# Characterizing the Moderate-Class Diagnostic Bottleneck in ML-Based Transformer Defect Diagnosis

**Tochukwu Udeh[1], Alexander Pirker[2], Markus Prossegger[1]**

[1]Carinthia University of Applied Sciences, Department of Engineering & IT, Villach, Austria,
+43 (0)665 65103133, udeh@fh-kaernten.at, www.fh-kaernten.at

[2]VUM Verfahren Umwelt Management GmbH, Graz/Klagenfurt, Austria,
+43 (0)664 88343073, alexander.pirker@vum.co.at, www.vum.co.at

**Summary:** Dissolved gas analysis (DGA) serves as the primary non-invasive method for diagnosing defects in power transformers. Conventional approaches for assessing and characterizing defects such as the Duval triangle and the IEC 60599 guidelines are based on threshold values and ratios of key gases. Due to various factors influencing the production of these gases, these methods can lead to incorrect conclusions. Machine learning (ML) models can overcome these constraints by capturing nonlinear patterns in DGA data; however, their performance across fault severities remains insufficiently explored, particularly within imbalanced real-world datasets where transitional fault states introduce diagnostic ambiguity.

This study evaluates ML classifiers for DGA-based fault diagnosis, revealing a key limitation in early fault detection. Using 1,097 expertly labeled transformer records and transformer-wise stratified cross-validation, nine ML models were benchmarked. XGBoost performed best (macro F1-score: 0.757 ± 0.073). A moderate-class bottleneck was identified: while severe faults were detected reliably (AUC ≥ 0.881), moderate fault classification was suboptimal. Quantitative analysis showed 83.3% of moderate condition were misclassified as normal, occurring within a diagnostic "gray zone" defined by low gas concentrations ($C_2H_4$ < 3.0, $CH_4$ < 3.2, expressed as z scores following standard scaling) and high feature overlap (Jensen–Shannon similarity = 76.5%). SHAP analysis confirmed model alignment with fault physics but limited discriminatory power for defects. Figure 1 shows the methodology flow diagram of this procedure. The findings highlight a core challenge: differentiating early faults condition (moderate) from normal. This necessitates integrating temporal or multimodal data for granular diagnosis. Practically, DGA alerts for moderate-faults indications require cautious interpretation, contextual data integration and experts review for reliable predictive maintenance.

**Keywords:** Power transformer; dissolved gas analysis; fault diagnosis; machine learning; explainable AI; class imbalance; performance stratification; early fault detection; uncertainty; predictive maintenance; dimensionality reduction; feature space analysis.

# 1 Introduction

Power transformers are critical components in electrical networks, where unexpected failures can lead to extensive outages and replacement costs exceeding several million dollars [1,2]. To prevent such incidents, utilities employ condition-based maintenance (CBM) practices that, in addition to other methods, rely on dissolved gas analysis (DGA), which is a non-invasive technique quantifying fault gases such as $H_2$, $CH_4$, and $C_2H_2$ that are generated by a defect during insulation degradation [3,4].

Conventional DGA interpretation schemes, including IEC 60599 [5] and IEEE C57.104.2008 [6], apply fixed gas ratios and thresholds to classify fault types. Due to various factors influencing the production of the key gases or because of multiple defects, these methods can lead to incorrect conclusions [7]. In response, machine learning (ML) approaches have emerged as data-driven alternatives capable of modeling nonlinear dependencies among gas concentrations and operational variables [8]. Ensemble models and synthetic sampling techniques have improved diagnostic balance under skewed fault distributions, where severe cases are typically underrepresented [9,10]. Similarly, recent studies have shown that modeling gas generation as a time series can enhance detection of incipient faults by capturing gradual concentration changes (temporal patterns) instead of static levels [11].

Nevertheless, most prior studies assess overall model performance without disaggregating results by fault severity [12,13]. In practice, insulation degradation evolves gradually from normal condition through moderate condition to severe states (fault condition). The moderate state, where gas levels overlap with normal state, presents the most persistent diagnostic challenge. Misclassification of this moderate state often results in missed detection of early faults and delayed maintenance action. Recent works have highlighted challenges in early or incipient fault detection using hybrid AI models and domain adaptation techniques [14,15,16]. However, the specific diagnostic bottleneck associated with moderate degradation states where gas signatures overlap strongly with normal operation remains underexplored in real-world datasets.

This study aims to address this gap by systematically evaluating ML-based diagnostic models across fault severity levels, with a focus on the moderate-class bottleneck. Using a real-world, expertly labeled dataset, we benchmark multiple classifiers, analyze feature-space overlap, and provide explainable insights into model behavior. Our work contributes to both the academic understanding of early fault detection limits and practical guidance for asset managers relying on DGA for predictive maintenance.

# 2 Methods

This section describes the dataset, preprocessing steps, feature construction, model training procedures, and explainability techniques used to develop and evaluate the proposed transformer health classification framework.

## 2.1 Dataset Description

A proprietary dissolved gas analysis (DGA) dataset was supplied by VUM Verfahren Umwelt Management GmbH (Austria). After record averaging and quality filtering during data

processing, 1,097 expert-labelled transformer-level samples remained and were used for model development. Each sample contains seven dissolved-gas concentrations ($H_2$, $CH_4$, $C_2H_2$, $C_2H_4$, $C_2H_6$, CO, $CO_2$), oil temperature, ambient temperature, transformer service age, and nominal voltage class. Operational health labels and counts are given in Table 2.1.

*Table 2.1: Transformer Samples Operational Health Labels*

| Condition | Count (n) | Percentage (%) |
|---|---|---|
| **Normal** | 944 | 86.1 |
| **Moderate** | 136 | 12.4 |
| **Fault** | 17 | 1.5 |

## 2.2 Missing Data and Imputation

Missing values occurred primarily in ambient temperature measurements (35%), largely because this parameter is often omitted from routine DGA sampling or lost during data transfer from older systems. Some values in the gases are missing because they were not measured in the past.

Imputation steps:

- Ambient temperature: daily-to-monthly averaging per transformer.
- Oil temperature and gases: multivariate iterative imputer with Random Forest (n_estimators = 100).

A comparison between complete-case (n = 671) and imputed (n = 1,097) datasets showed that imputation increased the sample size by 63 % but reduced the top F1-score by ≈ 9 %. The trade-off favored the imputed set for better class coverage.

## 2.3 Data Splitting and Imbalance Handling

To prevent leakage from repeated measurements of the same transformer, splits were performed at the transformer level. The partitioning used a split Training = 810, Validation = 233, Test = 54 transformer observations. Severe imbalance (Normal ≫ Moderate ≫ Fault) was corrected using the Synthetic Minority Over-sampling Technique (SMOTE) within training folds (k = 2, target 1:1). Three imbalance-handling strategies were compared: no correction, class weights, and SMOTE. SMOTE improved minority-class performance for ensemble models. After resampling, all features were normalized with robust scaling (median centering, IQR normalization).

## 2.4 Feature Engineering

Features were constructed from established DGA indicators and rate variables.

- Gas ratios: $R_1$ ($CH_4/H_2$), $R_2$ ($C_2H_2/C_2H_4$), $R_3$ ($C_2H_2/CH_4$), $R_5$ ($C_2H_6/C_2H_2$), and $CO_2/CO$.
- Gas generation rates: Δ concentration (ppm/months).

*Table 2.2: Feature Sets M1–M4*

| Feature Set | Components | Purpose |
|---|---|---|
| **M1** | Sample Age + seven gas concentrations | Baseline |
| **M2** | M1 + five ratios | Capture fault relations |
| **M3** | M2 + gas generation rates | Include temporal dynamics |
| **M4** | M3 + oil and ambient temperatures | Full operational context |

## 2.5 Summary of Data Preprocessing Steps

The preprocessing pipeline applied to all experiments consists of the following steps:

1. Load raw data.
2. Remove unlabeled records and records with more than 80% missing values.
3. Transformer samples with duplicate timestamps were averaged.
4. Impute missing values (Section 2.2).
5. Construct feature sets M1–M4.
6. Perform a transformer-wise (train, validation, and test) split.
7. Apply robust scaling using the median and interquartile range (IQR), fitted on the training data.
8. Apply SMOTE exclusively to the training folds to address class imbalance.
9. Perform hyperparameter tuning using stratified cross-validation within the training set.
10. Train models.
11. Evaluate on the held-out validation set and compute performance metrics and explainability outputs.

## 2.6 Model Training and Evaluation

Nine supervised classifiers were evaluated: Logistic Regression (LR), k-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), Naïve Bayes (NB), AdaBoost Classifier (ABC), Random Forest Classifier (RFC), Extra Trees Classifier (ETC), Extreme Gradient Boosting (XGB), and Multilayer Perceptron (MLP). Hyperparameters were optimized using stratified cross-validation on the training set, and final performance was assessed on a held-out validation set.

The primary performance metric was the macro-averaged F1-score. Secondary metrics included per-class precision and recall, overall accuracy, macro-averaged area under the receiver operating characteristic curve (AUC), and area under the precision–recall curve (AUC-PR). Ninety-five percent confidence intervals were computed across cross-validation folds.

Table 2.3 summarizes the hyperparameter configurations that achieved optimal macro-averaged F1-score performance during stratified cross-validation.

*Table 2.3: Hyperparameters for Selected Models*

| Model | Key Parameters |
|---|---|
| **XGBoost (XGB)** | n_estimators = 300, learning_rate = 0.1 |
| **Extra Tree (ETC)** | n_estimators = 300, max_depth = 20 |
| **Random Forest (RFC)** | n_estimators = 300 |
| **Multilayer Perceptron (MLP)** | hidden_layers = (50–150 nodes), activation = relu, α = 0.001 |
| **k-Nearest Neighbors (KNN)** | n_neighbors = 5, weights = distance, metric = manhattan |
| **Decision Tree Classifier (DTC)** | max_depth = 10, criterion = gini |
| **Logistic Regression (LR)** | C = 1, solver = lbfgs |
| **AdaBoost Classifier (ABC)** | n_estimators = 300, learning_rate = 0.01 |
| **Naïve Bayes (NB)** | var_smoothing = 1e-9 |

## 2.7 Explainability and Feature Space Analysis

Model interpretability was incorporated as an integral component of the proposed pipeline to ensure transparency, robustness, and trustworthiness of the predictive models. Post-hoc explainability techniques were applied after model training and evaluation to analyze both global feature importance and local decision behavior.

For tree-based models, SHAP (SHapley Additive exPlanations) TreeExplainer was employed due to its computational efficiency and exact formulation for ensemble trees. For non-tree-based models, SHAP KernelExplainer was used as for model-agnostic. Global feature importance was quantified using the mean absolute SHAP values, computed over 500 randomly selected training samples, providing a consistent measure of each feature's overall contribution to model predictions.

The local explainability analysis was focused on representative misclassified samples, which enabled a detailed examination of how individual feature contributions influenced incorrect predictions. SHAP Beeswarm plot and summary plots were analyzed to identify dominant features, potential feature interactions, and their impact on the model output, thereby supporting qualitative model validation and fault physics.

To investigate class structure in the learned feature space, we used dimensionality reduction for qualitative inspection and a distributional metric for quantification. PCA was applied to summarize the dominant linear variance directions, and t-SNE was used to visualize local, non-linear neighborhood relationships among samples. To quantify class overlap beyond visual inspection, we computed the Jensen–Shannon divergence between class-conditional feature distributions, providing a symmetric measure of distributional similarity between class representations.

Figure 2.1 illustrates the complete end-to-end approach and methodology that was adopted in this study. The pipeline begins with raw data ingestion from the oil database, followed by feature extraction, preprocessing, and feature engineering. The processed data are then normalized and split into training, validation, and test subsets prior to model training and

evaluation. Finally, post-hoc explainability and feature space analyses are applied to the trained models, ensuring that predictive performance is complemented by interpretable and diagnostically meaningful insights.

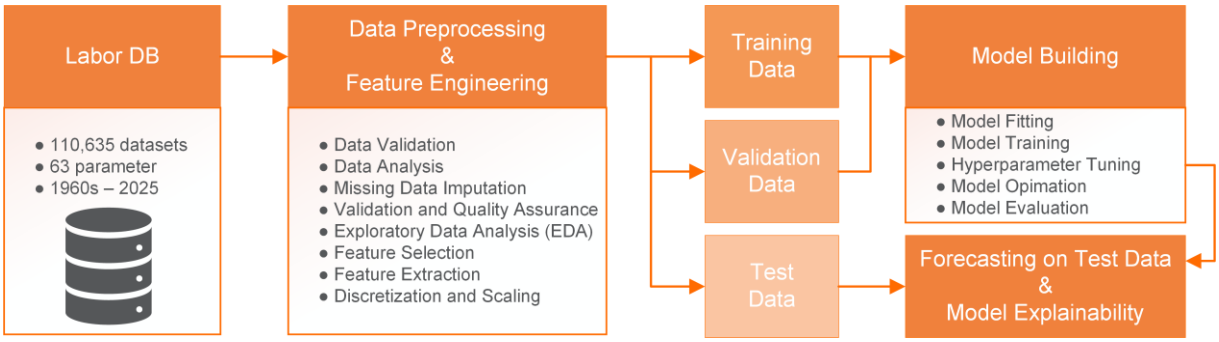


*Figure 2.1: Methodology flow diagram. Pipeline steps from raw data to final explainability outputs.*

# 3 Results

This section presents the empirical findings of the study, which includes, dataset characteristics, comparative model performance, class-specific behavior, and model interpretability analyses that explain observed performance trends.

## 3.1 Exploratory Data Analysis and Dataset Characteristics

The dataset comprised 1,097 transformer records, categorized as Normal (**86.1%**), Moderate (**12.4%**), and Fault (**1.5%**), indicating a pronounced class imbalance typical of real-world condition monitoring data. The age distribution of transformers is illustrated in **Figure 3(a)**, which showed a broad range from **0 to 75 years**, with a mean age of **24.2 years** (standard deviation ≈ **15.7 years**). The distribution is moderately right-skewed, with the majority of units clustered between early-life and mid-life operational stages. There is no meaningful relationship was observed between transformer age and fault occurrence (**r = −0.099**), this suggests that age alone is not a reliable predictor of failure in this dataset.

The correlation analysis among dissolved gas variables, presented in **Figure 3(b)**, revealed strong positive dependencies among hydrocarbon gases, particularly those associated with thermal and electrical faults. Notable correlations include $C_2H_4$–$C_2H_2$ (**r = 0.73**), $CH_4$–$C_2H_4$ (**r = 0.98**), and $C_2H_6$–$C_2H_4$ (**r = 0.97**). In contrast, transformer age exhibited consistently weak correlations with individual gas concentrations, reinforcing its limited diagnostic significance when considered in isolation.

Overall, the observed gas interdependencies align with established transformer fault chemistry, which confirms the physical consistency and reliability of the dataset. These characteristics provide a solid foundation for subsequent model development, feature selection, and interpretation.

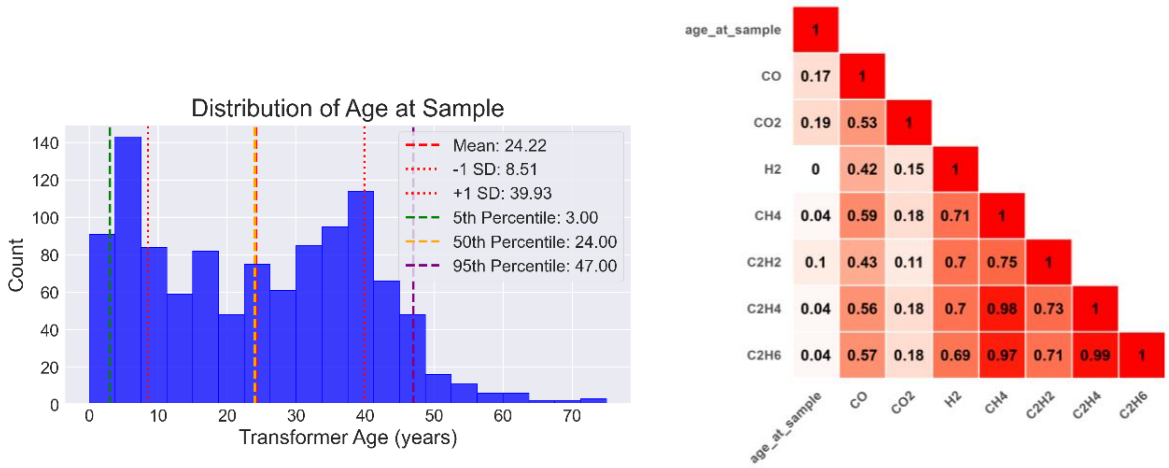


*Figure 3.1: (A) Histogram Distribution of transformer samples age. (B) Correlation heatmap of seven key DGA gases and operational parameters*

## 3.2 Model Performance Overview

A total of nine classifiers were benchmarked. Ensemble models consistently outperformed conventional learners. XGBoost attained the highest macro F1-score (0.757 ± 0.073), while Extra Trees yielded the highest accuracy (0.926 ± 0.004) and inter-rater reliability (κ = 0.686 ± 0.011). Naïve Bayes showed limited generalization (F1 = 0.218 ± 0.137).

Statistical testing confirmed model heterogeneity (Friedman $\chi^2 = 42.3$, $p < 0.001$). However, XGBoost and Extra Trees differences were not statistically significant (Cohen's d = 0.071), indicating comparable discriminative ability.

*Table 3.1: Top-performing classifiers (mean ± SD: 95% CI).*

| Model | Accuracy (95% CI) | F1-Score (95% CI) | ROC-AUC | PR-AUC | κ (95% CI) | MCC (95% CI) |
|---|---|---|---|---|---|---|
| **XGB** | 0.914 ± 0.009 (0.876–0.951) | 0.757 ± 0.073 (0.441–1.072) | 0.883 | 0.778 | 0.647 ± 0.033 (0.503–0.791) | 0.650 ± 0.032 (0.512-0.788) |
| **ETC** | 0.926 ± 0.004 (0.910–0.942) | 0.746 ± 0.098 (0.323–1.168) | 0.847 | 0.788 | 0.686 ± 0.011 (0.638–0.734) | 0.688 ± 0.011 (0.641-0.735) |
| **RFC** | 0.907 ± 0.006 (0.883–0.932) | 0.685 ± 0.072 (0.377–0.994) | 0.852 | 0.722 | 0.618 ± 0.007 (0.587–0.649) | 0.621 ± 0.007 (0.592-0.649) |

The Precision-Recall AUC (PR-AUC) metrics, which are more informative for imbalanced datasets, followed a similar trend to the macro F1-scores, with ETC achieving the highest value (0.788). This further confirms its superior performance in handling the class imbalance.

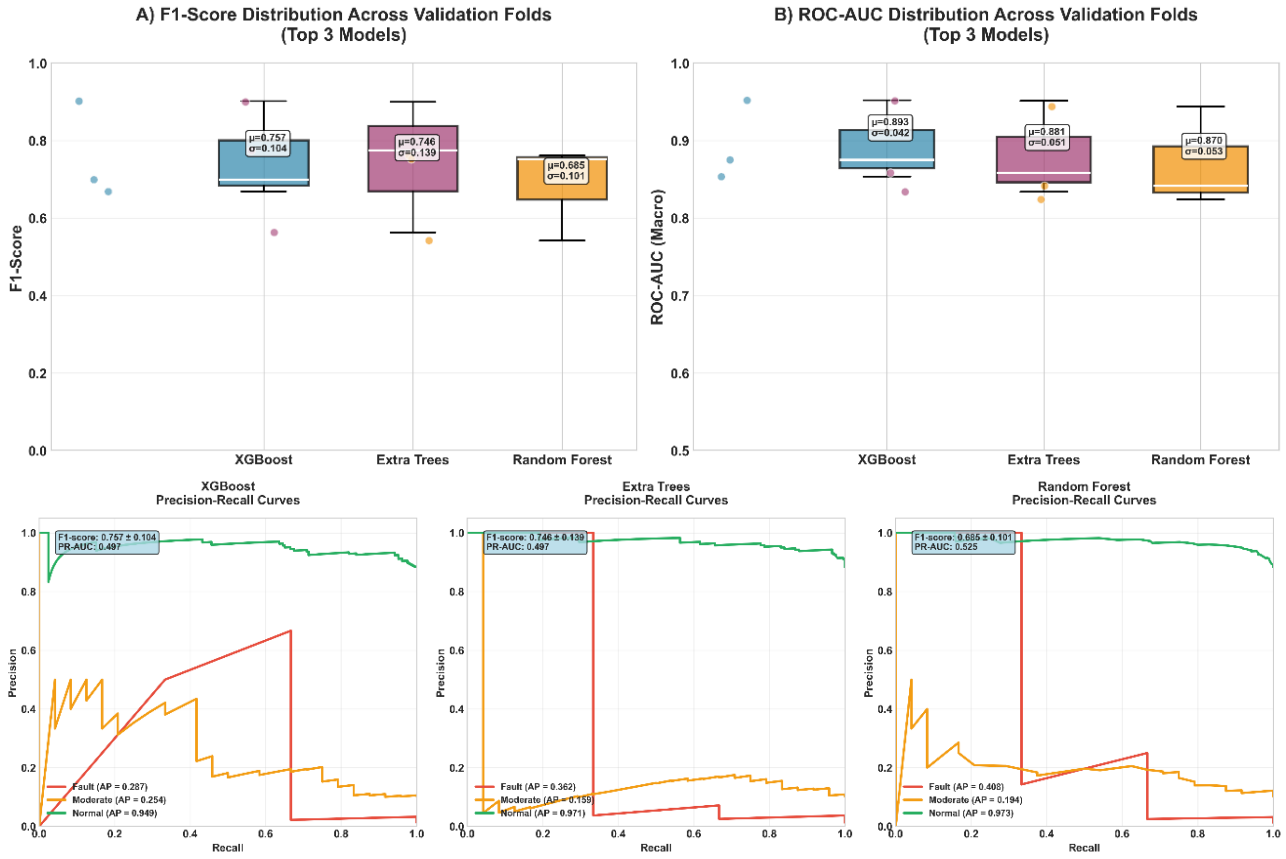


*Figure 3.2: (A) Box plots of macro F1, ROC-AUC, and (B) PR-AUC values across the top three models.*

## 3.3 Moderate-Class Detection

### 3.3.1 Class-Level Metrics

The class-level F1-scores and the corresponding normalized confusion matrices (Figure 3.3) provide complementary insights into model behavior across operating conditions. While the Normal class consistently achieves high and stable performance across all models, the Moderate class exhibits substantially reduced precision and recall, driven primarily by systematic misclassification as Normal. As shown in the confusion matrices, between 50 – 60% of Moderate samples and over 75% in the XGBoost model are incorrectly assigned to the Normal class, indicating pronounced overlap in their gas signature characteristics. Notably, this confusion pattern remains consistent across model architectures and data splits, suggesting that the observed performance degradation is not attributable to model capacity or class imbalance alone, but rather reflects intrinsic ambiguity in the underlying feature representations. These class-specific error structures directly motivate the feature space analyses and explainability presented in Section 3.3.2 and Section 3.4, where dimensionality reduction, distribution similarity metrics, and SHAP-based attribution, are used to investigate the source of overlap between Moderate and Normal conditions.

*Table 3.2: Per-class F1-scores for top models.*

| Model | Fault (F1) | Moderate (F1) | Normal (F1) |
|---|---|---|---|
| **XGB** | 0.633 ± 0.186 | 0.684 ± 0.032 | 0.952 ± 0.003 |
| **ETC** | 0.556 ± 0.294 | 0.723 ± 0.011 | 0.959 ± 0.002 |
| **RFC** | 0.444 ± 0.222 | 0.664 ± 0.008 | 0.948 ± 0.003 |

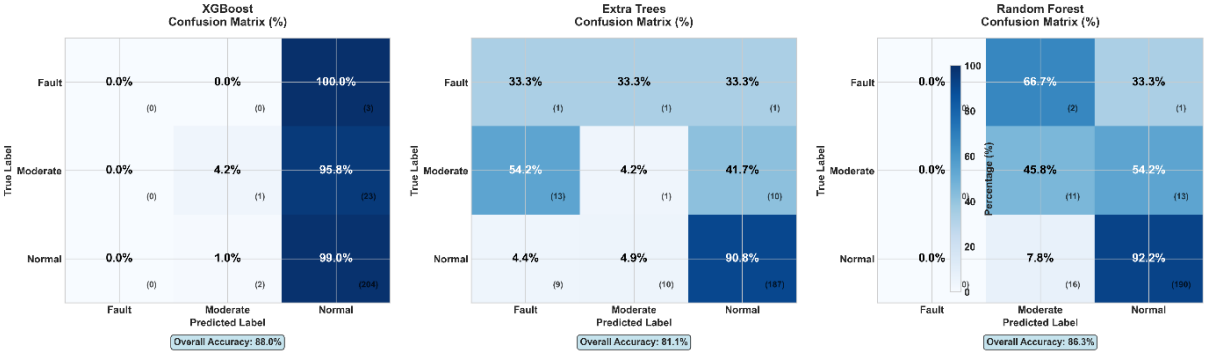


*Figure 3.3: Normalized confusion matrices for the top three models.*

### 3.3.2 Feature Space Analysis of Class Overlap

Dimensionality-reduction analyses using t-SNE and PCA were applied to investigate the geometric basis of Moderate-class confusion. Both methods consistently revealed substantial overlap between Normal and Moderate samples, with t-SNE showing no separable boundary and PCA (capturing 57.9 % of variance in the first two components) placing most Moderate samples within the convex hull of Normal samples. This intrinsic feature-space entanglement aligns with observed misclassification rates, where 55–60 % of Moderate samples were predicted as Normal. Jensen–Shannon divergence further quantified this proximity: the Normal–Moderate pair exhibited the highest similarity (JS = 0.235; 76.5 %), exceeding Normal–Fault and Moderate–Fault similarities by 12–20 %. Feature-level analysis confirmed that Moderate degradation states produce dissolved-gas patterns that mirror Normal conditions, differing mainly in magnitude rather than compositional structure, thereby creating a continuum that conventional classifiers struggle to disentangle.

Model-behavior diagnostics reinforced this interpretation. Prediction-uncertainty analysis showed that Moderate samples consistently exhibited higher entropy (0.4–0.8) compared to Normal samples (0.0–0.4), reflecting the model's low confidence when operating within the overlapping region of the feature space. The t-SNE uncertainty overlay localized these high-entropy zones precisely within the Normal–Moderate intersection, demonstrating a strong spatial correlation between geometric overlap and predictive ambiguity. Collectively, these findings indicate that the Moderate-class bottleneck arises from inherent data-distribution limitations rather than model deficiencies, underscoring the fundamental diagnostic challenge posed by the continuum between normal and moderately degraded states

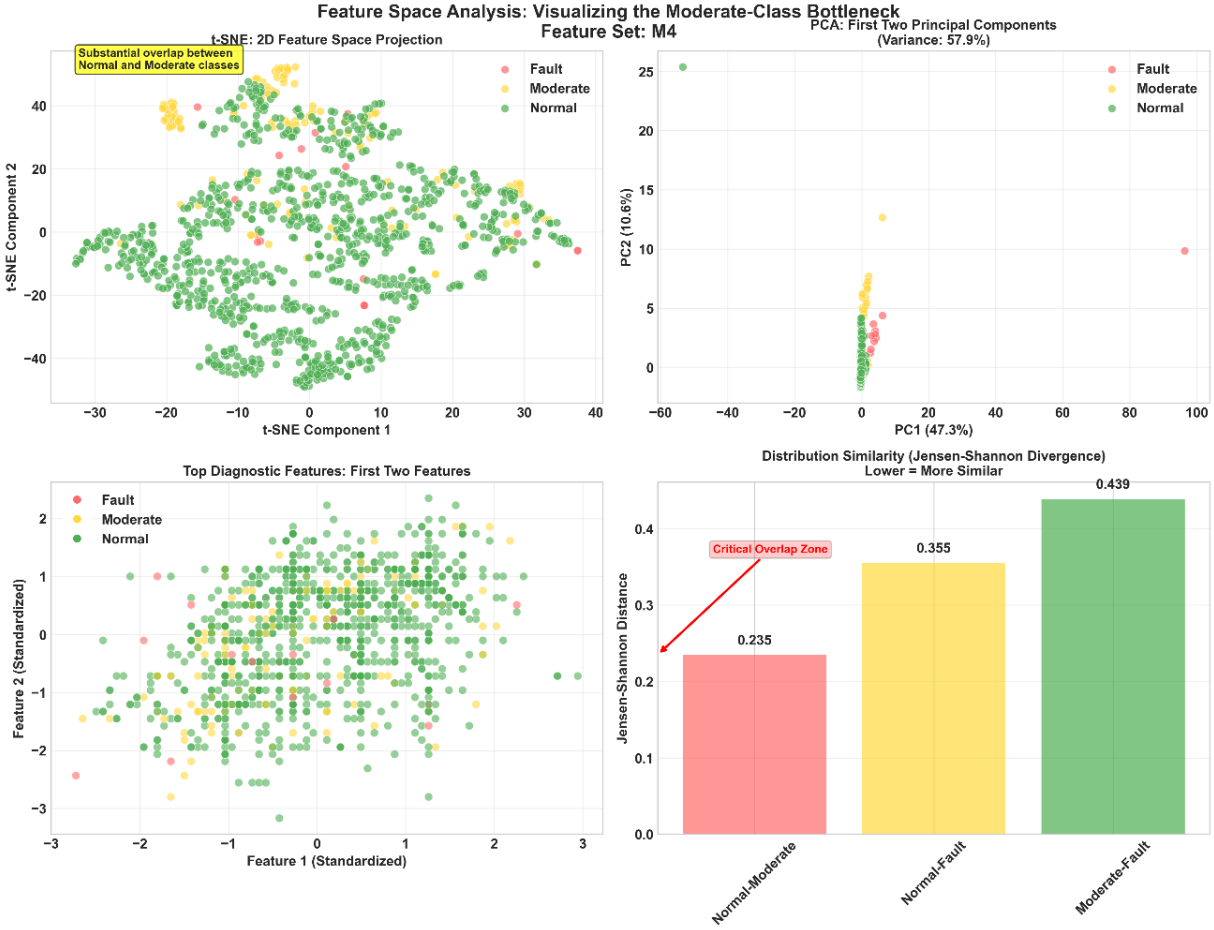


*Figure 3.4: Feature-space analysis showing class overlap and distribution similarity.*
*(A) t-SNE projection showing substantial overlap between Normal and Moderate classes.*
*(B) PCA projection (57.9 % variance explained) confirming the overlap pattern.*
*(C) Distribution in top diagnostic feature space.*
*(D) Jensen–Shannon divergence between class pairs, with Normal–Moderate showing highest similarity (76.5 %).*

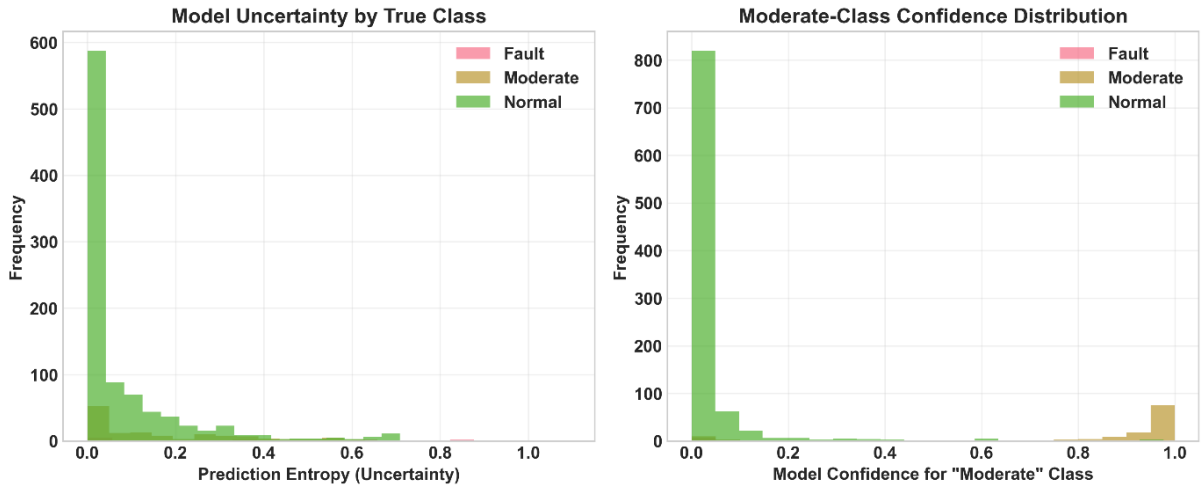


*Figure 3.5: Model uncertainty and confidence analysis. (Left) Prediction entropy by true class, showing higher uncertainty for Moderate samples. (Right) Model confidence for Moderate-class predictions, demonstrating low confidence for true Moderate cases due to feature-space overlap.*

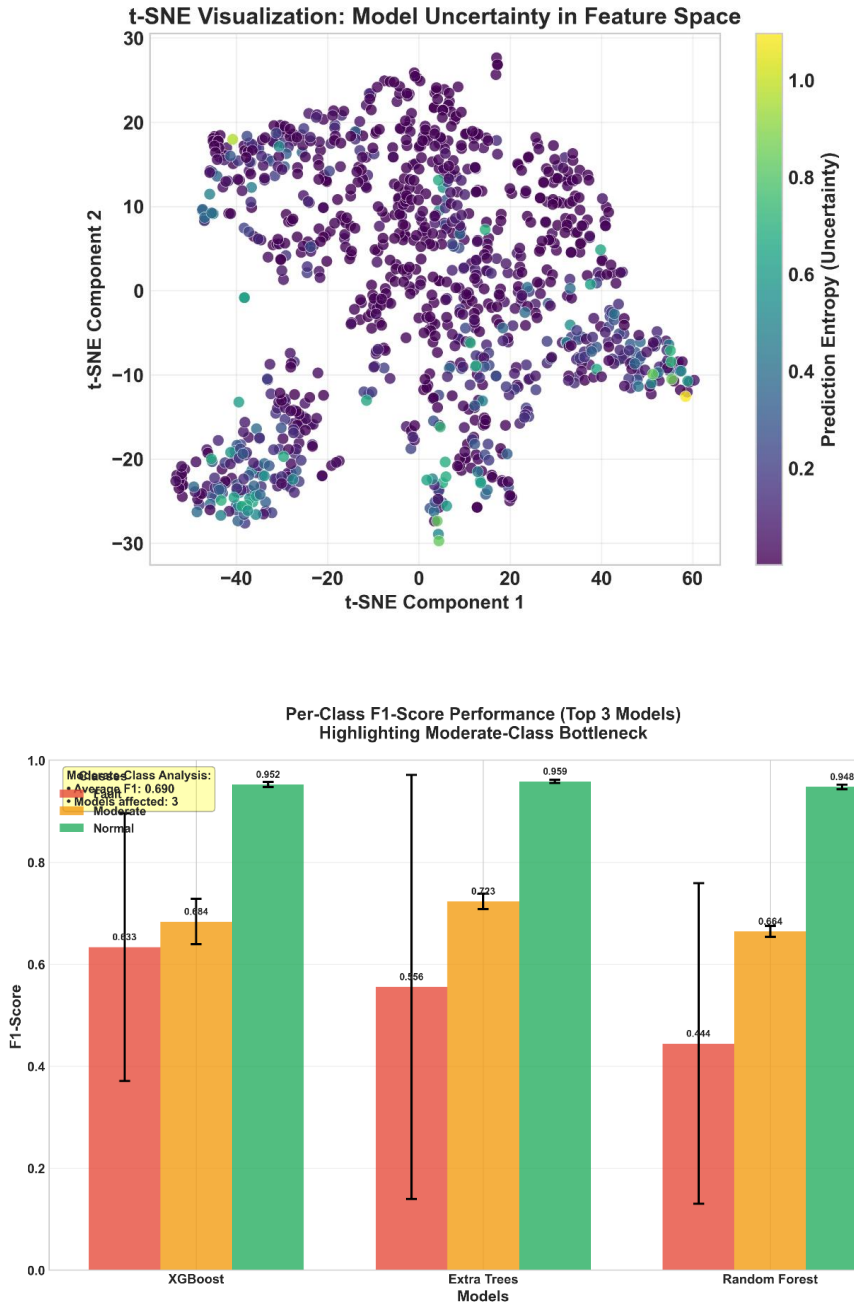


*Figure 3.6: (A) shows a t-SNE uncertainty overlay where high-entropy regions (warmer colors) align with the Normal–Moderate overlap, mapping the diagnostic bottleneck. (B) presents per-class F1-scores, emphasizing the Moderate-class performance deficit.*

*Table 3.3: Quantitative metrics of class-pair separability and prediction uncertainty*

| Metric | Normal–Moderate | Normal–Fault | Moderate–Fault |
|---|---|---|---|
| **Jensen–Shannon Distance** | 0.235 | 0.355 | 0.439 |
| **Distribution Similarity (%)** | 76.5 % | 64.5 % | 56.1 % |
| **Performance Gap (ΔF1)** | 26.8 % | 31.9 % | 27.9 % |
| **Model Uncertainty (Mean Entropy)** | 0.58 ± 0.12 | 0.32 ± 0.08 | 0.41 ± 0.10 |

### 3.3.3 Misclassification Patterns and Gas Signature Analysis

Analysis of misclassification patterns shows that Moderate faults are difficult to detect because their gas signatures often resemble Normal conditions. XGBoost misclassified 83.3% of Moderate samples as Normal, while Naïve Bayes achieved 95.8% Moderate-class accuracy despite weak overall performance, reflecting its bias toward majority-class patterns. Feature comparisons reveal that correctly classified Moderate samples exhibit much higher $C_2H_4$, $CH_4$, and $C_2H_6$ concentrations, whereas misclassified samples fall within Normal-like ranges, creating an ambiguous diagnostic zone. Confidence scores reinforce this: correctly classified Moderate samples show high confidence (0.69), while misclassified ones show very low confidence (0.12), similar to the model's response to true Normal samples. Gas-ratio patterns, including elevated R3 and $CO_2$/CO ratios in misclassified cases, further indicate mixed or overlapping fault signatures that obscure class boundaries and drive systematic diagnostic uncertainty.

*Table 3.4: Summary of Moderate-class misclassification proportions across models.*

| | XGB | ETC | RFC | DTC | ABC | MLP | KNN | LR | NB |
|---|---|---|---|---|---|---|---|---|---|
| **Misclassified as Normal (%)** | 83.3 | 66.7 | 62.5 | 66.7 | 66.7 | 50.0 | 33.3 | 8.3 | 4.2 |
| **Misclassified as Fault (%)** | 0.0 | 4.2 | 0.0 | 4.2 | 4.2 | 12.5 | 8.3 | 45.8 | 0.0 |
| **Correctly Classified (%)** | 16.7 | 29.2 | 37.5 | 29.2 | 29.2 | 37.5 | 58.3 | 45.8 | 95.8 |
| **Total Samples** | 24 | 24 | 24 | 24 | 24 | 24 | 24 | 24 | 24 |

**Caption:** "*Misclassification patterns for Moderate class across all models (validation set: 24 Moderate samples). XGBoost shows highest misclassification rate as Normal, while Naïve Bayes achieves highest Moderate-class accuracy despite poor overall performance*."

### 3.3.4 Computational Efficiency and Overfitting

All computational experiments were conducted on a standard workstation equipped with an Intel ProBook 6570s processor (2.74 GHz, 5 cores), 12 GB RAM, using Python 3.12. Reported training times are provided to ensure reproducibility and fair comparison across models under identical hardware and software conditions.

In terms of computational efficiency, the Extra Trees Classifier (ETC) achieved the fastest training time (8.29 s), followed by XGBoost (23.2 s) and the Multilayer Perceptron (MLP, 29.8 s). Despite its higher computational cost, XGBoost demonstrated strong generalization performance. XGBoost and Extra Trees exhibited the smallest overfitting gaps, indicating robust learning and effective control of model complexity.

Conversely, MLP showed the most severe overfitting, consistent with its high representational capacity and sensitivity to limited fault samples. AdaBoost (ABC), Decision Tree (DTC), and KNN also exhibited relatively large generalization gaps. In contrast, Logistic Regression (LR), Naive Bayes (NB), Random Forest (RFC), Extra Trees (ETC), and XGBoost (XGB) demonstrated only moderate overfitting, reflecting stable performance across training and testing datasets.

From an uncertainty perspective, Decision Tree and Naive Bayes produced the lowest prediction entropy, indicating highly confident—though potentially overconfident—outputs,

while AdaBoost exhibited the highest prediction uncertainty. These findings highlight the trade-offs between computational cost, generalization capability, and predictive confidence, reinforcing the suitability of ensemble-based methods for transformer fault classification.

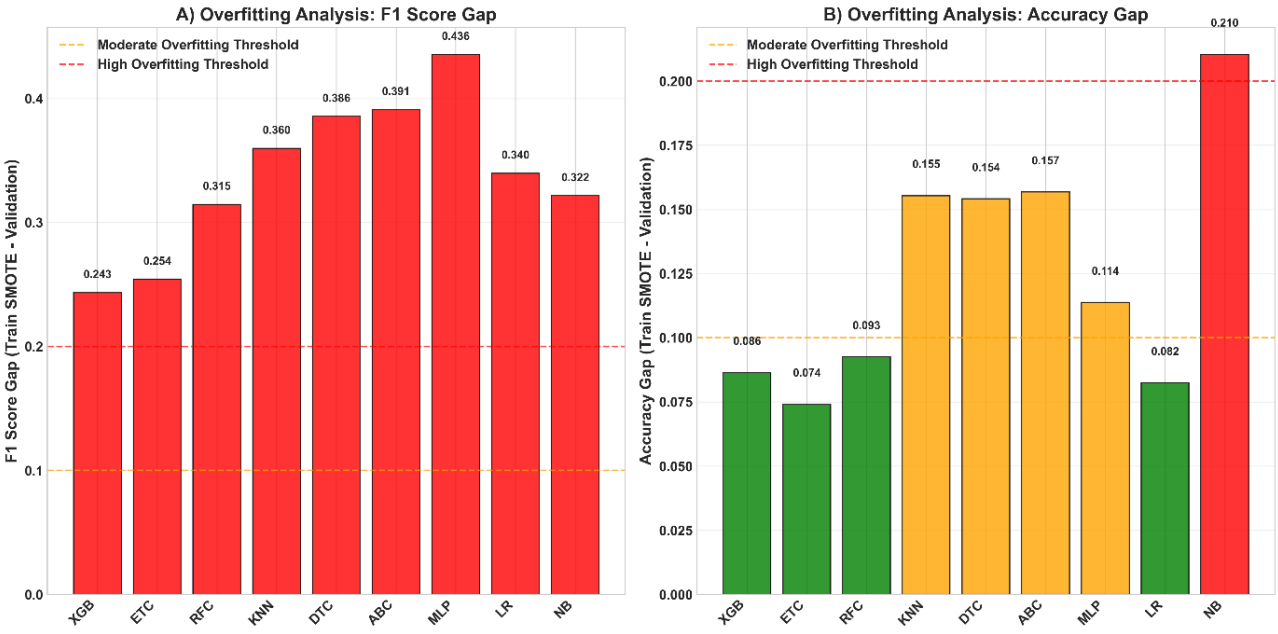


*Figure 3.7: Overfitting analysis showing (A) F1-score gap and (B) Accuracy gap (train vs. validation). Dashed lines indicate moderate (orange) and high (red) overfitting thresholds.*

## 3.4 Model Explainability

### 3.4.1 Rank Correlation Between Methods

SHAP analysis confirmed model alignment with fault physics. Across methods, ethene ($C_2H_4$), methane ($CH_4$), hydrogen ($H_2$), and acetylene generation rate were consistently the most influential predictors.

*Table 3.5: Top diagnostic features by mean SHAP importance.*

| Rank | Feature | Interpreted Fault Indication |
|---|---|---|
| **1** | $C_2H_2$ | Arcing |
| **2** | $C_2H_4$ | High-temperature thermal faults |
| **3** | $CH_4$ | Moderate overheating |
| **4** | TDCG | Overall gassing activity |
| **5** | $C_2H_2$_rate | Electrical arcing indicator |
| **6** | R5 ($C_2H_6/C_2H_2$) | High-energy discharge |

SHAP global values ranked $C_2H_4$ (|SHAP| = 0.23), $CH_4$ (0.18), and R3 (0.15) as dominant. For Moderate cases, SHAP distributions overlapped with Normal (K–S D = 0.12, p = 0.08), indicating limited gas distinction in early degradation. Local SHAP effects (e.g., $C_2H_4$ > 800 ppm → +0.20 fault probability) showed that moderate degradation often mimics thermal aging, explaining the observed misclassifications.

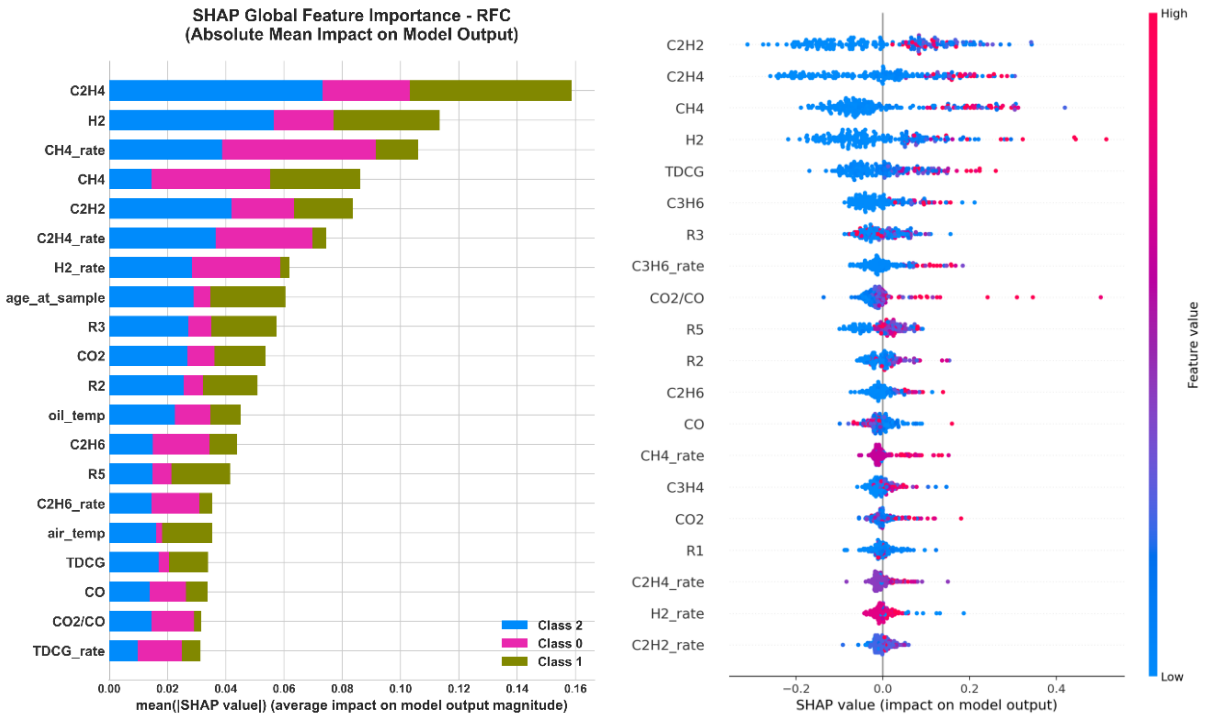


*Figure 3.8: SHAP global feature importance for Random Forest and MLP models.*

## 3.5 Comparison with Classical Diagnostics

Classical rule-based methods IEC 60599 [5] and IEEE C57.104[6] were applied for benchmarking. These methods were overly cautious, because they classify 87.9% of samples as "No Fault," leaving nearly half of the cases unresolved due to their rigid threshold boundaries.

In contrast, the optimized XGBoost model demonstrated superior diagnostic capability. Under stratified cross-validation, the model achieved a validation accuracy of 0.914 ± 0.012 and a macro F1-score of 0.757 ± 0.104, providing definitive and reliable classifications. This result from the model eliminates inconclusive outcomes. Our ML model result reduced the rate of diagnostic failure (inconclusive or incorrect results) by approximately 40% compared to the leading classical method.

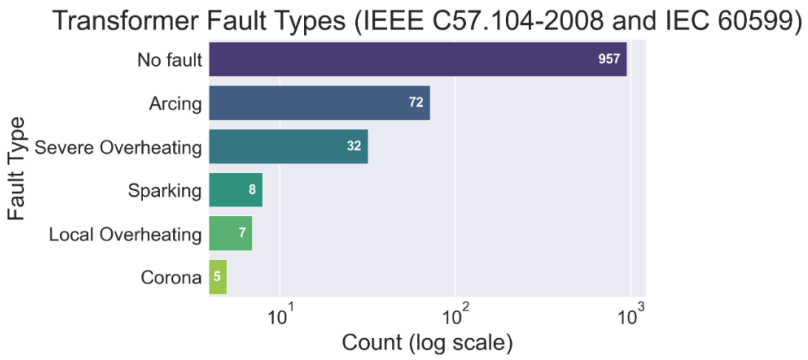


*Figure 3.9: Bar charts comparing diagnostic outcomes of IEC 60599 / IEEE C57.104 methods compared to ML-based results*

# 4 Discussion and Conclusion

The present study reveals a persistent diagnostic bottleneck associated with the Moderate fault class in transformer dissolved gas analysis, arising from intrinsic overlap between Normal and Moderate gas-signature distributions. Quantitative evidence from feature-space projections, Jensen–Shannon divergence, and prediction-uncertainty analysis demonstrates that Moderate degradation states form a transitional continuum rather than a separable class boundary. The observed 76.5 % distribution similarity between Normal and Moderate samples indicates that early degradation is characterized primarily by gradual magnitude changes in key gases, rather than by distinct compositional patterns. This finding aligns with earlier reports highlighting the difficulty of early fault discrimination using static DGA measurements [8,13] and provides a geometric and probabilistic explanation for these limitations.

The superior performance of ensemble learning methods, particularly XGBoost and Extra Trees, is consistent with prior studies on imbalanced DGA datasets [9,10]. However, our per-class evaluation shows that even the best-performing models experience systematic degradation in Moderate-class precision and recall. Importantly, this limitation persists despite extensive hyperparameter optimization and cross-validation, indicating that the bottleneck is not driven solely by class imbalance or model capacity. Similar conclusions were reported by [17], who demonstrated that early and moderate fault states cannot be reliably separated using static gas snapshots and instead require modeling of temporal gas evolution. Our findings complement this work by showing that, under static conditions, Moderate samples are embedded within the Normal feature manifold, leading to low model confidence and high predictive entropy.

The diagnostic “gray zone” identified in this study characterized by low to moderate concentrations of ethene ($C_2H_4$), methane ($CH_4$), and total dissolved combustible gases highlights a fundamental limitation of both machine-learning and classical threshold-based approaches. Classical methods, such as [5, 6], exhibited overly conservative behavior; these methods, assign the majority of samples to “No Fault” or leave them unresolved due to rigid rule boundaries. In contrast, the proposed ML-based framework produced definitive classifications and reduced diagnostic failure rates by approximately 40 %, which corroborates with the recent advances in data-driven DGA interpretation [7,8] techniques. Nonetheless, consistent with the observations of [17], static ML models remain constrained by the absence of temporal context when operating near early-fault boundaries.

Explainability analysis further supports these conclusions. SHAP-based feature attribution confirmed that model decisions align with established fault physics, with ethene, methane, hydrogen, and acetylene-related features dominating predictions. However, SHAP distributions for Moderate cases exhibited substantial overlap with Normal samples, reinforcing the interpretation that early degradation mimics thermal aging rather than manifesting as a distinct fault signature. Elevated predictive uncertainty for Moderate samples provides a quantitative signal of this ambiguity and underscores the need for uncertainty-aware diagnostic strategies in operational settings.

Taken together, these results suggest that overcoming the Moderate-class bottleneck requires diagnostic frameworks that extend beyond static classification. Hence, to address these challenges, domain-adaptive methods such as feature-weighted MMD-CORAL [15], temporal

learning approaches, such as hybrid CNN–LSTM architectures [16], and temporal gas-evolution modeling as demonstrated by [17], offer a promising pathway for resolving early fault ambiguity. Additionally, probabilistic decision thresholds, and calibrated confidence measures may improve robustness across utilities and transformer populations. Future work should prioritize the integration of temporal dynamics, uncertainty quantification, and expert oversight to ensure reliable and ethically aligned deployment of ML-based transformer condition monitoring systems.

# 5 References


[1] X. Zhao, F. Gui, H. Chen, L. Fan, and P. Pan, “Life cycle cost estimation and analysis of transformers based on failure rate,” *Applied Sciences*, vol. 14, no. 3, Art. no. 1210, 2024.

[2] M. Macmillan, K. Wilson, S. Baik, J. P. Carvallo, A. Dubey, and C. A. Holland, “Shedding light on the economic costs of long-duration power outages: A review of resilience assessment methods and strategies,” *Energy Research & Social Science*, vol. 99, Art. no. 103055, 2023.

[3] CIGRÉ, *Analysis of AC transformer reliability*, Technical Brochure 939, 2024.

[4] P. S. Georgilakis, I. Fofana, F. Olivares-Galvan, *et al*., “Environmental cost of transformer losses for industrial and commercial users of transformers,” in *Proc. North American Power Symp. (NAPS)*, 2011.

[5] IEC, *IEC 60599: Mineral Oil-Impregnated Electrical Equipment in Service—Guide to the Interpretation of Dissolved and Free Gases Analysis*, IEC Standard 60599, 2015.

[6] IEEE, *IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers*, IEEE Std C57.104-2008, 2009.

[7] A. Wajid, A. U. Rehman, S. Iqbal, M. Pushkarna, S. M. Hussain, H. Kotb, *et al*., “Comparative performance study of dissolved gas analysis (DGA) methods for identification of faults in power transformer,” *International Journal of Energy Research*, vol. 2023, pp. 1–14, 2023.

[8] H. Cao, C. Zhou, Y. Meng, *et al*., “Advancement in transformer fault diagnosis technology,” *Frontiers in Energy Research*, vol. 12, 2024.

[9] L. Wang, T. Littler, and X. Liu, “Hybrid AI model for power transformer assessment using imbalanced DGA datasets,” *IET Renewable Power Generation*, vol. 17, no. 8, 2023.

[10] L. T. Udeh, A. Pirker, and P. Markus, “AI-supported evaluation of transformer insulating oil data for fault diagnosis,” Carinthia University of Applied Sciences, 2025.

[11] C. Ding, W. Chen, D. Yu, and Y. Yan, “Research on transformer condition prediction based on gas prediction and fault diagnosis,” *Energies*, vol. 17, no. 16, no. 4082, 2024.

[12] T. G. Dietterich, “An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization,” *Machine Learning*, vol. 40, no. 2, 2000.

[13] Suwarno, H. Sutikno, R. A. Prasojo, and A. Abu-Siada, “Machine learning based multi-method interpretation to enhance dissolved gas analysis for power transformer fault diagnosis,” *Heliyon*, vol. 10, no. 4, Art. no. e25975, 2024.

[14] A. Pirker, M. Darmann, L. T. Udeh, and F. Belavić, “Artificial intelligence and classical methods in DGA interpretation: Hybrid approaches for practical transformer condition assessment,” to be published in *Proc. CIGRÉ Session*, Paris, France, 2026.

[15] H. Mahmoodiyan, M. Ahang, M. Abbasi, and H. Najjaran, “Feature-weighted MMD-CORAL for domain adaptation in power transformer fault diagnosis,” *arXiv* preprint, ver. 1, 2025.

[16] A. S. Alhanaf, H. H. Balik, and M. Farsadi, “Enhanced fault classification, detection, and location estimation in the IEEE 14-bus smart grid using a hybrid CNN-LSTM algorithm with adaptive learning rate,” *Arabian Journal for Science and Engineering*, 2025.

[17] M. Xing, W. Ding, H. Li, and T. Zhang, “A power transformer fault prediction method through temporal convolutional network on dissolved gas chromatography data,” *Security and Communication Networks*, pp. 1–11, 2022.