Journal of Intelligent Decision Making and Information Science

www.jidmis.org
eISSN: 3079-0875



HIRA-DGA: Hierarchical Invariant Representations and Abstention for Cross-Source Power Transformer Fault Diagnosis

Alibek Adilbekov[1,4,*], Luigi Carlo De Jesus[2,4], Joe Robert Paul Lucena[2,4], Moldir Baigulbayeva[1], Ruslan Alnazirov[3], Sayat Moldakhmetov[1]

1Department of Power Engineering and Radio Electronics, M. Kozybayev North Kazakhstan University, 86 Pushkin Street, Petropavlovsk 150001, Republic of Kazakhstan
2Asia Pacific College, 3 Humabon Place, Magallanes, Makati City 1232, Philippines
3Department of Science, National Guard Academy of the Republic of Kazakhstan, 6 Kizatov Street, Petropavlovsk 150009, Republic of Kazakhstan
4Naveen Jindal Youth Fellowship Program, O.P. Jindal Global University, Sonipat, Haryana, India

ARTICLE INFO

Article history:
Received 01 April 2026
Received in revised form 25 April 2026
Accepted 20 May 2026

Keywords: Abstention; Dissolved gas analysis; Domain shift; Power transformer; Transferability

ABSTRACT

High within-dataset accuracy does not establish that a dissolved-gas analysis model will remain valid in an independently sourced transformer population, especially when gas magnitudes, case mix, and label structures differ. This study proposes HIRA-DGA, a Hierarchical Invariant-Representation and Abstention framework, and evaluates it using two public real-data sources: 200 Malaysian development records and 25 Ukrainian cases, of which 19 represented single mechanisms and six were mixed or composite. Five common gases were harmonized. HIRA-DGA separates locally calibrated normal/fault screening from portable fault typing: Stage 0 uses centered log-ratio composition plus total gas magnitude; Stage 1 uses three log gas-ratios for thermal-versus-discharge recognition; Stage 2 uses centered log-ratio composition for discharge severity; and Stage 3 returns HOLD for unsupported label structures. Duplicate-aware 10-by-5 grouped validation, and a one-way, no-target-tuning external challenge were used. Median total gas magnitude differed by 65 to 1,322 times across harmonized classes. With logistic regression, direct four-class absolute-log and ratio models achieved 2/19 and 7/19 external accuracy, whereas HIRA-DGA achieved 18/19 (94.7%; Wilson 95% CI 75.4-99.1%). Stage 2 centered log-ratio features classified 11/11 discharge-severity cases versus 5/11 for its absolute-log ablation. A fixed LightGBM robustness check showed the same architectural direction, improving from 5/19 for direct ratio classification to 13/19 for the hierarchical path. All six mixed cases were assigned HOLD. The contribution is a representation-by-decision technology architecture, not a new classifier: it localizes magnitude-sensitive screening, transfers scale-reduced chemistry and makes abstention an explicit engineering output.

1. Introduction

Power transformers are high-value assets whose electrical and thermal faults can evolve before visible failure. Dissolved-gas analysis (DGA) remains one of the most established diagnostic tools because the relative production of hydrogen, methane, ethane, ethylene, and acetylene reflects

* *Corresponding author.*
E-mail address: aeadilbekov@ku.edu.kz

different energy and fault mechanisms. IEC 60599 provides engineering guidance for interpreting dissolved and free gases, while modern research increasingly applies machine learning to automate classification, handle incomplete observations, improve interpretability, and combine multiple diagnostic signals [1-10].

The dominant research pattern, however, remains benchmark-oriented. Recent studies improve classifier accuracy, feature selection, class balancing, synthetic sample generation, or model explanation within one source population [2, 3, 6, 11-16]. High internal accuracy is useful but does not answer a deployment question: will the learned decision boundary still be meaningful when the model encounters a different utility, laboratory practice, transformer population, concentration range, or fault taxonomy?

That transfer problem is now recognized explicitly. Domain-adaptation research has begun to align transformer DGA distributions by using target-domain samples during adaptation, including feature-weighted MMD-CORAL methods that prioritize features with larger distribution discrepancies [17]. Such methods are important, but they solve a different operational setting: target data are available for adaptation. Utilities also face a zero-target-tuning setting in which a previously developed model is challenged by a new population before any local retraining is allowed. In that setting, the first requirement is not a more complex learner; it is a defensible architecture that decides which information may transfer and which information must remain locally calibrated.

A second gap concerns task structure. A single four-class classifier usually receives one feature representation and is expected to solve normality screening, fault-mechanism recognition, and severity discrimination simultaneously. These decisions do not have the same transfer requirement. Total gas magnitude may be informative for local normal/fault screening but may be highly source-dependent. Ratios and compositions reduce global scale and may be more suitable for transferring fault chemistry. Mixed or composite faults may not belong to the mutually exclusive label space at all. Treating these distinct problems as one classification surface can therefore export a local threshold and force unsupported labels.

This problem is directly relevant to computer information systems and industrial management because transformer-diagnostic models become operational decision-support systems when they integrate measurements, encode model authority, and route uncertain cases for human review. Dhini et al. [18] established a multistage precedent by organizing transformer DGA diagnosis into five nested stages and comparing SVM, ELM-RBF, and BPNN on 266 cases using an 80:20 holdout. Their published results reported 81.60% overall test accuracy for SVM with SMOTE and 83.46% for BPNN, with stage-level testing accuracy ranging from 62.50% to 100.00%. That design addressed within-source class imbalance and staged fault separation, but it did not test an independent external source or define an abstention action for mixed or unsupported labels. The present study therefore extends the multistage question from within-source classification to cross-source information-system qualification: which information should remain locally calibrated, which chemistry representation can be challenged across sources without target tuning, and when should automation return HOLD rather than force a class?

The study addresses five questions. First, how large is the magnitude shift between two real DGA sources after harmonizing the same three fault classes? Second, does removing absolute scale improve transfer relative to direct four-class classification? Third, does a hierarchical representation-by-decision architecture outperform one-shot transfer without target tuning? Fourth, does the architectural direction persist when a fixed nonlinear gradient-boosting learner replaces logistic regression? Fifth, can mixed or unseen fault descriptions be handled by abstention rather than a forced single label? The contributions are: (1) a quantified real-data cross-source shift analysis; (2) HIRA-DGA, an original four-stage local/portable/abstention architecture with explicit formulations and algorithmic logic; (3) duplicate-aware internal validation plus a one-way external challenge using independently sourced data; (4) a learner-family robustness check using LightGBM under the same protocol; and (5) a reproducible supplementary package containing the exact source-derived files, mappings, code, and result tables used in the article.

2. Methodology

2.1. Study Design, Source Files, and Data Provenance

The research was a retrospective secondary analysis of two independently sourced real-data collections. The Malaysian source was used for development and grouped internal validation. The Ukrainian source was reserved for a one-time external challenge. No external case was used to select representations, model settings, or pseudocounts.

The development source was the public S1 Table accompanying Illias and Zhao Liang [19]. The parent article states that 400 actual-site diagnoses were obtained from an electrical utility in Malaysia, but the downloadable supporting file contains 200 rows. This study analyzes only those 200 publicly released rows and does not reconstruct the unreleased cases. The exact supporting file is identified by DOI 10.1371/journal.pone.0191366.s001 [20]. It contains six gases and four labels: 100 no-fault, 36 thermal-fault, 32 low-intensity-discharge, and 32 high-intensity-discharge records. Because the external source contains five gases, carbon monoxide was excluded and the five common gases were retained.

The external source was the 25-transformer dataset deposited by Shutenko and Kulyk [21], and associated with their fault-recognition study [22]. The repository states that the cases were obtained in cooperation with electrical laboratories and energy companies in Ukraine and reports five gas concentrations in ppm together with identified fault descriptions. Nineteen descriptions represented one dominant mechanism that could be prespecified as thermal, low discharge, or high discharge. Six descriptions were mixed, composite, or outside the mutually exclusive development taxonomy and were reserved for the abstention test.

Exact equality across the five development gases was audited before modeling. The 200 rows contained 186 unique vectors; 28 rows formed 14 exact duplicate pairs, and no duplicate pair had conflicting labels. Duplicate rows were retained because they are part of the released source file, but identical vectors were assigned to the same group identifier so they could not cross training and test folds. Table 1 summarizes the source provenance, files, records, model inputs, and analytical roles.

Table 1

Data-source provenance, exact files, and analytical roles

Element	Malaysia development source	Ukraine external source
Primary reference	Illias and Zhao Liang [19]	Shutenko and Kulyk [22]
Persistent data identifier	PLOS S1 Table DOI: 10.1371/journal.pone.0191366.s001	Mendeley Data V1 DOI: 10.17632/x22y4s9fks.1
Source-derived file used	malaysia_dga_200.csv (derived from PLOS_DGA_S1.docx)	ukraine_dga_25.csv (transcribed and mapped from repository table)
Records	200 released rows; 186 unique five-gas vectors	25 cases: 19 single-mechanism, 6 mixed/unseen
Common model inputs	H2, CH4, C2H6, C2H4, C2H2	H2, CH4, C2H6, C2H4, C2H2
Analytical role	Development; grouped validation; stage training	No-target-tuning transfer challenge; abstention stress set

2.2. Gap-Driven Novelty and Architecture

The central methodological gap is not a shortage of classifiers. It is the absence of an explicit rule for deciding which DGA information should remain local, which representation should be transferred, and when the system should refuse a label. HIRA-DGA addresses this gap by decomposing the diagnostic problem according to transfer role. The name denotes Hierarchical Invariant-Representation and Abstention: hierarchical because the decision proceeds from screening to mechanism to severity; invariant-representation because the portable stages deliberately reduce total-magnitude dependence; and abstention because unsupported label structures are routed to HOLD. Table 2 compares the proposed architecture with the principal multistage and direct-classification alternatives.

Table 2

Quantitative and architectural comparison

Approach	Architecture and representation	Validation setting
Dhini et al. [18], multistage system	Five nested FDD stages; seven raw gases; SVM, ELM-RBF, and BPNN; SMOTE for imbalance.	266 cases; IEC TC 10 plus Jakarta normal data; 80:20 holdout; no independent external source.
Direct four-class baselines in this study	One classifier receives either absolute-log concentrations or three log gas-ratios and predicts normal/thermal/low/high.	Duplicate-aware 10x5 grouped validation; one-way Ukraine challenge; no target tuning.

Approach	Architecture and representation	Validation setting
Direct fault-only ratio comparator	Single three-class ratio model after fault status is already known; no stage-specific representation.	Same 19-case external single-mechanism challenge.
HIRA-DGA	S0 local CLR+total; S1 portable ratios; S2 portable CLR; S3 HOLD for mixed/composite/unseen structures.	Same grouped development protocol; one-way external challenge; six mixed cases reserved for abstention.

2.3. Data Transformations and Formulations

Let x = (x1,...,x5) denote the concentrations of H2, CH4, C2H6, C2H4, and C2H2. To avoid undefined logarithms for zero-valued gases, each gas-specific pseudocount delta_j was fixed from development data only as one-half of the smallest positive concentration for that gas. The positive replacement is defined in Equation 1.

$$x*_j = max(x_j, delta_j) \tag{1}$$

For the centered log-ratio (CLR) representation, the log concentration of each gas is centered by the mean log concentration of all five gases. CLR preserves relative composition and removes multiplicative total-scale effects, which is appropriate when the interest is gas pattern rather than absolute burden [23].

$$CLR_j(x) = ln(x*_j) - (1/5)SUM(k = 1to5)ln(x*_k) \tag{2}$$

Stage 0 also uses total gas magnitude because local normal/fault screening may depend on overall burden. The magnitude variable is defined in Equation 3.

$$m(x) = log10[1 + SUM(j = 1to5)x_j] \tag{3}$$

The portable mechanism stage uses three log gas-ratios related to established DGA chemistry. The vector is defined in Equation 4 using the same development-fixed pseudocounts.

$$r(x) = [ln(CH4*/H2*), ln(C2H2*/C2H4*), ln(C2H4*/C2H6*)] \tag{4}$$

The primary learner was class-weighted logistic regression with standardized inputs and a maximum of 5,000 iterations. A deliberately low-capacity learner was selected so that the main contribution would remain attributable to the architecture and representation choices rather than an extensive hyperparameter search. To test whether the architectural direction was an artifact of linear decision boundaries, the identical protocol was repeated with LightGBM, a nonlinear gradient-boosting method [24]. LightGBM settings were fixed before external evaluation: 80 trees, learning rate 0.05, maximum depth 3, seven leaves, minimum child size 5, L2 regularization 1.0, and balanced class weights. No target-domain data were used for learner selection or tuning. For the primary multinomial logistic model with K classes, the class probability is given by Equation 5.

$$P(y = k|z) = exp(beta_k^T z)/SUM(l = 1toK)exp(beta_l^T z) \quad (5)$$

2.4. HIRA-DGA Training and Deployment Algorithm

Stage 0 is a local normal-versus-fault screen. It uses the five CLR coordinates plus m(x), thereby preserving local magnitude information. This stage is calibrated and validated only on the Malaysian development population. Because the external source contains no verified normal cases and its magnitude scale differs markedly, the study does not claim that Stage 0 is externally validated.

Stage 1 is the portable mechanism classifier. Among development faults, it distinguishes thermal from discharge cases using r(x). Stage 2 is activated only for discharge cases and distinguishes low from high discharge using CLR(x). The portable stages therefore avoid the direct use of total gas magnitude. Stage 3 is the abstention gate: a case with a known single mechanism follows the hierarchy, whereas a mixed, composite, or unseen description receives HOLD rather than a forced single class. Figure 1 summarizes these local, portable, and abstention stages.

Algorithm 1. HIRA-DGA training and deployment. Training: (1) group exact duplicate development vectors; (2) fit S0 on all development rows using [CLR(x), m(x)] and labels normal/fault; (3) fit S1 on development fault rows using r(x) and labels thermal/discharge; (4) fit S2 on development discharge rows using CLR(x) and labels low/high. Deployment: (5) apply S0 only within the locally calibrated population; (6) for a confirmed fault entering portable typing, use S1; (7) if S1 predicts thermal, return thermal; otherwise use S2 and return low or high discharge; (8) when the case is mixed, composite, or outside the supported hierarchy, return HOLD and do not force a class.

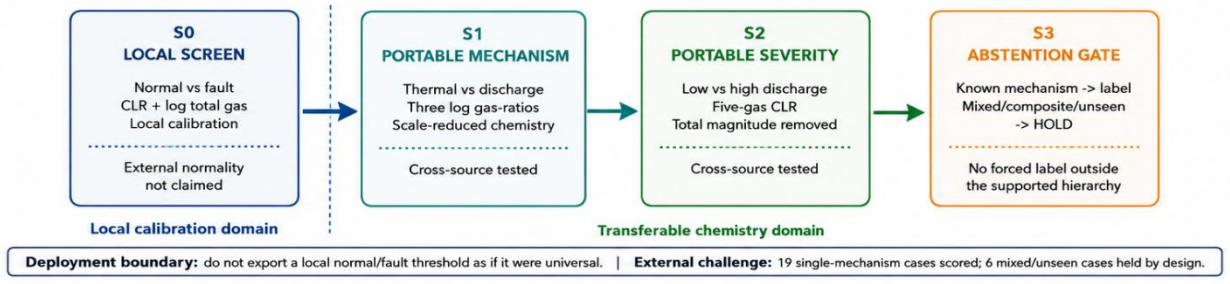


Fig. 1. HIRA-DGA architecture. Local abnormality screening is separated from portable fault typing, and unsupported label structures are routed to HOLD

2.5. Validation Design and External Label Harmonization

Internal validation used StratifiedGroupKFold with five folds and ten random seeds, giving 50 held-out folds per model. The grouping key was the exact five-gas vector, so duplicate measurements could never appear on both sides of a split. Direct four-class baselines used either absolute-log concentrations log10(1+x) or the three log gas-ratios. HIRA-DGA was also evaluated as a complete four-state pipeline inside the development population. Logistic regression was the primary transparent learner; the complete direct and hierarchical protocol was then repeated with the fixed LightGBM specification as a nonlinear robustness check.

External validation was strictly one-way. After internal evaluation, the models were fitted on all development rows and applied to the 19 prespecified single-mechanism external faults without tuning. The harmonization rule was fixed from the textual fault descriptions: T1, T2, and T3 were thermal; PD, critical PD, corona discharge, and sustained developed PD were low discharge; D2, high-intensity spark discharge, and extensive creeping discharge were high discharge. The full row-level mapping is included in the supplementary file external_predictions.csv. Six mixed or composite descriptions were not assigned a fabricated single ground-truth class and were evaluated only for the HOLD decision.

The external source contains no verified normal cases. Accordingly, the external challenge evaluates fault typing, not external normal/fault detection. To ensure a fair comparison, the study reports both one-shot four-class baselines and a direct fault-only three-class ratio classifier trained on the same development fault subset used by the portable HIRA stages.

2.6. Performance Measures, Statistical Tests, and Reproducibility

Internal performance was summarized by accuracy, balanced accuracy, and macro-F1 across 50 grouped folds. External accuracy and macro-F1 were reported with Wilson 95% confidence intervals for accuracy. Two stage ablations replaced the proposed portable representation with absolute-log concentrations while keeping the same logistic learner and training subset. Two paired exact binomial tests were prespecified as mechanism checks: direct four-class ratio versus final HIRA-DGA, and Stage 2 absolute-log versus centered log-ratio. Because these tests are exploratory evidence about why transfer changes rather than primary confirmatory hypotheses, both raw and Holm-adjusted p-values across the two comparisons are reported. The strongest representation ablation is Stage 2 because both models address exactly the same low-versus-high discharge task.

The analysis was implemented in Python using pandas, NumPy, SciPy, and scikit-learn. The supplementary package contains the exact source-derived CSV files, the original PLOS S1 supporting file, transformation and mapping notes, code, grouped validation outputs, external row-level predictions, and figures. No proprietary measurements and no reconstructed unreleased cases are included.

3. Results

3.1. Cross-Source Magnitude Shift

The first result is a data-property result, not a classifier result. After harmonizing thermal, low-discharge, and high-discharge classes, the median total of the five common gases differed dramatically between sources. The development-to-external median ratio was 97.5 for thermal faults, 65.0 for low-discharge faults, and 1,322.4 for high-discharge faults. Figure 2 shows the shift on a logarithmic scale.

This finding changes the modeling problem. A classifier that learns absolute magnitude can achieve excellent internal separation and still treat external faults as numerically similar to a local low-burden region. The result is therefore not interpreted as generic evidence that cross-source machine learning

fails. The safer interpretation is that raw-scale transfer is confounded by severe magnitude incompatibility, case-construction differences, and potentially different measurement or reporting contexts. HIRA-DGA is designed around this observation: magnitude remains local, while the portable stages emphasize relative chemistry.

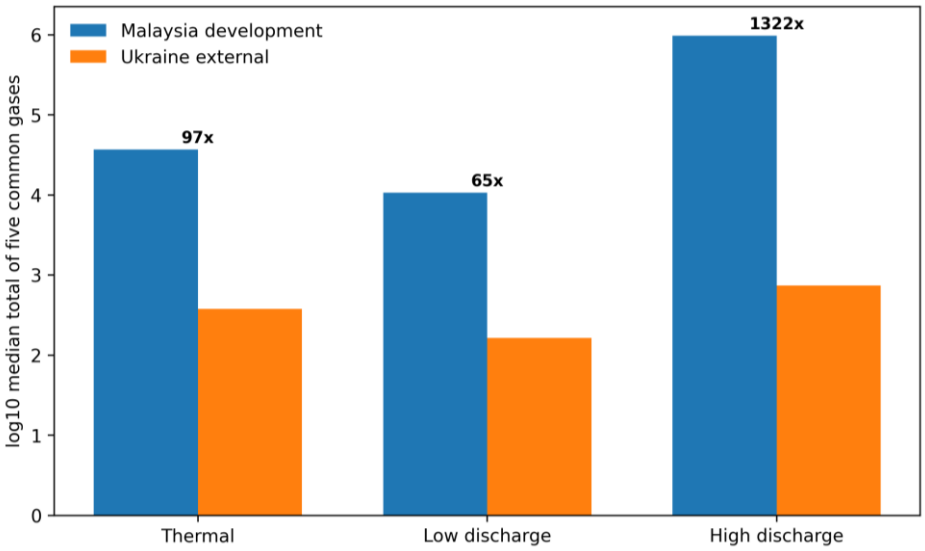


Fig. 2. Cross-source median total gas magnitude (log10 scale) by harmonized fault class. Labels show the development-to-external median ratios

3.2. Duplicate-Aware Internal Validation

Grouped validation prevented the 14 exact duplicate pairs from crossing folds. Table 3 shows that the primary logistic HIRA-DGA pipeline retained strong internal performance, with mean accuracy 0.951, macro-F1 0.954, and balanced accuracy 0.967. Stage 0 achieved 0.954 mean accuracy, while Stages 1 and 2 were perfectly separated across the grouped development folds. The nonlinear LightGBM check produced even higher internal scores: mean accuracy 0.984 for direct absolute-log classification and 0.983 for the full HIRA-DGA pipeline. These values should be interpreted cautiously because the subsets are modest and source-specific. More importantly, the high internal LightGBM scores did not guarantee external transfer, reinforcing the need for independent-source evaluation.

Table 3

Duplicate-aware internal validation of direct baselines and HIRA-DGA stages

Model / stage	Representation and task	Accuracy mean +/- SD	Macro-F1 mean +/- SD	Balanced accuracy mean +/- SD
LR: Direct 4-class ABS-LOG	log10(1+x); normal/thermal/low/high	0.943 +/- 0.044	0.943 +/- 0.041	0.952 +/- 0.031
LR: Direct 4-class RATIO	three log ratios; normal/thermal/low/high	0.913 +/- 0.049	0.917 +/- 0.045	0.941 +/- 0.034

Model / stage	Representation and task	Accuracy mean +/- SD	Macro-F1 mean +/- SD	Balanced accuracy mean +/- SD
LR: HIRA S0	CLR + total; normal/fault	0.954 +/- 0.037	0.953 +/- 0.037	0.954 +/- 0.037
LR: HIRA S1	three log ratios; thermal/discharge	1.000 +/- 0.000	1.000 +/- 0.000	1.000 +/- 0.000
LR: HIRA S2	CLR; low/high discharge	1.000 +/- 0.000	1.000 +/- 0.000	1.000 +/- 0.000
LR: HIRA full pipeline	hierarchical four-state pipeline	0.951 +/- 0.038	0.954 +/- 0.033	0.967 +/- 0.024
LightGBM: Direct 4-class ABS-LOG	same direct task and grouped splits	0.984 +/- 0.023	0.983 +/- 0.024	0.981 +/- 0.024
LightGBM: Direct 4-class RATIO	same direct task and grouped splits	0.978 +/- 0.032	0.978 +/- 0.032	0.984 +/- 0.025
LightGBM: HIRA full pipeline	same hierarchical representations and grouped splits	0.983 +/- 0.021	0.984 +/- 0.020	0.982 +/- 0.019

3.3. External Transfer and Abstention

The external challenge demonstrates why the decision structure matters. With logistic regression, the direct four-class absolute-log model classified only 2 of 19 faults correctly (10.5%). Replacing absolute magnitude with three gas-ratios improved the direct four-class result to 7 of 19 (36.8%). This shows that relative chemistry recovers some transferability, but the one-shot classifier still competes against a normal class whose boundary was learned only from the local development population.

A fair fault-only comparator removed that normal class and used the same three gas-ratios on the 19-fault challenge. Logistic-regression accuracy increased to 17 of 19 (89.5%). HIRA-DGA reached 18 of 19 (94.7%; Wilson 95% CI 75.4-99.1%). The one-case difference is not presented as proof that hierarchy is universally superior. The stronger result is structural: separating local screening from portable fault typing accounts for most of the transfer recovery, and assigning a different representation to discharge severity resolves an additional case without target tuning.

The stage-level ablation supports the representation-by-decision claim. For thermal-versus-discharge recognition, the ratio stage scored 18/19 while the absolute-log ablation scored 17/19. For low-versus-high discharge severity, centered log-ratio features scored 11/11 whereas the absolute-log

ablation scored 5/11. Six discordant cases improved and none worsened. The exact paired p-value was 0.03125, which remained 0.03125 after Holm adjustment across the two prespecified mechanism tests. Thus, the clearest empirical contribution is not a new logistic-regression classifier; it is the use of composition for a severity decision strongly distorted by source magnitude.

The nonlinear learner comparison in Figure 3 extends the robustness check. Under the same no-target-tuning protocol, direct four-class LightGBM achieved 0/19 with absolute-log inputs and 5/19 with ratios; the fault-only ratio comparator reached 15/19, and the hierarchical path reached 13/19. The hierarchy therefore improved substantially over the direct four-class LightGBM design, although it did not outperform the fault-only LightGBM comparator. This learner sensitivity is important: HIRA-DGA should be interpreted as a technology architecture that controls transfer roles and model authority, not as a guarantee that one classifier family will dominate. For the logistic primary analysis, HIRA-DGA improved 11 cases and worsened none relative to direct four-class ratio classification; the exact paired p-value was 0.00098 and the Holm-adjusted value was 0.00195. Table 4 reports the complete external-transfer, ablation, and abstention results.

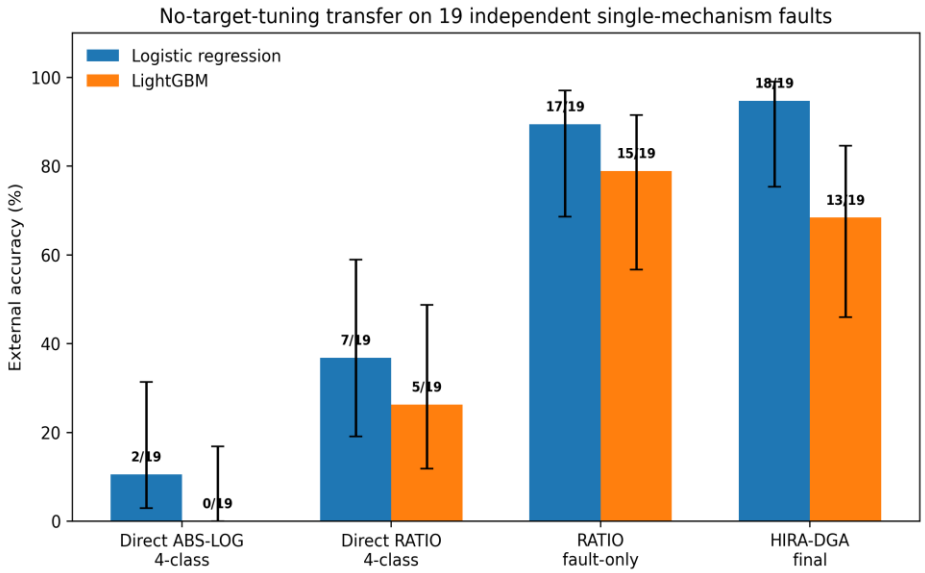


Fig. 3. External transfer without target-domain tuning for logistic regression and LightGBM. Error bars are Wilson 95% confidence intervals; values above bars are correct/total cases

Table 4

External transfer, stage ablations, and abstention outcomes

Method / decision	Task	Correct / n	Accuracy / outcome	Macro-F1 / interpretation
LR: Direct 4-class ABS-LOG	normal/thermal/low/high	2 / 19	0.105 (95% CI 0.029-0.314)	Macro-F1 0.167
LR: Direct 4-class RATIO	normal/thermal/low/high	7 / 19	0.368 (95% CI 0.191-0.590)	Macro-F1 0.369

Method / decision	Task	Correct / n	Accuracy / outcome	Macro-F1 / interpretation
LR: Direct fault-only RATIO	thermal/low/high	17 / 19	0.895 (95% CI 0.686-0.971)	Macro-F1 0.872
LR: HIRA-DGA final	thermal/low/high	18 / 19	0.947 (95% CI 0.754-0.991)	Macro-F1 0.941
LightGBM: Direct 4-class ABS-LOG	normal/thermal/low/high	0 / 19	0.000 (95% CI 0.000-0.168)	Macro-F1 0.000
LightGBM: Direct 4-class RATIO	normal/thermal/low/high	5 / 19	0.263 (95% CI 0.118-0.488)	Macro-F1 0.289
LightGBM: Direct fault-only RATIO	thermal/low/high	15 / 19	0.789 (95% CI 0.567-0.915)	Macro-F1 0.744
LightGBM: HIRA-DGA final	thermal/low/high	13 / 19	0.684 (95% CI 0.460-0.846)	Macro-F1 0.650
LR: HIRA S2 CLR	low/high discharge	11 / 11	1.000 (95% CI 0.741-1.000)	Primary representation
LR: S2 ABS-LOG ablation	low/high discharge	5 / 11	0.455 (95% CI 0.213-0.720)	Holm-adjusted p=0.03125
HIRA S3 abstention	mixed/composite/unseen	6 / 6	HOLD	No forced single label

The six remaining Ukrainian cases contained combined thermal and discharge descriptions, including T3+D2 and other mixed patterns. A conventional closed-set model would be forced to select one of the known classes even though the ground-truth description contains multiple mechanisms. HIRA-DGA instead assigned HOLD to all six by design. This is not counted as classification accuracy because no fabricated single target label was created.

4. Discussion

The abstention concept is important for industrial information systems because it defines the boundary of model authority. HOLD means that the single-mechanism hierarchy is structurally insufficient for the case; the next action is engineering review, multi-label analysis, complementary diagnostics, or local data collection. This differs from a low-confidence threshold. A model may be confident in a class that is absent from the true label structure. The abstention rule therefore depends on support and taxonomy, not probability alone.

The framework also clarifies what should happen after transfer failure. If a utility later releases labeled target data, a domain-adaptation method such as feature-weighted MMD-CORAL may be appropriate [17]. HIRA-DGA is the preceding qualification layer: before adaptation, it prevents the

organization from treating a local normality boundary as universal and provides a transparent portable path for the fault-typing tasks that can be defended with scale-reduced chemistry.

The methodological contribution is fourfold. First, HIRA-DGA introduces a representation-by-decision architecture: total magnitude is retained only where a locally calibrated screening decision needs it, while portable mechanism and severity decisions use scale-reduced chemistry. Existing one-shot classifiers generally ask one representation to solve all labels simultaneously.

Second, the framework separates local validation from zero-target-tuning transfer. Stage 0 is deliberately not claimed as externally portable. This boundary is methodological: a diagnostic system should state which component is locally calibrated instead of implying that every trained threshold transfers.

Third, HIRA-DGA treats abstention as an engineering class of action. Mixed and unseen descriptions are not silently collapsed into a known label. The output HOLD is traceable and directs the case to another diagnostic process.

Fourth, the contribution is grounded in real secondary data and exact source files. The same five gases were analyzed across an electrical-utility source from Malaysia and a repository of cases obtained with Ukrainian electrical laboratories and energy companies. The two-source experiment shows that when source magnitudes differ by up to three orders of magnitude, the diagnostic workflow itself must decide what to localize, what to transfer, and when to abstain.

Relative to the prior multistage system of Dhini et al. [18], HIRA-DGA does not claim superiority from a cross-study accuracy comparison because the underlying datasets and validation protocols differ. The advance is instead in the technology-design question: Dhini et al. optimized staged within-source diagnosis, whereas HIRA-DGA assigns transfer roles to the stages, tests an independent source without target tuning, adds a nonlinear learner check, and defines HOLD as an auditable endpoint. This makes the contribution a performance-qualification architecture rather than another classifier ranking exercise.

A utility implementation can follow four records in a diagnostic passport. Record 1 stores raw gases, units, sample date, laboratory method, transformer identity, and model source. Record 2 executes the locally calibrated S0 screen and records that its authority is local. Record 3 applies S1 and S2 only to cases entering the portable fault-typing path and stores the representation and model version used. Record 4 stores the final label or HOLD disposition together with the reason for abstention.

The practical lesson is that model portability should be decomposed rather than assumed. Local normality thresholds may be updated with confirmed utility data. Portable fault typing can be challenged across organizations without target tuning. Mixed cases should remain held until a multi-label model is developed from adequate confirmed examples. This creates a controlled migration path from a public-data prototype to a locally validated industrial system.

5. Limitations and Future Work

The study has important limitations. The development analysis uses the 200 rows present in the public PLOS supporting file, not the 400 cases described in the parent article. The external challenge contains only 25 cases, of which 19 can be scored as single mechanisms, and it contains no verified normal examples; Stage 0 therefore cannot be externally validated. The evidence base is one development source and one external source. Accordingly, the title and claims use cross-source rather than implying general multi-utility robustness. The exact measurement units and laboratory procedures of the Malaysian supporting table are not independently documented in the released file, whereas the Ukrainian repository specifies ppm. The observed magnitude shift therefore cannot be attributed to utility population alone.

The label harmonization also compresses detailed fault descriptions into three common single-mechanism classes. One description, sustained developed PD with overlapping, is semantically less clear than the standard T1/T2/T3 and D2 codes. The full mapping is disclosed so that reviewers can inspect it, and future work should repeat the analysis with harmonized IEC or Duval labels assigned prospectively by domain experts.

HIRA-DGA is not a claim of universal 94.7% cross-source accuracy. The external confidence interval is wide, the fault-only ratio comparator already reaches 89.5% with logistic regression, and the LightGBM results show meaningful learner-family sensitivity. The strongest empirical support for representation assignment comes from the Stage 2 ablation, not from the one-case difference between logistic HIRA-DGA and the fault-only comparator. Future work should reserve entire utilities as test domains, include verified normal cases, retain multi-label faults, report target-free and target-adapted results separately, and evaluate coverage, false-HOLD cost, and false-AUTO cost.

A next methodological extension is a prospective multi-utility HIRA passport. Stage 0 would be calibrated per organization; S1 and S2 would be tested as portable chemistry modules; S3 would expand from rule-based abstention to conformal or open-set uncertainty; and confirmed mixed cases would support a multi-label branch. Such work would connect DGA with recent multi-source, physics-informed, digital-twin, and edge-inference research while preserving explicit control over model authority [1, 5, 25-27].

6. Conclusions

Two independently sourced transformer DGA collections showed that the same harmonized fault classes could differ in median total gas magnitude by 65 to 1,322 times. Under that shift, direct four-class logistic models achieved only 2/19 with absolute-log inputs and 7/19 with ratios. A fixed nonlinear LightGBM check produced the same warning: high grouped internal accuracy coexisted with 0/19 absolute-log and 5/19 ratio performance on the external fault challenge.

The original contribution is HIRA-DGA. The framework separates a locally calibrated normal/fault screen from portable mechanism and discharge-severity stages, assigns each stage a chemistry representation matched to its transfer role, and returns HOLD for mixed or unsupported label structures. Without target-domain tuning, the primary logistic path classified 18/19 single-

mechanism external cases (94.7%; 95% CI 75.4-99.1%), while the centered log-ratio discharge-severity stage achieved 11/11 compared with 5/11 for its absolute-log ablation. The LightGBM check improved from 5/19 for direct ratio classification to 13/19 for the hierarchical path but remained below its 15/19 fault-only comparator, demonstrating that the architecture controls the decision process without guaranteeing learner-independent superiority. All six mixed or unseen cases were held rather than forced into a class.

The practical conclusion is not that one public model is universally deployable. It is that transformer diagnostic technology should explicitly decide what remains local, what can be challenged across sources, and when automation must abstain. The contribution is a reproducible technology-design concept supported by real data, formal transformations, transparent algorithmic logic, duplicate-aware validation, an independent-source challenge, a nonlinear robustness check, and source-traceable supplementary files.

Abbreviations

CLR, centered log-ratio; DGA, dissolved gas analysis; HIRA-DGA, Hierarchical Invariant-Representation and Abstention for dissolved-gas analysis; HOLD, abstention output for unsupported label structures; LR, logistic regression; ML, machine learning.

Author Contributions

Conceptualization, A.A., L.C.D.J. and R.A.; methodology, A.A., L.C.D.J., J.R.P.L., R.A. and S.M.; software, L.C.D.J. and J.R.P.L.; validation, J.R.P.L., M.B., R.A. and S.M.; formal analysis, A.A., L.C.D.J. and M.B.; investigation, A.A., M.B. and S.M.; resources, R.A. and S.M.; data curation, J.R.P.L., M.B. and S.M.; writing-original draft preparation, A.A., L.C.D.J., J.R.P.L. and M.B.; writing-review and editing, all authors; visualization, L.C.D.J., J.R.P.L. and M.B.; supervision, R.A. and S.M.; project administration, A.A. and R.A. All authors have read and agreed to the submitted version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The development data are publicly available as the PLOS ONE supporting file at https://doi.org/10.1371/journal.pone.0191366.s001. The external dataset is available from Mendeley Data at https://doi.org/10.17632/x22y4s9fks.1. Source-derived tables, mappings, analysis code, validation outputs, external predictions, and figures are included in the supplementary submission package described below.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgement

The authors acknowledge the researchers who publicly released the transformer DGA sources used for secondary analysis. The study does not disclose proprietary transformer measurements.

Institutional Review Board Statement

Not applicable. The study used publicly available secondary data and did not involve human participants or animals.

Informed Consent Statement

Not applicable.

Supplementary Materials

The submission package contains: (S1) PLOS_DGA_S1.docx, the exact public supporting file identified by DOI 10.1371/journal.pone.0191366.s001; (S2) malaysia_dga_200.csv, the source-derived development table; (S3) ukraine_dga_25.csv, the source-derived external table and mapping; (S4) hira_analysis_v2.py, which reproduces logistic and LightGBM analyses; (S5) grouped validation, external result, Holm-adjusted test, and row-level prediction files; (S6) graphical abstract and analysis figures; and (S7) README_DATA_SOURCES.txt with persistent identifiers, transformations, model settings, and file notes. The original external dataset remains available from Mendeley Data, DOI 10.17632/x22y4s9fks.1.

Declaration of Generative AI Use

OpenAI ChatGPT was used only to improve language and readability and to assist with reference-style conversion and journal formatting. No generative AI tool was used to create or alter the figures or to perform the reported analysis. The authors reviewed and verified all technical content, citations, analyses, and conclusions and accept full responsibility for the manuscript.

References

[1] Huang, Y., Huang, Z., & Chen, J. (2026). Multi-source data fusion and multi-task physics-informed transformer for power transformer fault diagnosis. Energies, 19(3), 599, https://doi.org/10.3390/en19030599.

[2] Nkwambe, M. S., Thango, B. A., & Ogudo, K. A. (2026). Power transformer fault classification from dissolved gas analysis using machine learning. Energies, 19(13), 2947, https://doi.org/10.3390/en19132947.

[3] Lin, J., Yeo, L. Y., Afrouzi, H. N., Ektesabi, M. M., & Tavalaei, J. (2025). From limited data to reliable diagnosis: An interpretable deep learning framework for transformer fault analysis. International Journal of Electrical Power & Energy Systems, 172, 111227, https://doi.org/10.1016/j.ijepes.2025.111227.

[4] Moradi, E., Elsisi, M., Mahmoud, K., Lehtonen, M., & Darwish, M. M. (2025). Robust deep neural network-based internet of things for power transformer fault diagnosis under imbalanced data and uncertainties. International Journal of Electrical Power & Energy Systems, 168, 110731, https://doi.org/10.1016/j.ijepes.2025.110731.

[5] Hu, D., Yang, Y., Dai, H., Tang, C., & Xie, J. (2025). An interpretable machine learning method for fault diagnosis of oil-immersed transformers based on edge inference. International Journal of Electrical Power & Energy Systems, 168, 110647, https://doi.org/10.1016/j.ijepes.2025.110647.

[6] Adekunle, A. A., Fofana, I., Picher, P., Rodriguez-Celis, E. M., Arroyo-Fernandez, O. H., & Simard, H. (2025). Multiclass fault diagnosis in power transformers using dissolved gas analysis and grid search-optimized machine learning. Energies, 18(13), 3535, https://doi.org/10.3390/en18133535.

[7] Sutikno, H., Prasojo, R. A., Abu-Siada, A., & Suwarno. (2024). Machine learning based multi-method interpretation to enhance dissolved gas analysis for power transformer fault diagnosis. Heliyon, 10(4), e25975, https://doi.org/10.1016/j.heliyon.2024.e25975.

[8] Wang, T., Li, Q., Yang, J., Xie, T., Wu, P., & Liang, J. (2023). Transformer fault diagnosis method based on incomplete data and TPE-XGBoost. Applied Sciences, 13(13), 7539, https://doi.org/10.3390/app13137539.

[9] Zhang, D., Li, C., Shahidehpour, M., Wu, Q., Zhou, B., & Zhang, C. (2022). A bi-level machine learning method for fault diagnosis of oil-immersed transformers with feature explainability. International Journal of Electrical Power & Energy Systems, 134, 107356, https://doi.org/10.1016/j.ijepes.2021.107356.

[10] IEC. (2022). IEC 60599:2022 Mineral oil-filled electrical equipment in service - Guidance on the interpretation of dissolved and free gases analysis. International Electrotechnical Commission, Geneva.

[11] Abdelwahab, S. A. M., Taha, I. B. M., Fahim, R., & Ghoneim, S. S. M. (2025). Transformer fault diagnose intelligent system based on DGA methods. Scientific Reports, 15, 8263, https://doi.org/10.1038/s41598-024-78293-7.

[12] Azmi, P. A. R., Yusoff, M., & Sallehud-din, M. T. M. (2025). Improving transformer failure classification on imbalanced DGA data using data-level techniques and machine learning. Energy Reports, 13, 264-277, https://doi.org/10.1016/j.egyr.2024.12.006.

[13] Al-Sakini, S. R., Bilal, G. A., Sadiq, A. T., & Al-Maliki, W. A. K. (2025). Dissolved gas analysis for fault prediction in power transformers using machine learning techniques. Applied Sciences, 15(1), 118, https://doi.org/10.3390/app15010118.

[14] Guan, S., Wu, T., & Yang, H. (2024). Research on transformer fault diagnosis method based on ACGAN and CGWO-LSSVM. Scientific Reports, 14, 17676, https://doi.org/10.1038/s41598-024-68141-z.

[15] Wang, L., Chi, J., Ding, Y., Yao, H., Guo, Q., & Yang, H. (2024). Transformer fault diagnosis method based on SMOTE and NGO-GBDT. Scientific Reports, 14, 7179, https://doi.org/10.1038/s41598-024-57509-w.

[16] Guan, S., Yang, H., & Wu, T. (2023). Transformer fault diagnosis method based on TLR-ADASYN balanced dataset. Scientific Reports, 13, 23010, https://doi.org/10.1038/s41598-023-49901-9.

[17] Mahmoodiyan, H., Ahang, M., Abbasi, M., & Najjaran, H. (2025). Feature-weighted MMD-CORAL for domain adaptation in power transformer fault diagnosis. 2025 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), https://doi.org/10.1109/CCECE64018.2025.11364516.

[18] Dhini, A., Faqih, A., Kusumoputro, B., Surjandari, I., & Kusiak, A. (2020). Data-driven fault diagnosis of power transformers using dissolved gas analysis (DGA). International Journal of Technology, 11(2), 388-399, https://doi.org/10.14716/ijtech.v11i2.3625.

[19] Illias, H. A., & Zhao Liang, W. (2018a). Identification of transformer fault based on dissolved gas analysis using hybrid support vector machine-modified evolutionary particle swarm optimisation. PLOS ONE, 13(1), e0191366, https://doi.org/10.1371/journal.pone.0191366.

[20] Illias, H. A., & Zhao Liang, W. (2018b). S1 Table. Dissolved gas analysis (DGA) data. PLOS ONE Supporting Information, https://doi.org/10.1371/journal.pone.0191366.s001.

[21] Shutenko, O., & Kulyk, O. (2023b). DGA results for 25 transformers. Mendeley Data, V1, https://doi.org/10.17632/x22y4s9fks.1.

[22] Shutenko, O., & Kulyk, O. (2023a). Method of fault-type recognition based on the dissolved gas analysis using a set of diagnostic criteria. IET Generation, Transmission & Distribution, 17, 5511-5523, https://doi.org/10.1049/gtd2.13068.

[23] Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2), 139-177, https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.

[24] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146-3154.

[25] Duz, F. H. D. S., Zacarias, T. G., Ribeiro Junior, R. F., Steiner, F. M., Assuncao, F. D. O., & Bonaldi, E. L. (2025). Smart monitoring of power transformers in Substation 4.0: Multi-sensor integration and machine learning approach. Sensors, 25(17), 5469, https://doi.org/10.3390/s25175469.

[26] Cui, J., Kuang, W., Geng, K., & Jiao, P. (2025). Intelligent fault diagnosis and operation condition monitoring of transformer based on multi-source data fusion and mining. Scientific Reports, 15, 7606, https://doi.org/10.1038/s41598-025-91862-8.

[27] Yao, H., Zhang, X., Guo, Q., Miao, Y., & Guan, S. (2024). Fault diagnosis method for oil-immersed transformers integrated digital twin model. Scientific Reports, 14, 20355, https://doi.org/10.1038/s41598-024-71107-w.