$

$$\text{TDCG}_5 = \text{H}_2 + \text{CH}_4 + \text{C}_2\text{H}_6 + \text{C}_2\text{H}_4 + \text{C}_2\text{H}_2 \quad (4)$$

Here, TDCG_7 denotes total dissolved combustible gas under the seven-gas condition. CO_2 is not combustible and is therefore excluded from TDCG_7 in Equation (1), but it can still provide auxiliary evidence of cellulose ageing or thermal degradation in solid insulation. THC denotes the total hydrocarbon concentration in Equation (2). In Equation (3), $r_i(a, b)$ denotes the concentration ratio between gases a and b in sample i, and $g_{i,a}$ and $g_{i,b}$ are the corresponding gas concentrations. $\varepsilon = 1\text{e-}6$ was used to avoid division by zero. For five-gas records, CO was unavailable, so TDCG was calculated as TDCG_5 in Equation (4).

2.3. Model training and validation

Light Gradient Boosting Machine (LightGBM) and Extreme Gradient Boosting (XGBoost) were used as boosted-tree classifiers [19,20]. Random Forest (RF) and support vector machine (SVM) with a radial basis function (RBF) kernel were included as benchmark models [21,22]. SVM inputs were standardized before training using the scikit-learn workflow [23]. These baselines cover tree-ensemble and statistical-learning methods commonly used for nonlinear classification [24,25]. For SHAP-stability analysis on Kaggle, eight 80/20 stratified splits were used with seeds 3, 7, 13, 31, 42, 57, 73, and 101. For the audited benchmark summary, the tree-ensemble rows use the ten-run source table retained with the revision files, whereas the RBF-SVM row is reported as the single standardized reference run used in the figure-level baseline. Lewis Tier 1 used a 75/25 stratified split. Tier 2 used fivefold StratifiedGroupKFold grouped by transformer identity. Tier 3 used the first 75% of temporal records for training and the final 25% for testing. Confidence and MCS thresholds were used only as examples of review triggers, not as optimized deployment thresholds.

2.4. SHAP and Mechanism-Consistency Score

For model output $f_c(x_i)$ for class c , SHAP decomposes the prediction into a baseline term and feature contributions, as shown in Equation (5) [12,13]:

$$f_c(x_i) = E[f_c(X)] + \sum_{j=1}^M \varphi_{i,j,c} \quad (5)$$

MCS measures the fraction of positive evidence for class c contributed by the mechanism-supported feature set $M(c)$, as defined in Equation (6):

$$MCS_{i,c} = \frac{\sum_{j \in M(c)} \max(0, \varphi_{i,j,c})}{\sum_{j=1}^M \max(0, \varphi_{i,j,c}) + \varepsilon} \quad (6)$$

Only positive SHAP values were used to calculate MCS because they support the current assessment class [12,13]. Negative SHAP values represent counter-evidence and are not part of the supporting rationale chain [15,16]. The mechanism-supported feature set $M(c)$ was specified a priori before the SHAP calculation from DGA gas-generation logic and from the five-gas feature space available in the Kaggle benchmark; it was not optimized to maximize MCS. Correlated raw and derived features were retained when they represented complementary diagnostic expressions of the same gas-generation mechanism rather than independent causal variables. A low-MCS sample was defined as $MCS < 0.5$. MCS is therefore an auditability metric, not a causal metric, attribution-similarity metric, or substitute for classification performance [14]. Table S1 reports the exact class-specific feature sets used for Equation (6).

2.5. AI-method innovation: evidence-level auditable learning

The methodological contribution of this study is the conversion of public DGA assessment into a structured trustworthy-AI learning problem. For sample i , the gas-feature vector x_i and assessment label y_i are augmented by three evidence variables: the label-provenance state s_i , the validation protocol v_i , and the class-conditional explanation vector φ_i . The state s_i distinguishes public assembled labels, rule-derived weak labels, and future maintenance-confirmed labels; v_i distinguishes random,

transformer-grouped, chronological, and external field-confirmed validation; ϕ_i describes the model evidence supporting the predicted assessment category.

Under this representation, the AI workflow is composed of four coupled operators. The label-provenance operator P maps each sample to an evidence tier and prevents weak-label reproducibility from being interpreted as field validity. The leakage-aware partitioning operator L controls sample dependence caused by asset identity, time ordering, and public benchmark overlap. The tree-ensemble prediction operator F learns nonlinear gas-pattern representations from concentrations and ratio features. The explanation-audit operator A converts class-conditional SHAP values into positive supportive evidence and evaluates whether that evidence is compatible with mechanism-supported DGA feature sets.

This architecture places an explicit audit layer around the classifier. Consequently, classification performance, label credibility, leakage sensitivity, mechanism-compatible evidence, and expert-review decisions can be reported as separate but linked outputs. The AI innovation therefore lies in the evidence-audit formulation of weak-label industrial assessment, elevating the contribution from model selection to evidence-governed trustworthy industrial-AI evaluation.

$$D_i = \text{Report, if } p_i \geq \tau_p \text{ and } MCS_i, \hat{y}_i \geq \tau_m; \text{ otherwise Expert Review} \quad (7)$$

Equation (7) defines a selective-reporting rule, where p_i is the model confidence for the predicted class, τ_p is the confidence threshold, τ_m is the MCS threshold, and D_i is the resulting reporting or expert-review decision. The thresholds are not optimized deployment cut-offs in the present public-data study. Instead, they operationalize the requirement that a directly reportable AI-assisted DGA output should satisfy both statistical confidence and mechanism-compatible supportive evidence. Results derived from Equation (7) therefore quantify benchmark-label agreement under different coverage levels, review rates, and review burdens rather than clinical or utility-level deployment performance.

3. Public benchmark evaluation and evidence audit

3.1. Evidence hierarchy and datasets

The experiments were organized around three public datasets with different evidential roles. The Kaggle DGA dataset served as the main five-gas public-label benchmark. The Lewis dataset provided seven-gas time-series records with transformer identities and timestamps for weak-label workflow reproduction and leakage-sensitivity analysis [26]. The alan-456 dataset was used for public-benchmark overlap auditing and cross-dataset sensitivity testing. Because these datasets differ in label provenance, gas completeness, metadata availability, and overlap status, Table 2 defines the evidential boundary of each data source before model results are interpreted. Figure 1 then presents the evidence-hierarchy audit framework. Figure 1a shows the workflow from public DGA data, label-source control, leakage-sensitive splitting, tree-ensemble evaluation, SHAP evidence extraction, and MCS review. Figure 1b distinguishes the evidential strength associated with random splits, transformer-grouped splits, temporal splits, and future field-confirmed validation. This structure allows later performance results, explanation results, and validation boundaries to be interpreted within one evidential system [10,11].

Table 2. Public DGA data sources, gas completeness, label provenance, and evidential boundaries.

Dataset	Gas completeness	Label source	Use in this study	Evidential boundary
Kaggle DGA	H ₂ , CH ₄ , C ₂ H ₆ , C ₂ H ₄ , C ₂ H ₂	Publicly compiled labels	Main public-label benchmark; SHAP/MCS analysis	Public-label agreement only; CO and CO ₂ unavailable; source checksums listed in Data availability
Lewis <i>et al.</i> 2022	H ₂ , CH ₄ , C ₂ H ₆ , C ₂ H ₄ , C ₂ H ₂ , CO, CO ₂	Transparent rule-derived weak labels	Weak-label workflow and leakage-sensitivity analysis	Rule-label reproduction; maintenance-confirmed outcomes unavailable
alan-456	H ₂ , CH ₄ , C ₂ H ₆ , C ₂ H ₄ , C ₂ H ₂	Publicly compiled labels	Benchmark overlap and cross-dataset sensitivity audit	Not independent after exact overlap with Kaggle-derived fingerprints

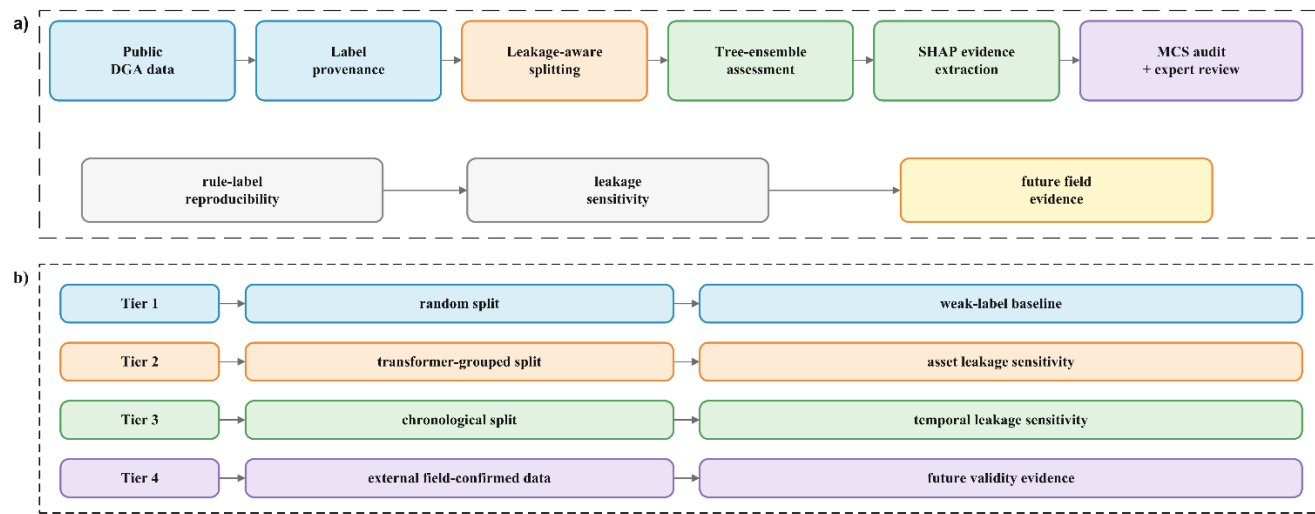


Figure 1. Evidence-hierarchy audit framework for trustworthy AI-assisted DGA assessment.

(a) Workflow from public DGA data, label-source control, leakage-sensitive splitting, tree-ensemble evaluation, SHAP evidence extraction, and MCS review; **(b)** Four validation tiers formed by random splits, transformer-grouped splits, temporal splits, and external field-confirmed data.

3.2. Public-label benchmark performance

On the Kaggle five-gas public-label benchmark, the audited benchmark table showed that tree-ensemble models achieved higher benchmark-label agreement than the RBF-SVM reference baseline. To evaluate both model performance and explanation stability, the TreeSHAP stability of LightGBM was reported together with multi-model classification results [27–29]. Table 3 shows that the top five SHAP-dominant features were identical across eight random seeds, and the coefficients of variation for the mean SHAP values of the main features were below 10%. This indicates that the explanation evidence was insensitive to random initialization. Table 4 further shows that LightGBM achieved the highest mean Macro-F1, while Random Forest and XGBoost maintained similar tree-ensemble performance. Therefore, LightGBM and XGBoost were treated as complementary tree-ensemble audit workflows rather than as competing field-diagnostic claims. Table 3 reports the LightGBM cross-seed TreeSHAP stability check, whereas the subsequent SHAP/MCS and selective-reporting analyses use the retained XGBoost source outputs.

This separation keeps the explanation-stability and MCS evidence streams traceable to their original source records.

Table 3. Cross-seed stability of LightGBM-TreeSHAP feature importance (eight random seeds).

Feature	Mean SHAP	Coefficient of variation (%)	Mean rank	Top-5 occurrences
C ₂ H ₂ /CH ₄	1.060	5.2	1.6	8/8
C ₂ H ₂ /C ₂ H ₄	1.040	8.2	1.6	8/8
CH ₄ /H ₂	0.917	4.8	2.8	8/8
C ₂ H ₄ /C ₂ H ₆	0.797	3.7	4.0	8/8
C ₂ H ₄	0.679	3.8	5.0	8/8
TDCG	0.577	5.7	6.1	0/8
H ₂	0.516	7.7	6.9	0/8
C ₂ H ₆	0.395	9.8	8.5	0/8
C ₂ H ₂	0.376	7.2	9.1	0/8
CH ₄	0.346	11.9	9.4	0/8

Table 4. Repeated random-seed model performance on the Kaggle five-gas public-label benchmark.

Model	Mean Accuracy	Standard deviation of Accuracy	Mean Macro-F1	Standard deviation of Macro-F1	Number of runs
LightGBM	0.9099	0.0080	0.8982	0.0089	10
Random Forest	0.8993	0.0097	0.8884	0.0111	10
XGBoost	0.9012	0.0098	0.8878	0.0104	10
RBF-SVM	0.4345	NA	0.4140	NA	1

To avoid comparison only among machine-learning models, transparent DGA rule baselines were also included. The International Electrotechnical Commission (IEC) ratio method and Rogers ratio method were approximated using five-gas ratios [2,3], and samples outside rule coverage were treated as Unknown. Table 5 reports both full-sample results and classified-subset results. This separates two types of information: the decision coverage with which a rule method can assign an effective class, and the class-matching ability on samples that the rule method classifies. The table shows that traditional ratio rules have clear engineering meaning, but under a public five-gas, seven-class label system they have limited coverage and constrained class mapping.

Figure 2 places machine-learning models [19–21,24] and traditional DGA rule baselines [2,3] in the same comparison framework. Figure 2a shows that the Macro-F1 scores of LightGBM, XGBoost, and Random Forest were substantially higher than that of RBF-SVM, indicating that tree ensembles were better suited to nonlinear class boundaries in five-gas DGA features. Figure 2b shows a tradeoff between decision coverage and full-sample Macro-F1 for the IEC and Rogers rules on both Kaggle and alan-456. Although rule methods are physically interpretable, they cannot assign a mappable seven-class decision to every sample. The comparison therefore does not simply aim to show that machine learning outperforms rule methods. Instead, it brings classification performance, rule coverage, and evidential interpretability into a unified public-benchmark reliability assessment.

Table 5. Results of traditional DGA ratio-rule baselines on public five-gas datasets.

Dataset	Rule method	Coverage	Full-sample Accuracy	Full-sample Macro-F1	Classified-subset Accuracy	Classified-subset Macro-F1	n
Kaggle	IEC ratio method	0.8186	0.4366	0.4500	0.5334	0.5121	3397
Kaggle	Rogers ratio method	0.6773	0.4354	0.4743	0.6428	0.5407	2811
alan-456	IEC ratio method	0.8574	0.3628	0.4493	0.4231	0.4989	1990
alan-456	Rogers ratio method	0.5924	0.3529	0.4606	0.5956	0.5133	1375

Note: Classified-subset results are calculated only for samples assigned a mapped class by the rule method; n is the number of those classified samples. Coverage is the proportion of all samples assigned a mapped class. Unknown decisions are treated as errors in the full-sample results.

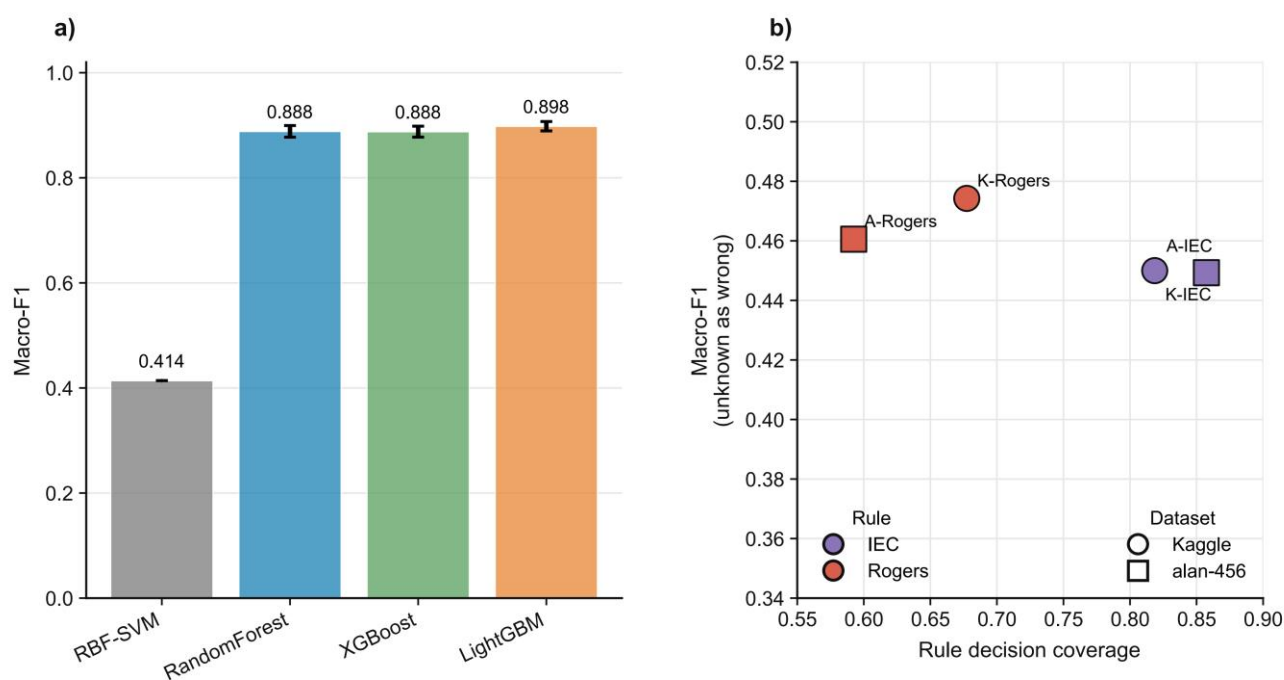


Figure 2. Public-benchmark performance comparison between machine-learning models and traditional DGA ratio-rule baselines. **(a)** Macro-F1 comparison of RBF-SVM, Random Forest, XGBoost, and LightGBM on the Kaggle public benchmark; **(b)** Relationship between decision coverage and full-sample Macro-F1 for IEC and Rogers ratio rules on Kaggle and alan-456, with Unknown treated as an error.

3.3. Benchmark overlap and leakage sensitivity

To determine whether alan-456 could serve as an independent external test set, an exact gas-label fingerprint audit was first conducted for the Kaggle and alan-456 public five-gas datasets. Figure 3a shows the audit workflow. Sample fingerprints were constructed from the five DGA gas concentrations and their class labels after consistent gas-column ordering and numeric standardization. Gas concentrations were rounded to three decimal places before string matching; row-level duplicates were retained to preserve the public benchmark composition, while unique fingerprints and row-level overlaps were counted separately. Figure 3b quantifies the audit result: 100.0% of alan-456 samples had identical gas-label fingerprints in Kaggle, while 69.2% of Kaggle samples overlapped with alan-456. Based on

this finding, Table 6 compares models trained on alan-456 and tested on Kaggle before and after overlapping test fingerprints are removed. All alan-456 samples overlap with Kaggle, so the reverse direction leaves no overlap-free alan-456 test samples. The results show that, without public-benchmark overlap auditing, a nominal cross-dataset test may be mistaken for independent external validation and may overestimate evidence of generalization [10,11].

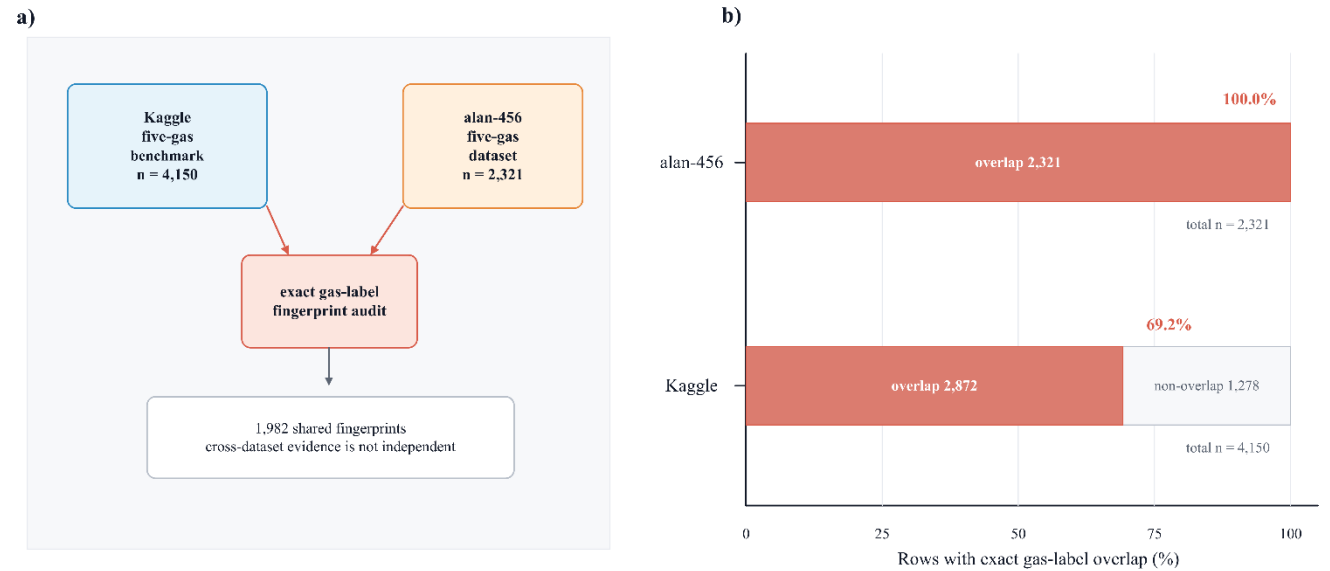


Figure 3. Exact fingerprint-overlap audit of public DGA benchmark datasets. **(a)** Five-gas label-fingerprint audit workflow between the Kaggle and alan-456 datasets; **(b)** Overlap proportions in alan-456 and Kaggle samples.

Table 6. Cross-dataset sensitivity before and after exact public-benchmark overlap removal.

Training source	Test source	Overlap handling	XGBoost Macro-F1	Random Forest Macro-F1	LightGBM Macro-F1
alan-456	Kaggle	No overlap removal (n = 4150)	0.8281	0.8404	0.8405
alan-456	Kaggle	Overlapping test fingerprints removed (n = 1278)	0.4688	0.5015	0.4781

Note: Both rows train on all 2321 alan-456 samples and test on Kaggle (4150 samples before filtering; 1278 afterward). Each result is from a single run with seed 42. Macro-F1 averages all seven encoded classes, assigning zero where a class F1 is undefined. The difference reflects overlap filtering and changed test-set composition, not independent field-confirmed validation.

The Lewis seven-gas weak-label data provided leakage-sensitivity evidence that differed from the public five-gas benchmark [26]. Table 7 compares three protocols by validation tier: random splitting, transformer-grouped splitting, and temporal splitting. Random splitting produced near-deterministic weak-label reproduction, whereas transformer-grouped and temporal splits both reduced performance, with the grouped folds showing a larger Macro-F1 dispersion across held-out transformer groups. Tier 2 reports unweighted fold-mean metrics and sample standard deviations from fivefold asset-grouped validation, rather than metrics calculated from pooled predictions or one temporal holdout. This result shows that repeated samples from the same transformer and temporal dependence can affect

performance estimates for public DGA models [10,11]. Accordingly, Table 7 should be interpreted as a protocol-sensitivity demonstration rather than as a strict head-to-head comparison of equivalent resampling designs. Reporting only sample-level random-split results is therefore insufficient to support stronger generalization claims.

Table 7. Leakage-sensitive validation tiers on the Lewis weak-label data.

Tier	Protocol	Test samples	Accuracy	Macro-F1	Notes
1	Stratified random split	2532	0.9988	0.9992	Weak-label reproduction
2	Transformer grouped fivefold	10126 out-of-fold evaluations	0.9846 ± 0.0140	0.8691 ± 0.1756	Device-level leakage sensitivity
3	Temporal split	2532	0.9609	0.9549	Forward-time weak-label reproduction

Note: Tier 2 reports the unweighted mean $\pm$ sample standard deviation across five folds (ddof = 1), not metrics calculated from pooled predictions. Its test-sample count is the sum of the five held-out fold sizes. Tiers 1 and 3 are single holdout results. Within each fold, Macro-F1 averages the classes present in the true or predicted labels; classes absent from both are excluded. Class composition can therefore differ across folds, limiting direct comparison across protocols.

3.4. Mechanism-constrained simulation scenario

To further simulate pre-field validation, a mechanism-constrained synthetic DGA scenario was constructed. The scenario contained 126 simulated transformers, 10 time points per transformer, and 1260 seven-gas samples. Each transformer was assigned a known latent fault state. Gas concentrations were generated jointly from DGA mechanism templates, device random effects, temporal severity trends, mixed-evidence perturbations, and measurement noise. Table 8 reports the performance of three splitting protocols in this synthetic scenario. The aim was to test protocol sensitivity to device dependence and temporal dependence under controlled conditions with known latent classes [10,11]. The simulation does not replace field-confirmed validation, but it provides a reproducible environment for stress-testing validation protocols when field data are limited.

Table 8. Pre-field simulation validation results in the mechanism-constrained synthetic DGA scenario.

Validation protocol	Training samples	Test samples	Accuracy	Macro-Precision	Macro-Recall	Macro-F1	Notes
Random sample split	944	316	0.9810	0.9821	0.9801	0.9809	Adjacent samples from the same transformer may appear in both training and test sets
Transformer-grouped split	920	340	0.9765	0.9763	0.9752	0.9755	Transformers in the test set are unseen during training
Temporal split	882	378	0.9630	0.9642	0.9574	0.9600	First 70% of time points for training and final 30% for testing

The simulation results showed that random sample splitting produced the highest Macro-F1, followed by transformer-grouped splitting and temporal splitting. This trend was consistent with the

leakage-sensitivity analysis on public data, indicating that evaluation protocols can affect performance estimates even in a synthetic scenario with known latent classes. Figure 4 presents the Lewis weak-label data and the mechanism-constrained synthetic scenario side by side. Figure 4a corresponds to real public weak-label time series, and Figure 4b corresponds to the controlled synthetic scenario. Together, the two sets of results show that random splitting tends to produce optimistic performance estimates, whereas grouped or temporal splits better expose evaluation risks caused by device dependence and temporal dependence.

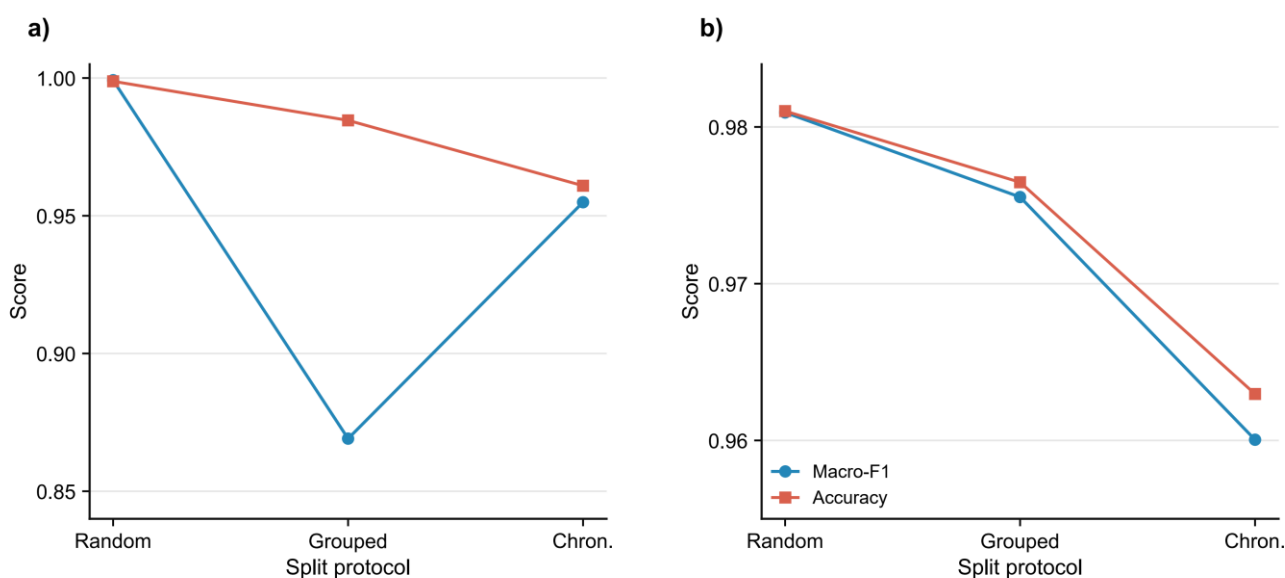


Figure 4. Leakage sensitivity and mechanism-constrained simulation validation results. **(a)** Changes in Accuracy and Macro-F1 on the Lewis weak-label data under random, transformer-grouped, and temporal splits; **(b)** Changes in Accuracy and Macro-F1 in the mechanism-constrained synthetic DGA scenario under the same splitting protocols. The grouped point in panel **(a)** represents the unweighted five-fold mean.

3.5. Mechanism-audited SHAP evidence

The SHAP results on the Kaggle benchmark showed that model discrimination did not mainly depend on single gas concentrations [12,13]. Instead, it relied more on gas-ratio features linked to DGA mechanisms [1–3]. This mechanism-oriented interpretation is consistent with Duval-type gas-pattern evidence [4,5]. Figure 5a shows that the largest global contributors were C_2H_2/CH_4 , CH_4/H_2 , C_2H_2/C_2H_4 , and C_2H_4/C_2H_6 , followed by TDCG, C_2H_4 , H_2 , and C_2H_2 . This indicates that acetylene-, hydrogen-, and ethylene-related evidence played major discriminative roles in the public five-gas benchmark [8,9]. Figure 5b further shows differences in evidence structure across assessment classes. The Normal class was mainly influenced by TDCG, the partial discharge (PD) class by CH_4/H_2 and H_2 , low-energy discharge (D1) and high-energy discharge (D2) classes by C_2H_2 -related ratios, and medium-temperature (T2) and high-temperature (T3) thermal-fault classes by C_2H_4/C_2H_6 and C_2H_4 . Table 9 maps these dominant features to DGA gas-generation mechanisms, showing that the model evidence structure was broadly consistent with typical gas-evolution patterns for partial discharge, discharge faults, and thermal faults. Feature contributions still overlapped across classes, indicating blurred boundaries in

public DGA classification. SHAP explanations should therefore be treated as mechanism-audit evidence, not as proof of physical causality [12–14].

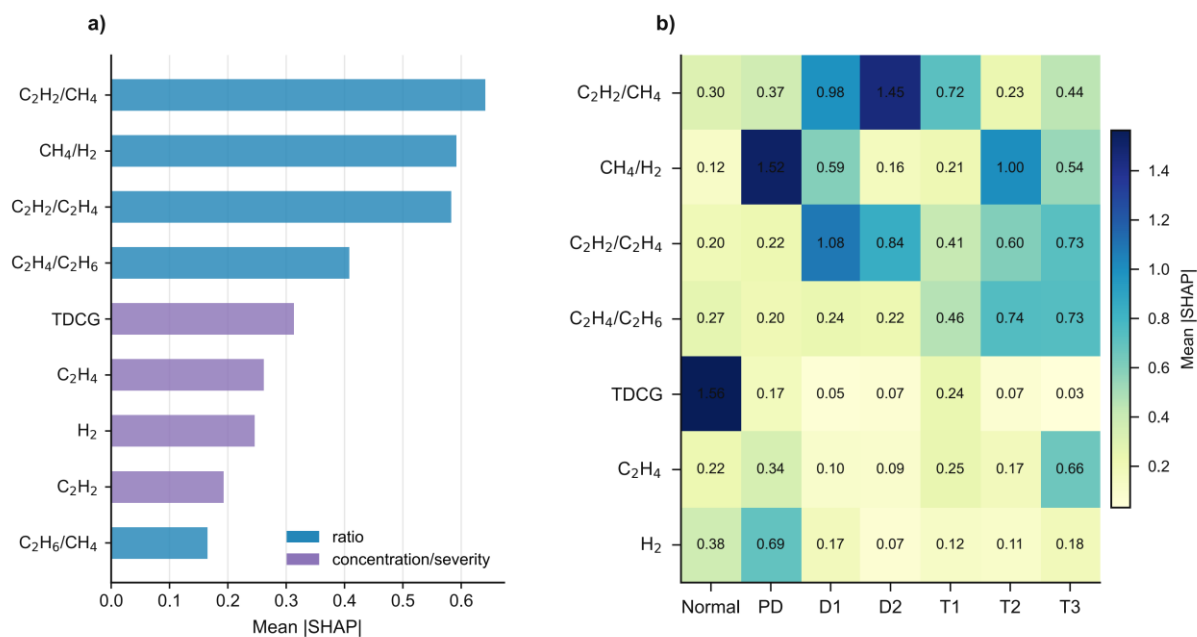


Figure 5. Global SHAP evidence and class-level feature-contribution patterns from the retained XGBoost-SHAP audit workflow. **(a)** Ranking of global mean absolute SHAP values, showing dominant gas-ratio and gas-concentration/severity features; **(b)** Class-level SHAP evidence heat map, showing the dominant attribution patterns for different assessment classes.

Table 9. Main global SHAP features of the XGBoost audit workflow on the Kaggle benchmark.

Rank	Feature	Mean SHAP (XGBoost test split)	DGA interpretation
1	C ₂ H ₂ /CH ₄	0.642	Acetylene relative to methane, supporting discharge discrimination
2	CH ₄ /H ₂	0.592	Methane-hydrogen relationship, supporting PD and low-energy discharge separation
3	C ₂ H ₂ /C ₂ H ₄	0.583	Evidence separating discharge-dominant and thermal-fault patterns
4	C ₂ H ₄ /C ₂ H ₆	0.408	Ethylene-ethane relationship associated with thermal-fault temperature level
5	TDCG	0.313	Overall combustible-gas severity
6	C ₂ H ₄	0.262	Evidence of thermal oil decomposition
7	H ₂	0.246	Evidence of low-energy ionization and discharge
8	C ₂ H ₂	0.193	Evidence of high-energy discharge or arcing
9	C ₂ H ₆ /CH ₄	0.165	Ethane-methane relationship supporting low-temperature thermal separation
10	CH ₄	0.128	Appears across several fault classes and has limited standalone discrimination

MCS was designed as an audit metric for supporting evidence, rather than as a causal metric or a conventional feature-importance metric [12,13]. It asks whether the positive SHAP evidence supporting an assessment class comes from gas features supported by DGA mechanisms [14]. For the XGBoost public-label workflow, Table 10 reports class-level mean MCS, dispersion, and low-MCS proportions to

quantify mechanism consistency across classes. Figure 6 further illustrates the review value of MCS. In Figure 6a, bars denote the mean MCS of each assessment class, and black error bars denote the corresponding standard deviation. Longer error bars indicate greater within-class variation in mechanism consistency and less stable model explanations for that class. Figure 6b shows the class-level relationship between the low-MCS proportion and error rate. Figure 6c compares mean MCS and model confidence between correctly and incorrectly predicted samples. Table S2 summarizes representative thresholds for Equation (7), showing the tradeoff between automatic reporting coverage, review burden, and accepted-case benchmark agreement. These results show that MCS extends SHAP attribution from general feature importance to an auditable signal of supporting evidence, providing experts with an auxiliary mechanism-review basis rather than a replacement for confidence or field-confirmed validation [15,16].

Table 10. Class-level Mechanism-Consistency Score in the Kaggle XGBoost workflow.

Class	Mean MCS	Standard deviation of MCS	Low-MCS proportion	Notes
Normal	0.7876	0.0509	0.0055	Low-gas and severity features support the Normal class in the public-label benchmark
PD	0.7779	0.1238	0.0375	Hydrogen-related support is present, with some contribution from hydrocarbon-ratio separation
D1	0.7514	0.1652	0.1060	Low-energy discharge evidence shows moderate dispersion
D2	0.8694	0.0696	0.0000	Acetylene and discharge evidence are relatively consistent
T1	0.5769	0.1039	0.1780	Thermal-boundary evidence is mixed under five-gas input
T2	0.8865	0.1030	0.0000	Thermal-fault evidence is relatively consistent
T3	0.5401	0.1499	0.3476	The lack of CO and CO ₂ limits interpretation of high-temperature thermal faults and paper-insulation evidence

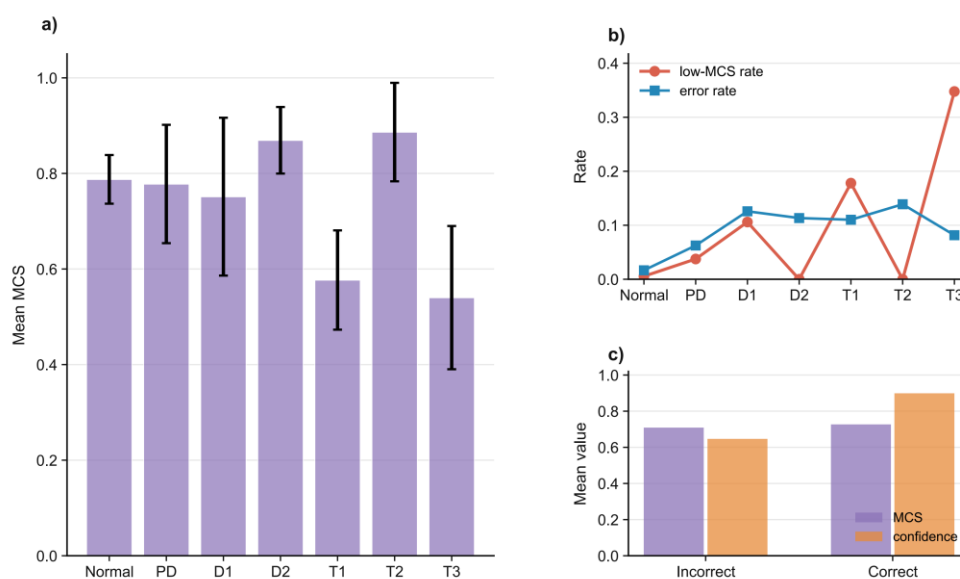


Figure 6. MCS-based analysis of class-level mechanism consistency and review signals. **(a)** Mean MCS and standard deviation for each class; **(b)** Class-level relationship between low-MCS proportion and error rate; **(c)** Comparison of mean MCS and model confidence between correct and incorrect predictions.

Representative Equation (7) threshold settings are reported in Table S2. The values show that confidence thresholding reduces accepted-case error as coverage falls, while MCS filtering adds a mechanism-review criterion. The complete threshold grid, bootstrap intervals, and class-wise accepted-sample metrics are retained in the supplementary source-data package. These results are based on public labels and should be interpreted as benchmark-label agreement under selective reporting, not as field deployment performance.

4. Discussion

The main contribution of this study is an evidence-hierarchy audit framework for public DGA benchmarks. The results show that public-benchmark performance cannot be interpreted from random-split accuracy alone [10,11]. Exact dataset overlap, repeated measurements from the same device, temporal dependence, and weak labels can all cause evidence for independent validation to be overestimated [17]. These risks are not unique to transformer assessment. They are common to weakly supervised industrial condition-monitoring tasks.

The framework also clarifies the role of machine-learning methods within established engineering DGA rule systems [1–3]. It also connects flexible classifiers with Duval-type gas-pattern interpretation [4,5]. In this framework, the machine-learning workflow supplements rule systems through evidence auditing and benchmark analysis. It reveals public-benchmark behaviour, quantifies sensitivity to data splitting and dataset overlap, and audits the evidence structure of flexible classifiers under explicit label-source conditions. This makes the evaluation process more methodologically transparent, especially for public-data studies in which field-confirmed labels are not yet available.

The framework reserves a fourth evidence tier for external field-confirmed validation. This tier can be advanced only when maintenance-confirmed field-label data are available from operating utilities. The present study is bounded by public datasets, and this tier is left for future field-collaboration studies.

MCS adds a reliability-oriented audit layer to standard SHAP analysis [12,13]. SHAP can explain model output, but statistically important attributions are not necessarily physically persuasive [14–16]. MCS evaluates whether the positive evidence chain supporting the final assessment class is compatible with DGA mechanisms. A high MCS indicates good consistency between the model's supporting evidence and expected gas-generation mechanisms. Its field validity still requires further verification with maintenance-confirmed labels and expert review. A low MCS can serve as a trigger for expert review, even when the model assigns a high classification probability.

The proposed MCS is also useful as a methodological bridge between attribution and auditability. A SHAP vector is an attribution object: it explains how features move a model output relative to a baseline. MCS transforms this attribution object into an evidence-quality signal by asking whether the positive evidence used to support a specific output comes from feature groups that are mechanistically meaningful for that output. This distinction is important in weak-label industrial AI because a model may be accurate with respect to proxy labels while relying on evidence that is unstable, leakage-prone, or physically weak. Similar evidence-audit logic can be transferred to other condition-monitoring tasks in which proxy labels, repeated assets, and incomplete field confirmation limit the interpretation of benchmark accuracy.

The main limitation of this study lies in the evidence tier, not in the algorithmic form itself. The public data used here do not yet include utility-level maintenance-confirmed outcomes, complete

transformer identities, timestamps, operating conditions, or oil-treatment history [26]. The added mechanism-constrained synthetic scenario can simulate latent fault states and a pre-field validation workflow, but it remains a simulation stress test. The next stage of field-confirmed diagnostic-performance evaluation should combine maintenance-confirmed labels, transformer-level grouping, forward temporal testing, external utility data, probability calibration, uncertainty estimation, and expert closed-loop review [29,30]. The label-source control, leakage-sensitivity validation, simulation stress testing, and mechanism-consistency audit provided here can serve as a preparatory methodological framework for moving public DGA benchmark studies toward field-confirmed diagnostic evaluation [10,11].

5. Conclusions

This study addressed the evidential reliability of public DGA benchmarks by constructing an evidence-hierarchy audit framework for trustworthy AI-assisted transformer assessment. Through label-source control, public-benchmark overlap auditing, leakage-sensitivity validation, and SHAP-MCS mechanism auditing, the study extends public-data experiments from single classification-accuracy reporting to systematic evidence-quality assessment. Methodologically, the framework integrates evidence-state augmentation, leakage-aware validation, class-conditional attribution auditing, and confidence-aware expert review into a reproducible trustworthy industrial-AI workflow.

The results show that tree-ensemble models can achieve high benchmark-label agreement on public five-gas data, but public-sample overlap, weak-label sources, and sample-level random splitting strongly affect how that performance should be interpreted. MCS further provides an auditable measure of whether supporting SHAP evidence is consistent with DGA gas-generation mechanisms. Future studies should use maintenance-confirmed labels, transformer-level metadata, and external field data to validate field-confirmed diagnostic performance, probability calibration, and expert closed-loop review.

Data availability statement

This study used public DGA datasets from the Kaggle DGA public data portal (<https://www.kaggle.com/datasets/luizflaviopereira/dga-dissolved-gas-analysis>), Lewis *et al.*'s Zenodo archive (<https://doi.org/10.5281/zenodo.5796243>), and the alan-456 transformer-fault-dataset GitHub repository (<https://github.com/alan-456/transformer-fault-dataset>), accessed on 4 September 2026. The processed analysis files were retained as data/raw/kaggle_dga_labeled_5gas.csv (SHA-256: 59E1852E52D2CEB8F5605CE20D617C2E04C87407AB177AF3268BD3A58B0E1132), data/raw/lewis_2022_dga_standard_weak_labels.csv (SHA-256: C3E9F0D2A83A0A4A995C15509F7608C710C30D1868A26C00D3C6C11B1DE66AA7), and data/raw/alan456_dga_labeled_5gas.csv (SHA-256: 17EECC602470F7BE6288028C1B7FDC042D81C420B281CE58967C8670CE7D3B27). The raw archive files were retained as data/external/kaggle_dga_dissolved_gas_analysis.zip (SHA-256: 3DD44B0D79BAAC310CB7273420C3165CAC764D6B613AA1496A1D1A7CF1182BF6), data/external/Lewis_et_al_2022_DGA_data.zip (SHA-256: B16186AA9D64B6262E863F812A479963ED7A72B493BFC9847E57D34071C65EB9), and data/external/alan456_transformer_fault_dataset/DGA-master.zip (SHA-256: 3D31852F38A726B99DFA35841F183F9F4DD0E9A108B357CB4BA45A3067DECCBF). Source tables and scripts for the reported benchmark, overlap, leakage, SHAP, MCS, and selective-reporting results are retained in the outputs/ and src/ folders of the revision package. The provenance audit table reports public dataset titles, repository

or uploader names, URLs, version metadata where available, access date, license status, retained source files, checksums, and preprocessing history. Public-label and weak-label metrics in this study quantify benchmark-label agreement or weak-label reproduction, not maintenance-confirmed field diagnostic validity. Supplementary material accompanies this article. The supplementary document contains Table S1, which specifies the class-specific mechanism-supported feature sets used to calculate MCS, and Table S2, which reports representative confidence-MCS selective-reporting results. The separate source-data and code archive supplied with the revised manuscript contains the provenance audit, analysis source tables, complete threshold grid, bootstrap intervals, class-wise accepted-sample metrics, and reproducibility scripts.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used ChatGPT (OpenAI), for English translation and language editing of the manuscript. The authors reviewed and edited the AI-assisted text and take full responsibility for the content of the manuscript.

Acknowledgments

The authors thank the providers of the public DGA datasets and the data-sharing platforms. This study received no external funding.

Authors' contribution

Conceptualization, Y.Z. and R.X.; methodology, Y.Z., R.X. and Y.Q.; software, Y.Z. and Y.Q.; validation, Y.Z., Z.C. and T.Y.; formal analysis, Y.Z. and Z.C.; investigation, Y.Z., Y.Q. and T.Y.; data curation, Y.Z. and Y.Q.; writing—original draft preparation, Y.Z.; writing—review and editing, R.X., Z.C., T.Y. and C.J.; visualization, Y.Z. and Z.C.; supervision, R.X.; project administration, R.X. and C.J. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

Tao Yu and Chuancai Jing are employees of Shandong Huineng Electric Co., Ltd. The authors declare that this affiliation does not influence the design, analysis, interpretation, or presentation of the results reported in this manuscript. The authors declare no other conflicts of interest.

References

- [1] IEEE. IEEE Std C57.104-2019: IEEE guide for the interpretation of gases generated in mineral oil-immersed transformers. 2019. Available: <https://standards.ieee.org/ieee/C57.104/7476/> (accessed on 4 September 2026).
- [2] IEC. IEC 60599:2022: mineral oil-filled electrical equipment in service—guidance on the interpretation of dissolved and free gases analysis. 2022. Available: <https://webstore.iec.ch/en/publication/66491> (accessed on 4 September 2026).
- [3] Rogers RR. IEEE and IEC codes to interpret incipient faults in transformers, using gas in oil analysis. *IEEE Trans. Electr. Insul.* 1978, EI-13(5):349–354.

- [4] Duval M. A review of faults detectable by gas-in-oil analysis in transformers. *IEEE Electr. Insul. Mag.* 2002, 18(3):8–17.
- [5] Duval M, dePablo A. Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases. *IEEE Electr. Insul. Mag.* 2001, 17(2):31–41.
- [6] 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(3):1007–1012.
- [7] IEC. IEC 60422:2013: mineral insulating oils in electrical equipment—supervision and maintenance guidance. 2013. Available: <https://webstore.iec.ch/en/publication/2110> (accessed on 4 September 2026).
- [8] Cheng L, Yu T. Dissolved gas analysis principle-based intelligent approaches to fault diagnosis and decision making for large oil-immersed power transformers: a survey. *Energies* 2018, 11(4):913.
- [9] Chen Z, Zhang X, Xiong H, Chen D, Cheng H, *et al.* Dissolved gas analysis in transformer oil using Pt-doped WSe₂ monolayer based on first principles method. *IEEE Access* 2019, 7:72012–72019.
- [10] Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans. Knowl. Discovery Data* 2012, 6(4):Article 15.
- [11] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 2023, 4(9):100804.
- [12] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In *Proceedings of Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, USA, December 4–9, 2017, pp. 4765–4774.
- [13] Lundberg SM, Erion G, Chen H, DeGrave AJ, Prutkin JM, *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2020, 2(1):56–67.
- [14] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 2019, 1(5):206–215.
- [15] Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, August 13–17, 2016, pp. 1135–1144.
- [16] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning, PMLR*, Sydney, Australia, August 6–11, 2017, pp. 3145–3153.
- [17] Northcutt C, Jiang L, Chuang I. Confident learning: estimating uncertainty in dataset labels. *J. Artif. Intell. Res.* 2021, 70:1373–1411.
- [18] Ratner AJ, De Sa CM, Wu S, Selsam D, Ré C. Data programming: creating large training sets, quickly. In *Proceedings of Advances in Neural Information Processing Systems 29 (NIPS 2016)*, Barcelona, Spain, December 5–10, 2016, pp. 3567–3575.
- [19] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA, August 13–17, 2016, pp. 785–794.
- [20] Ke G, Meng Q, Finley T, Wang T, Chen W, *et al.* LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, USA, December 4–9, 2017, pp. 3146–3154.
- [21] Breiman L. Random forests. *Mach. Learn.* 2001, 45(1):5–32.

- [22] Cortes C, Vapnik V. Support-vector networks. *Mach. Learn.* 1995, 20(3):273–297.
- [23] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 2011, 12:2825–2830.
- [24] Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 2001, 29(5):1189–1232.
- [25] Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
- [26] Lewis ZM, Wild JA, Allcock M. UK power station transformer dissolved gas analysis data (2010–2015). 2021. Available: <https://zenodo.org/records/5796243> (accessed on 4 September 2026).
- [27] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015, 10(3):e0118432.
- [28] Fawcett T. An introduction to ROC analysis. *Pattern Recognit. Lett.* 2006, 27(8):861–874.
- [29] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* 2009, 45(4):427–437.
- [30] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, PMLR*, Sydney, Australia, August 6–11, 2017, pp. 1321–1330.