CHARACTERIZING THE MODERATE-CLASS DIAGNOSTIC BOTTLENECK IN ML-BASED TRANSFORMER DEFECT DIAGNOSIS

L.T. UDEH*[1], A. PIRKER[2], M. Prossegger[1]

Introduction

Power transformers are high-value, high-criticality assets. Dissolved gas analysis (DGA) is widely used as a non-invasive indicator of insulation and oil degradation and fault analysis because gases form under thermal and electrical stress. Conventional interpretation schemes (e.g., IEC 60599, IEEE C57.104 and Duval methods) rely on fixed thresholds and ratios. In practice, these rules can become ambiguous under mixed fault mechanisms, operational variability, and low-concentration gassing, which can delay early intervention.

This work shows that ML-based DGA diagnosis does not perform uniformly across fault severities. We focus on the “Moderate” (early degradation) state and demonstrate a systematic bottleneck where Moderate samples are frequently classified as Normal. The study combines transformer-wise splitting and cross-validation to avoid leakage from repeated measurements of the same asset, severity-stratified performance reporting, and quantitative + explainable analyses (feature-space overlap, uncertainty, SHAP) to explain why the bottleneck persists.

Data

A dataset of 110,635 oil samples supplied by VUM Verfahren Umwelt Management GmbH (Austria) was curated to 3,387 power transformer and 19,251 instrument transformer samples after quality filtering. Each sample contains the dissolved gas concentrations (H_2, CH_4, C_2H_2, C_2H_4, C_2H_6, CO, CO_2, O_2, N_2) and additional operational context (e.g., temperatures, service age, voltage). Expert labels define three operational states: Normal (86.1%), Moderate (12.4%), and Fault (1.5%). This distribution reflects the imbalance typical of field data and motivates evaluation procedures that are robust to minority classes.

Methods

Preprocessing removed unlabelled records, records with more than 80% missing values and duplicate timestamps per transformer. Missing values occurred mainly in ambient temperature (≈35%) and in legacy gas measurements; gaps were addressed via transformer-wise averaging and a multivariate iterative imputer based on Random Forest. The Methodology flow diagram is shown in Figure 1.

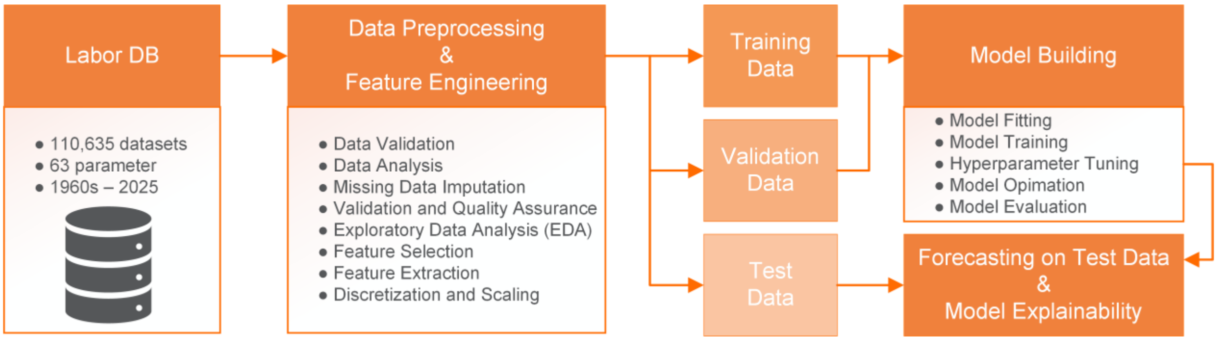

Figure 1: Methodology flow diagram. Pipeline steps from raw data to final explainability outputs.

[1] Carinthia University of Applied Sciences - Applied Data Science, Villach, Austria, tochukwulivinus.udeh@alumni.fh-kaernten.at

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

Feature engineering evaluated nested sets that progressively added established DGA ratios (e.g., CH_4/H_2, C_2H_2/C_2H_4, CO_2/CO), gas generation rates (ppm/month) to approximate dynamics, and temperature variables for operational context. Features were robust-scaled (median/IQR) using training data only.

To prevent asset leakage, data splits and cross-validation were performed transformer-wise. Severe class imbalance (Normal ≫ Moderate ≫ Fault) was handled by comparing imbalance strategies (no correction, class weights, and SMOTE), with SMOTE applied strictly within training folds. Nine supervised classifiers were benchmarked, including linear, instance-based, neural, and ensemble models; macro-averaged F1 was used as the primary metric, complemented by ROC-AUC, PR-AUC, κ and MCC. Explainability used SHAP for global and local feature attributions; PCA and t-SNE visualizations plus Jensen–Shannon divergence (JSD) quantified class overlap; prediction entropy summarized uncertainty.

Key results

Ensemble models performed best overall. XGBoost achieved the highest macro F1-score (0.757 ± 0.073), while Extra Trees delivered the highest accuracy (0.926 ± 0.004) and κ (0.686 ± 0.011).

Despite strong aggregate performance, per-class analysis revealed a consistent Moderate-class bottleneck. Severe Fault cases were detected comparatively reliably (Fault AUC reported ≥ 0.881), but Moderate detection remained suboptimal across architectures. In the validation set, XGBoost misclassified 83.3% of Moderate samples as Normal; several other models showed similar confusion patterns, indicating a systematic limitation rather than a single-model weakness.

Feature-space analysis explains the bottleneck: Moderate samples largely lie within the Normal feature manifold in PCA and t-SNE projections. Quantitatively, Normal–Moderate similarity was highest (76.5%) with the smallest JSD distance (0.235), demonstrating intrinsic overlap. A practical diagnostic “gray zone” was associated with low hydrocarbon concentrations (reported for example as $C_2H_4 < 3.0$ and $CH_4 < 3.2$ in z-score units after scaling), where Moderate and Normal signatures converge.

Uncertainty and explainability further support the interpretation. Moderate samples exhibited higher prediction entropy than Normal, and high-entropy regions aligned with the Normal–Moderate overlap in t-SNE space, suggesting entropy/confidence as a useful operational “review required” indicator. SHAP confirmed physically plausible drivers (notably C_2H_4, CH_4, H_2 and acetylene-related indicators), yet SHAP distributions for Moderate overlap strongly with Normal, explaining limited discriminatory power for early degradation.

As a benchmark, classical IEC 60599 and IEEE C57.104 methods were applied and showed conservative behaviour (large shares of “No Fault” and unresolved outcomes due to rigid thresholds). The ML pipeline produced definitive classifications and reduced diagnostic failure by about 40% compared with the classical benchmark, while still facing the Moderate/Normal ambiguity in the identified gray zone.

Suggestions and outlook

The findings indicate that early/moderate fault detection is limited less by classifier choice than by intrinsic overlap of static DGA signatures between normal conditions and early degradation. For practical asset management, Moderate-level alerts should therefore be interpreted cautiously and supported by calibrated confidence/entropy indicators, contextual operating information, repeat sampling and expert review—especially when predictions fall into the low-gas gray zone. Overcoming the bottleneck will likely require information that breaks the static overlap, most notably temporal evolution of gassing and/or multimodal condition indicators (e.g., operational loading context or complementary diagnostic measurements), combined with uncertainty-aware decision policies.

Keywords

Power Transformer; Condition Assessment; Dissolved Gas Analysis; Fault Diagnosis; Machine Learning; Explainable AI; Class Imbalance; Performance Stratification; Predictive Maintenance.