American Journal of Innovation in Science and Engineering (AJISE)

ISSN: 2158-7205 (ONLINE)

VOLUME 5 ISSUE 1 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Innovation in Science and Engineering (AJISE)

Volume 5 Issue 1, Year 2026
ISSN: 2158-7205 (Online)
DOI: https://doi.org/10.54536/ajise.v5i1.6829
https://journals.e-palli.com/home/index.php/ajise

AI-enabled Prediction of Power Transformer Remaining Useful Life Using Dissolved Gas Analysis and Random Forest Regression

Daniel Kumi Owusu[1*], Curtis Ekow Baidoo[2]

Article Information

Received: November 12, 2025
Accepted: January 29, 2025
Published: February 21, 2026

Keywords

Dissolved Gas Analysis, Power Transformers, Predictive Maintenance, Random Forest Regression, Remaining Useful Life

ABSTRACT

Power transformers represent vital components within electrical power networks, and unexpected breakdowns can lead to significant financial losses and operational disruptions. Dissolved Gas Analysis remains a widely adopted technique for monitoring transformer health, yet most diagnostic methods and machine learning applications concentrate on fault identification rather than ongoing prognostic evaluation. Persistent challenges include the nonlinear nature of fault development, correlations among gas variables, and limited transparency in model outputs. This research introduces a data-driven framework designed to estimate transformer Remaining Useful Life through Dissolved Gas Analysis combined with an optimised Random Forest regression approach. The framework is validated using a publicly accessible dataset containing 2,100 labelled instances of dissolved gas readings, obtained from the Kaggle repository on transformer faults and Remaining Useful Life prediction. Statistical descriptors of hydrogen, carbon monoxide, acetylene, and ethylene are employed after targeted feature selection to address multicollinearity. Relative to traditional Dissolved Gas Analysis interpretation and baseline machine learning techniques, the proposed model achieves higher predictive accuracy, lower estimation error, and enhanced robustness. An analysis of feature importance further differentiates the framework by offering clear insight into the gas parameters most strongly associated with Remaining Useful Life. The findings substantiate the effectiveness of Random Forest regression in delivering dependable, interpretable, and practically applicable predictions of transformer service life.

INTRODUCTION

Power transformers represent essential elements within electricity transmission and distribution networks. Their dependable performance is fundamental to sustaining grid stability and delivering a consistent, high-quality supply of power (Shaqaq & Alghadeer, 2025). Despite this importance, transformers remain susceptible to stresses such as temperature variation, excessive voltage, and load conditions, which may trigger faults including overheating, partial discharge, and low-energy discharge. According to Liu and Yang (2025), when timely diagnostic measures are lacking, transformer failures can escalate into widespread outages, system breakdowns, and equipment degradation, leading to social disruption and substantial financial losses

Predictive monitoring approaches grounded in machine learning have emerged as effective strategies for evaluating transformer condition and reducing operational risks (Khan, 2025). Such methods contribute to improved energy generation planning, enhanced distribution network efficiency, and more reliable market operations. Within the spectrum of available algorithms, Random Forest regression has demonstrated strong potential in predictive analysis owing to its capacity to manage complex datasets with high accuracy. An investigation utilising IEEE data-port records reported that the Random Forest model attained a testing accuracy of 94.4% (David *et al.*, 2023). The algorithm operates by constructing an ensemble of decision trees and combining their outputs, thereby strengthening prediction consistency while also facilitating the assessment of feature importance. These outcomes provide valuable insights into the variables most strongly influencing transformer service life and operational performance (Sintiya *et al.*, 2025).

Advanced diagnostic approaches such as Dissolved Gas Analysis (DGA) offer valuable information on electrical and thermal stresses affecting oil-immersed power transformers. By monitoring dissolved gases in transformer oil, utilities can assess asset condition and anticipate potential failures (Gao *et al.*, 2025). As noted by Radu and Năvrăpescu (2025), DGA has been established as a principal technique for transformer diagnostics, relying on the measurement of fault-indicative gases including Hydrogen (H_2), Carbon monoxide (CO), Acetylene (C_2H_2), and Ethylene (C_2H_4). The integration of machine learning with DGA extends its application beyond conventional fault identification, enabling continuous prediction of Remaining Useful Life (RUL). This advancement supports improved lifecycle management of transformers and facilitates proactive maintenance practices.

Although Random Forest and other machine learning techniques provide notable advantages, several limitations persist. Hyperparameter optimisation, frequently reliant on iterative trial-and-error procedures, can undermine diagnostic reliability and elevate operational expenditure (Yang & Wang, 2025). In addition, the computational demands associated with Random Forest models

[1] Department of Electrical and Electronic Engineering, Faculty of Engineering, Takoradi Technical University, Ghana
[2] Jubilee Technical Training Center, Takoradi Technical University, Ghana
[*] Corresponding author's e-mail: daniel.owusu@ttu.edu.gh

necessitate deliberate resource management during deployment (Biradar *et al.*, 2024). The complex and nonlinear progression of transformer faults is also inadequately represented by conventional DGA methods, including key gas and ratio analyses, which remain restricted by fixed thresholds and linear assumptions (Aciu *et al.*, 2024). Furthermore, existing studies reveal a shortage of comprehensive frameworks capable of linking DGA gas behaviour with precise RUL estimation (Velásquez, 2024).

Even though artificial intelligence has improved diagnostic accuracy, much of the current research remains concentrated on fault identification and categorical health evaluation, with continuous estimation of RUL receiving limited attention (Mashifane *et al.*, 2025). In addition, challenges such as class imbalance, multicollinearity among (DGA) variables, and insufficient feature selection techniques hinder the creation of predictive models that are both dependable and transparent (Ibrahim & Hebala, 2025). To address these constraints, a comprehensive data-driven framework that integrates DGA with advanced machine learning is required, enabling accurate, interpretable, and practically applicable RUL predictions to support proactive transformer lifecycle management.

This research introduces a comprehensive data-driven framework for predicting the RUL of power transformers through DGA combined with Random Forest regression. Conventional diagnostic practices based on fixed thresholds are surpassed by employing targeted feature selection to mitigate multicollinearity, constructing an optimised Random Forest model for continuous RUL estimation, and assessing performance with established regression metrics. In addition, the study highlights the significance of specific gas parameters (H_2, CO, C_2H_4, and C_2H_2) in influencing RUL prediction, thereby improving model transparency and providing valuable guidance for proactive maintenance strategies.

LITERATURE REVIEW

The use of deep learning approaches to forecast transformer health indices and life expectancy was investigated by (El-Rashidy *et al.*, 2025). The study assessed a range of algorithms, including CNN, LSTM, and GRU networks, alongside hybrid configurations, and reported gains in predictive accuracy based on performance metric evaluations. A central focus was placed on the importance of explainable AI methods, particularly LIME and SHAP, which provided transparent insights into model outputs and highlighted the relevance of individual features. These methods strengthened engineering decision-making by clarifying the reasoning processes of complex models. In addition, the research employed hyperparameter optimisation and feature selection techniques to further enhance model effectiveness

Improvements in transformer condition monitoring, particularly the use of energy-autonomous sensors that facilitate continuous remote data acquisition, was explored by (Bajwa *et al.*, 2025). The research underscored the advantages of sensor fusion, the integration of real-time IoT platforms, and the application of digital twins to support predictive maintenance. Although challenges such as financial constraints and data management complexities were identified, the study recommended intelligent monitoring approaches incorporating machine learning to strengthen power system reliability and operational efficiency.

Dladla and Thango (2025) examined machine learning applications in diagnosing power transformer faults through DGA. The review encompassed 124 publications spanning 2014-2024 and showed that China accounted for the largest share of contributions at 53 per cent, with India and Saudi Arabia following. Frequently employed algorithms included support vector machines and artificial neural networks, while publication activity reached its highest levels in 2021 and 2023. The assessment also highlighted notable gaps in the literature, particularly the limited investigation of hybrid modelling approaches and comprehensive performance metrics. The study concluded that further research is required to advance transformer fault detection and strengthen maintenance practices.

Machine learning classifiers for transformer defect diagnosis were evaluated using accuracy, precision, recall, and F1 metrics (Adekunle *et al.*, 2025). When compared to more straightforward algorithms like Naïve Bayes and SVM, the analysis showed that ensemble techniques like Random Forest, XGBoost, and LightGBM produced better results. Moisture has a negative impact on dielectric strength, according to the study's findings, which also highlighted the significance of integrating many diagnostic methods to guarantee accurate transformer problem diagnosis.

Emme and Moola (2025) used DGA data to assess transformer health through data science techniques. Principal Component Analysis (PCA), oversampling, and exploratory analysis were all used in the study. Health indicators were predicted using a variety of machine learning classifiers, such as SVM, Random Forest, XGBoost, and k-NN. Random Forest achieved an accuracy of 96.9%. The results highlighted the significance of feature selection and data balance in assisting predictive maintenance procedures.

Research on transformer condition monitoring employing self-organising neural networks with incremental learning was reported by (Liu *et al.*, 2025). In contrast to traditional techniques that depend on expert involvement, the study proposed a structured methodology incorporating data preprocessing, k-means clustering, and adaptive neural networks. This framework facilitated continuous learning, allowing real-time fault detection to be achieved without imposing extra computational demands.

A review of transformer health monitoring systems and fault detection techniques was presented by (Dubey *et al.*, 2025). The investigation concentrated on IoT-enabled monitoring, employing sensors and Arduino platforms to measure voltage, temperature, and load continuously in real

time. The primary objective was to enhance transformer reliability, facilitate predictive maintenance, and minimise outages through early fault identification. The study also discussed emerging transformer technologies, including smart insulation and AI-based diagnostic tools, alongside advances in electrical engineering such as grid automation and improved fault detection mechanisms.

MATERIALS AND METHODS

This study introduces a data-driven strategy for predicting transformer RUL through the application of DGA data and Random Forest regression. The methodological framework is organised into three sequential stages: Acquisition and Preparation of Data, Selection and Development of Features, and Model Creation and Enhancement. Each stage was structured to promote accuracy, robustness, and interpretability, thereby supporting reliable diagnostic practices for transformers in operation.

Acquisition and Preparation of Data

Origin and Organisation of Data

The file rul_train_features.csv provides a dataset consisting of 2,100 records. Each record is defined by 16 floating-point variables that capture the statistical descriptors (mean, standard deviation, maximum, and minimum) of four dissolved gases, namely H_2, CO, C_2H_2, and C_2H_4. The dataset also includes a target column containing continuous RUL values expressed in operational units, which has been relabelled from its original designation of "predicted".

Assessment of Data Quality

An initial review established that the dataset features were complete, with no missing values detected. Outlier assessment was conducted through z-score evaluation combined with domain-specific criteria, ensuring that irregularities within the data did not distort the modelling process. Records corresponding to genuine extreme operational scenarios were maintained to safeguard the diversity of fault representation.

Exploratory Assessment of Data

The exploratory assessment of data involved a systematic review of the dataset to verify its quality and to examine the significance of individual features, as detailed below:

RUL Distribution

The histogram of the RUL target revealed a bimodal distribution, reflecting the presence of two separate transformer groups. One group was characterised by elevated RUL values in the range of approximately 1,050 to 1,100 units, whereas the other displayed markedly reduced values, falling between about 500 and 550 units.

Feature Distribution

The histograms of individual DGA features revealed right-skewed behaviour in the standard deviation values, whereas the mean measurements presented multimodal distributions indicative of varying fault severities.

Correlation Analysis

The correlation heatmap demonstrated pronounced multicollinearity among the statistical descriptors of each gas, with the strongest associations observed between the minimum, maximum, and mean values within individual species.

Selection and Development of Features

To mitigate multicollinearity, the feature selection strategy retained only the mean and standard deviation values for each gas, discarding the minimum and maximum descriptors that exhibited strong correlations. Subsequent correlation analysis verified that the eight preserved features (H_2_mean, H_2_std, CO_mean, CO_std, C_2H_4_mean, C_2H_4_std, C_2H_2_mean, C_2H_2_std) conveyed unique and non-redundant information pertinent to RUL prediction. For classification purposes, quantile-based discretisation was employed to generate categorical RUL bins.

Model Creation and Enhancement

Data Segmentation

The processed dataset was divided into input features (x) and the target variable (y). To preserve balanced representation across RUL ranges, a stratified partition assigned 80% of the data to training and 20% to testing. Reproducibility of outcomes was secured through random seed initialisation.

Implementation of Random Forest Regression

The Random Forest Regressor was chosen owing to its capacity to limit overfitting, its robustness against noisy data, and its effectiveness in modelling non-linear relationships. Randomisation is introduced through the selection of features at each node split, while predictions are aggregated from numerous decision trees trained on bootstrapped subsets of the dataset.

Construction of Pipelines and Parameter Optimisation

A scikit-learn pipeline was constructed by combining StandardScaler with RandomForestRegressor to normalise the DGA feature parameters. Hyperparameter optimisation was carried out using GridSearchCV, employing neg_mean_squared_error as the evaluation metric and applying 5-fold cross-validation.

Evaluation of Model Performance

The Mean Absolute Error (MAE) measures the mean difference between estimated and observed transformer service lives, functioning as a standard indicator of prediction accuracy within maintenance planning, as presented in equation (1).

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i| \quad (1)$$

e total number of test samples, yi

denotes the actual RUL of the ith transformer, denotes the predicted RUL of the ith transformer, and denotes the absolute difference between the actual and predicted RUL.

In addition, the Root Mean Square Error (RMSE) provides a measure of the variation between predicted and observed RUL values, as illustrated in equation (2).

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (2)$$

total number of test samples, yi signifies the actual RUL, ŷi signifies the predicted RUL, and (yi - i)2 also signifies the squared prediction error for each sample.

Likewise, the explanatory strength of the model was assessed using the Coefficient of Determination (R^2 score), as presented in equation (3).

$$R^2 = 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (3)$$

he test samples., yi represents the actual RUL, represents the predicted RUL, represents the mean of actual RUL values.

Methods for RUL Estimation

The Random Forest algorithm employed DGA profiles to estimate transformer RUL values. Confidence intervals for these predictions were derived from the distribution of outputs across the ensemble trees, providing a probabilistic perspective. To enhance interpretability, the key gas concentration variables influencing RUL estimation were determined through feature importance rankings based on the Mean Decrease in Gini metrics

RESULTS AND DISCUSSION

Correlation Heatmap for Pre-processing Stage

Figure 1 illustrates pronounced multicollinearity among the DGA variables, most evident within the statistical descriptors of individual gases. The deep red diagonal blocks highlight strong positive associations between the mean, standard deviation, maximum, and minimum values of each gas. In contrast, the deep blue squares located in the bottom row indicate that C_2H_4 and C_2H_2 exhibit strong correlations with RUL predictions.

Analysis of Remaining Useful Life Target Value

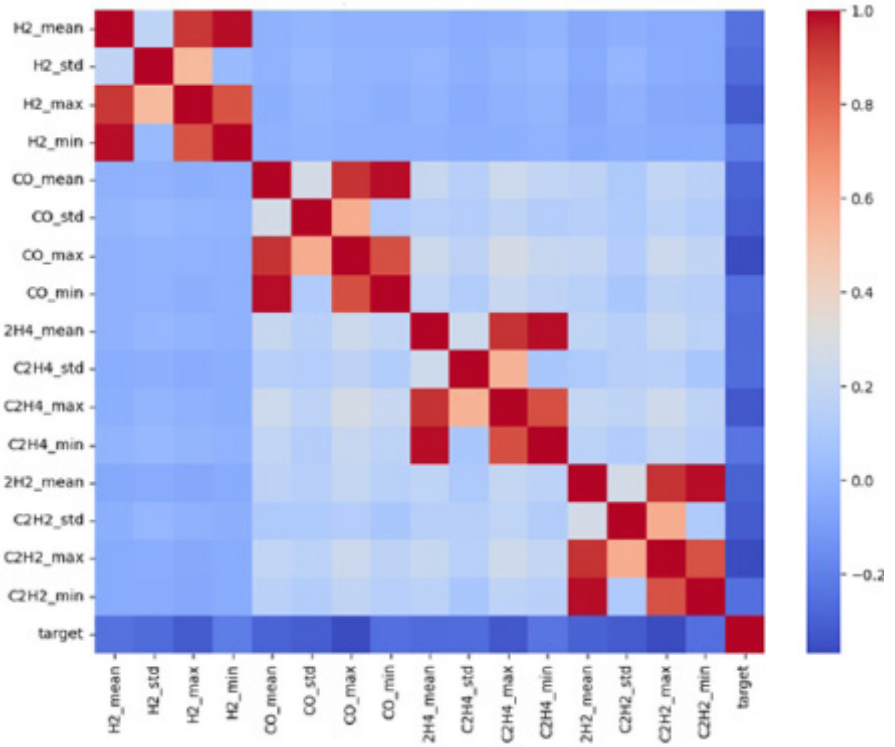

Figure 1: Correlation Heatmap for Pre-treatment Analysis

Distribution

Figure 2 depicts a bimodal distribution of RUL, characterised by two distinct groups: healthy transformers (n>700) concentrated within the 1050-1100 unit range, and degraded transformers (n=300) clustered between 500 and 550 units. The limited presence of values in the intermediate region suggests a pattern of rapid deterioration rather than a progressive decline. This distribution introduces class imbalance, presenting challenges for Random Forest regression in addressing the variability of transformer conditions.

Pattern of C_2H_2 Mean Concentration

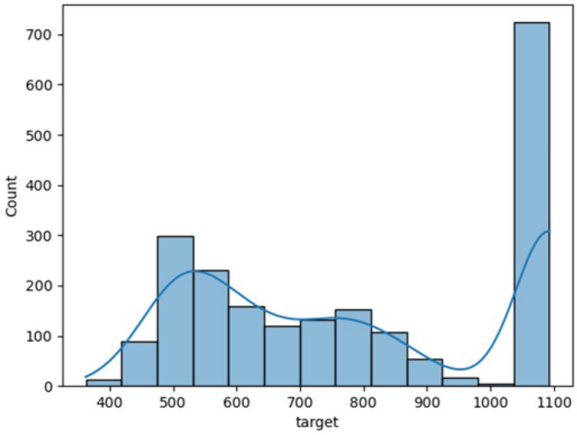

Figure 2: Remaining Useful Life Prediction Targets

Figure 3 presents C_2H_2 concentration values spanning from near zero to 0.0004, with frequencies remaining consistently between 40 and 60 across most intervals. This balanced distribution offers adequate feature diversity for training the Random Forest model. Nevertheless, the pronounced drop beyond 0.0004 reflects the limited occurrence of severe fault conditions, which may constrain the accuracy of predictions for heavily degraded transformers.

Pattern of C_2H_2 Standard Deviation

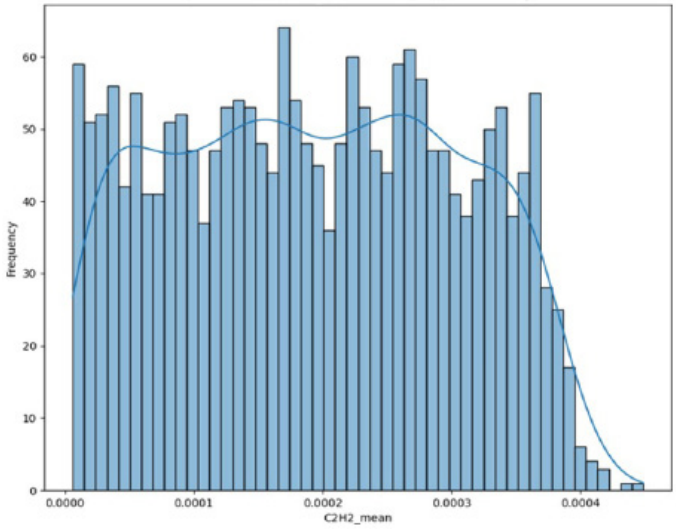

Figure 3: Distribution Pattern of Mean C_2H_2 Concentrations

Figure 4 illustrates a right-skewed distribution of C_2H_2 standard deviation values concentrated close to zero, reflecting stable acetylene behaviour in many transformers with limited evidence of arcing. The pronounced

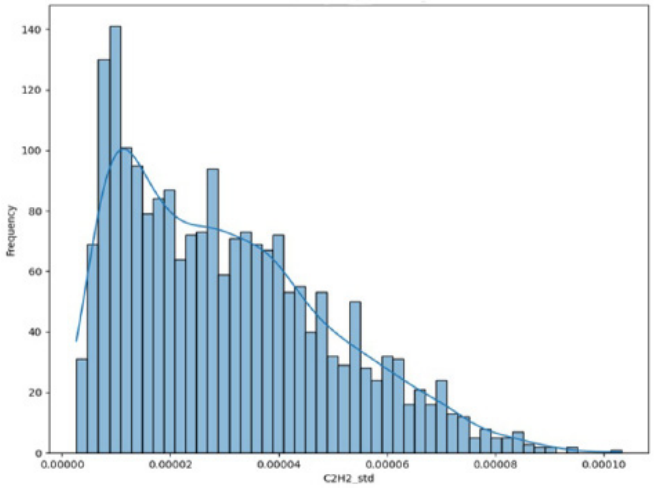

Figure 4: Distribution Pattern of C_2H_2 Standard Deviation

extension of the right tail highlights units exhibiting considerable variability, which points to intermittent fault occurrences. This measure of variability strengthens Random Forest modelling by incorporating temporal dynamics in addition to absolute concentration levels.

Pattern of C_2H_4 Mean Concentration

Figure 5 illustrates a uniform distribution of C_2H_4 concentrations within the range of 0.000-0.010 units, with most intervals exhibiting frequencies between 40 and 60. This balanced representation provides sufficient diversity to support effective Random Forest training across different thermal fault scenarios. Nevertheless, the marked decline beyond 0.009 units reflects the scarcity of high-concentration data, which may constrain predictive accuracy in cases of severe thermal faults.

Pattern of C_2H_4 Standard Deviation

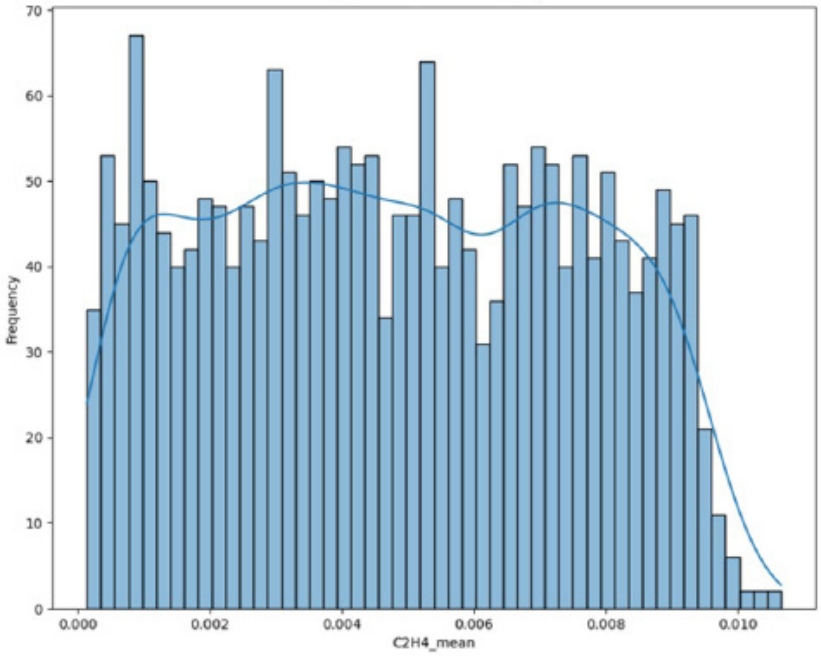

Figure 5: Distribution Pattern of Mean C_2H_4 Concentrations

Figure 6 presents a right-skewed distribution of C_2H_4 standard deviation values, with the majority concentrated within the lower variability range of 0.0001-0.0004. The sharp decline generates an elongated tail extending to 0.0025, signifying that substantial fluctuations in ethylene are uncommon and generally linked to intermittent thermal faults or temporary operating conditions in a limited number of transformers.

Pattern of H_2 Mean Concentration

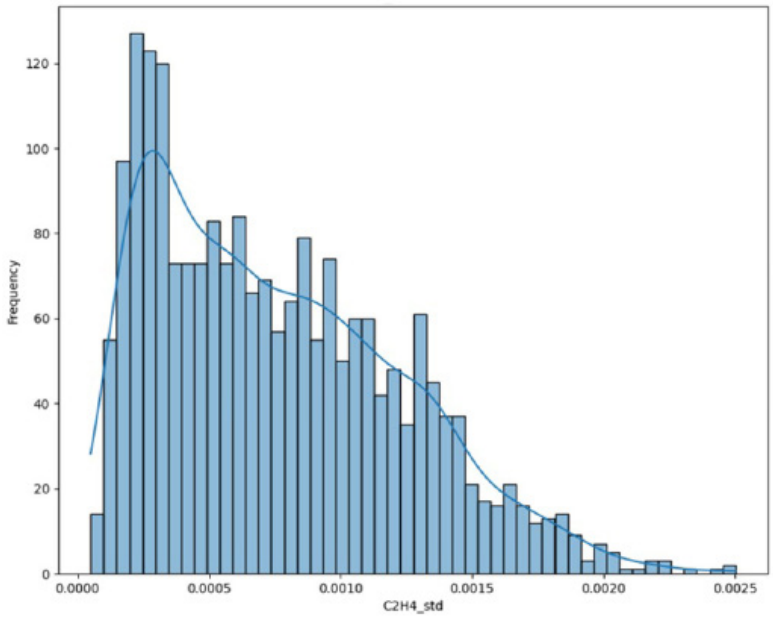

Figure 6: Distribution Pattern of C_2H_4 Standard Deviation

Figure 7 illustrates a uniform distribution of H_2 concentrations within the range of 0.0000-0.0030 units, with most intervals recording frequencies between 40 and 70. This balanced pattern provides adequate variability to support effective Random Forest training across different hydrogen levels. Nonetheless, the steep reduction beyond 0.0030 units highlights the scarcity of high-concentration data, which may limit predictive accuracy for transformers experiencing severe degradation.

Pattern of H_2 Standard Deviation

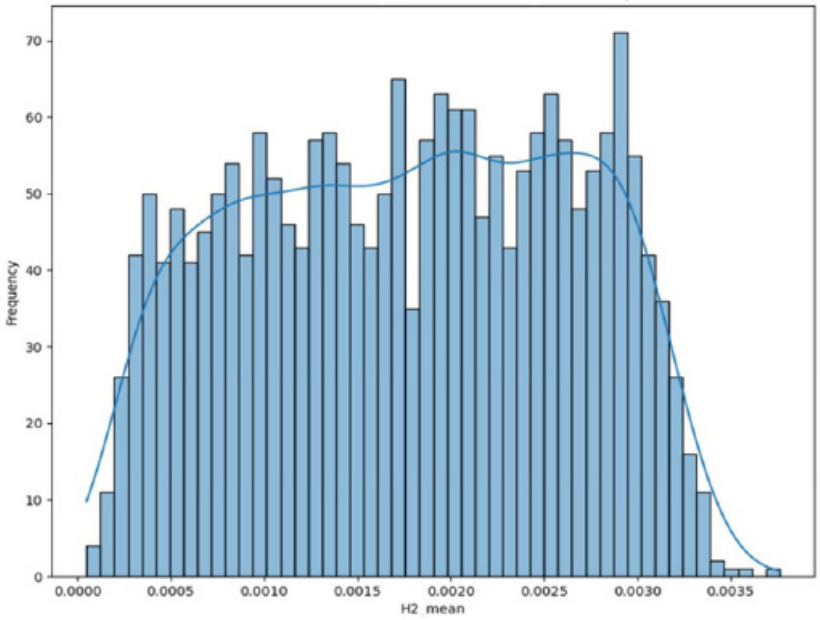

Figure 7: Distribution Pattern of H2 Mean Concentrations

Figure 8 illustrates the distribution of H_2 standard deviations, which displays a pronounced right skew. The majority of values are concentrated at the lower end, reflecting stable hydrogen concentrations across the dataset. As the standard deviation increases, the frequency of observations falls steeply, producing a long tail characterised by few high-value points. This behaviour indicates that large fluctuations in H_2 occur infrequently and are generally associated with rare or unstable fault conditions.

Pattern of CO Standard Deviation

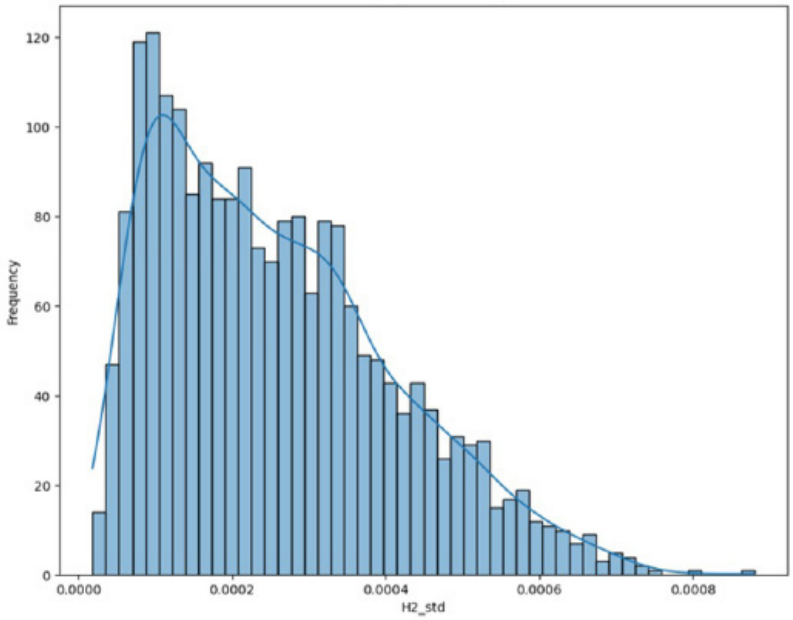

Figure 8: Distribution Pattern of H2 Standard Deviation

Figure 9 presents the frequency distribution of the standard deviations in CO concentration measurements, an important parameter for RUL estimation. The distribution exhibits positive skewness, with most values concentrated at the lower range, reflecting stable CO behaviour characteristic of healthy operating states. As the standard deviation rises, the number of observations decreases markedly, producing a long right tail with limited data points. This pattern indicates that substantial CO fluctuations occur infrequently and are linked to unstable fault scenarios. Such evidence supports

the application of a statistical threshold to separate normal variability from pronounced deviations that signify fault progression and impending risk to RUL.

Correlation Heatmap for Post-Processing Stage

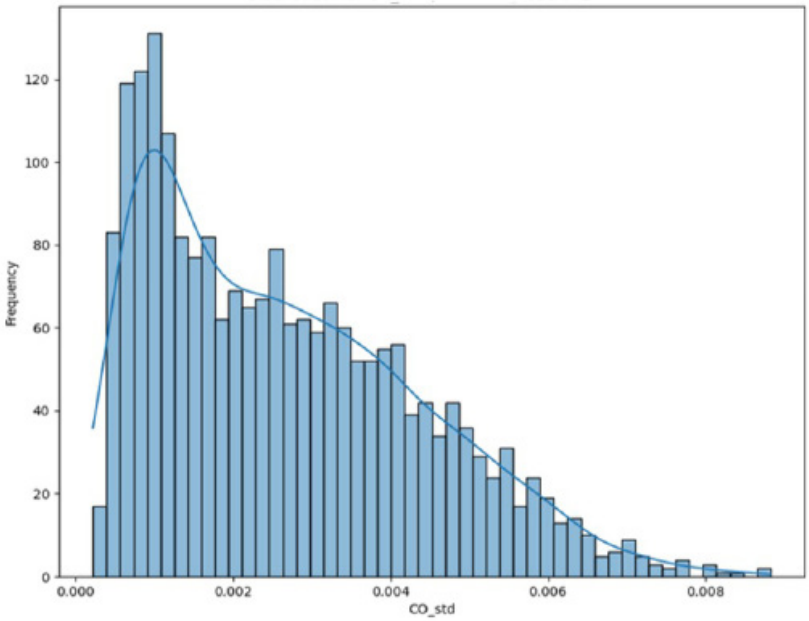

Figure 9: Distribution Pattern of CO Standard Deviation

The correlation heatmap in Figure 10 illustrates the associations between features and the RUL target following the resolution of multicollinearity. Diagonal entries indicate perfect positive correlation, reflecting the self-correlation of each feature. Off-diagonal values have diminished markedly, represented in light blue or grey, which confirms the successful elimination of linear dependencies among predictors. This adjustment enables the Random Forest model to operate with a statistically valid set of independent variables, thereby enhancing both model stability and the reliability of feature importance measures. Correlations with the RUL target remain unchanged, with C_2H_4 mean and C_2H_2 mean continuing to exhibit strong negative relationships. Their significance as predictors of RUL is reinforced, as elevated concentrations of these fault gases signal progression towards failure and reduced remaining useful life.

Evaluation of CO Variability with Respect to RUL

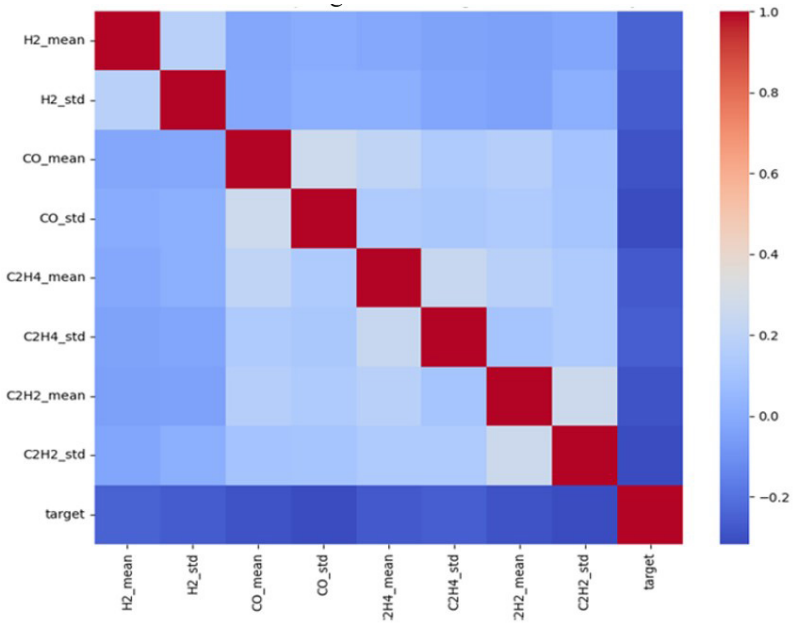

Figure 10: Correlation Heatmap for Post-treatment Analysis

Classes

Figure 11 depicts a boxplot illustrating the distribution of CO standard deviation across transformers classified into Low, Medium, and High RUL groups. The median CO standard deviation shows a distinct downward trend from Low to High RUL, suggesting that reduced variability in CO is associated with extended service performance. Interquartile ranges contract progressively with

increasing RUL, reinforcing the evidence of diminished dispersion in transformers operating under healthier conditions. The presence of numerous outliers within the Medium and High categories indicates that even units with comparatively better RUL may exhibit marked CO fluctuations, pointing to early thermal disturbances that could precede fault development.

Evaluation of C_2H_4 Variability with Respect to RUL Classes

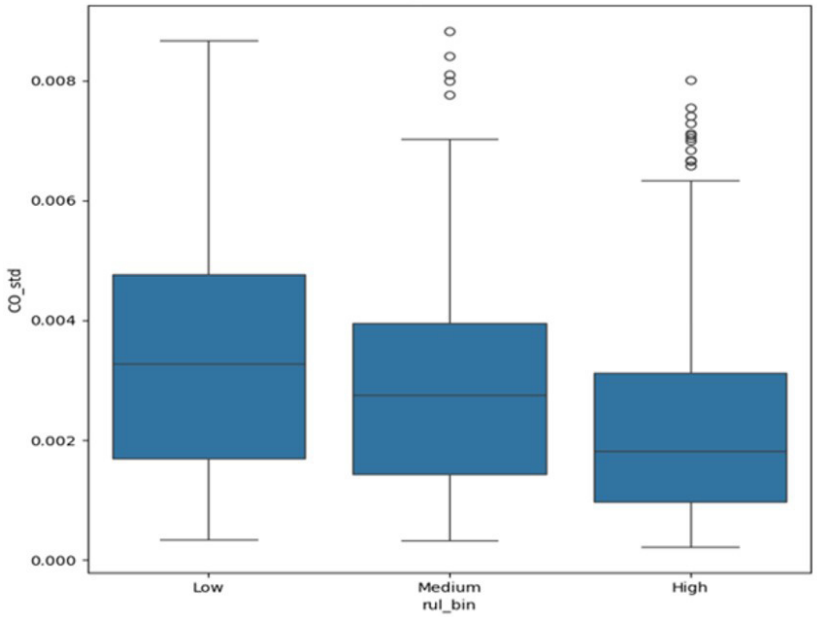

Figure 11: Distribution of CO Standard Deviation by RUL Categories

Figure 12 illustrates the distribution of Ethylene standard deviation (C_2H_4_std) across transformers grouped into Low, Medium, and High RUL categories. An evident inverse relationship emerges, as median C_2H_4_std values decline progressively with increasing RUL. The interquartile ranges follow the same trend, demonstrating that transformers with greater remaining service life exhibit reduced variability and tighter dispersion. This outcome reinforces the role of C_2H_4_std as a reliable inverse prognostic indicator. A considerable number of outliers, most notably within the High RUL group, are noteworthy for anomaly detection. These cases reveal that transformers, even when classified with favourable RUL, can display pronounced temporal fluctuations in Ethylene concentration. Such deviations may reflect short-lived thermal disturbances or the onset of localised faults, thereby serving as potential precursors to fault progression and deserving closer monitoring.

Evaluation of C_2H_2 Variability with Respect to RUL Classes

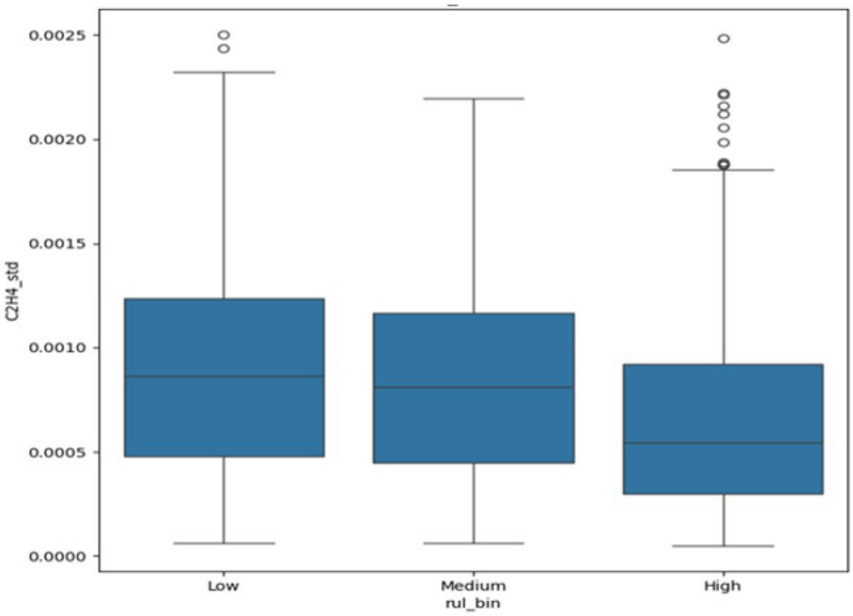

Figure 12: Distribution of C_2H_4 Standard Deviation by RUL Categories

Classes

Figure 13 presents the distribution of Acetylene standard deviation (C_2H_2_std) across transformers classified into Low, Medium, and High RUL groups. The results demonstrate a distinct inverse association, with median C_2H_2_std values declining progressively as RUL increases. The interquartile ranges show a comparable contraction, signifying that transformers with greater remaining service life display reduced variability and more limited temporal dispersion. This behaviour indicates that healthier units sustain steadier operating conditions, consistent with the absence of intense electrical stress. A notable number of outliers appear within the High RUL category, representing transformers that, despite extended remaining life, exhibit marked temporal fluctuations in Acetylene concentration. These deviations serve as early signals of intermittent high-energy discharge phenomena and merit close consideration as warning indicators in RUL monitoring.

Evaluation of H_2 Variability with Respect to RUL

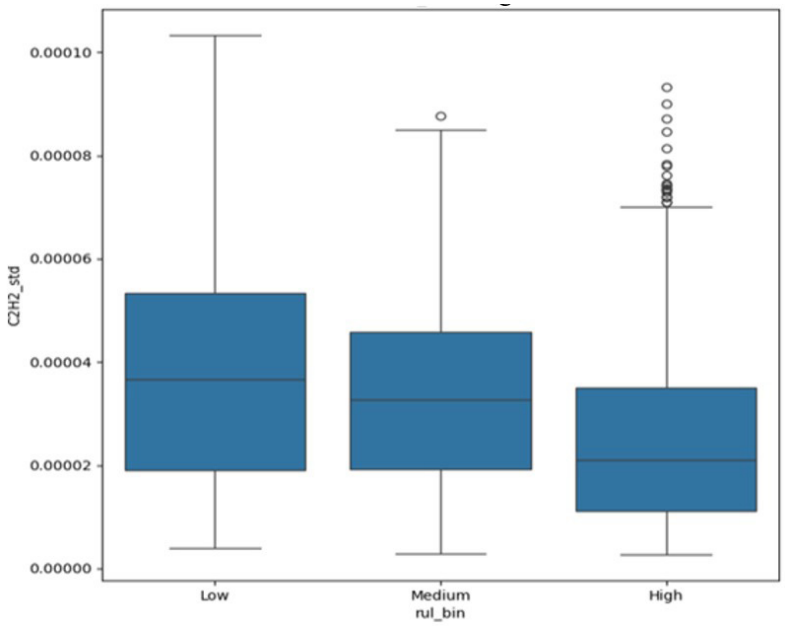

Figure 13: Distribution of C_2H_2 Standard Deviation by RUL Categories

Classes

Figure 14 illustrates the distribution of hydrogen concentration variability (H_2_std) across transformers classified according to RUL categories. The results reveal a distinct inverse association, with median H_2_std values decreasing progressively as RUL rises. A similar trend is observed in the interquartile ranges, which contract with increasing RUL, signifying that transformers with greater remaining service life display reduced variability and more stable behaviour. This outcome indicates that units with

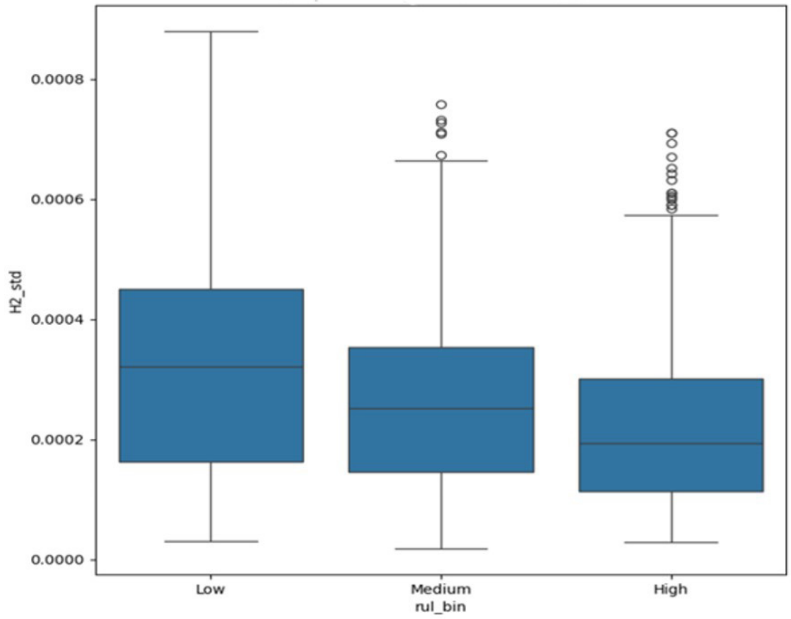

Figure 14: Distribution of H2 Standard Deviation by RUL

extended service potential sustain consistent hydrogen concentrations, aligning with steady operating conditions and the absence of persistent low-energy fault activity. A considerable number of outliers appear within the High RUL group, underscoring their importance for anomaly detection. These cases show that certain transformers, despite favourable RUL classification, still undergo pronounced transient fluctuations in hydrogen levels. Such anomalies may correspond to intermittent fault events or operational irregularities, providing valuable early warning signals of possible degradation.

Evaluation of Mean H_2 Variability with Respect to RUL Classes

Figure 15 depicts the relationship between mean Hydrogen concentration (H_2_mean) and RUL, with data points distinguished by colour coding for Low (blue), Medium (orange), and High (green) classifications. The plot indicates a weak inverse association, showing that RUL decreases marginally as H_2_mean rises. Although accumulated hydrogen concentration is physically consistent with reduced service life, H_2_mean demonstrates limited effectiveness in discriminating between RUL categories because of the considerable overlap across all three groups. This overlap underscores the inadequacy of relying solely on gas mean for RUL estimation and supports the inclusion of variability-based features. The High RUL category displays the clearest pattern, forming a compact horizontal cluster at elevated RUL values. This establishes a reliable baseline for healthy operation and confirms that H_2_mean remains stable in well-performing units, yet lacks the sensitivity required to capture the acceleration of fault progression.

Evaluation of Mean C_2H_2 Variability with Respect to RUL Classes

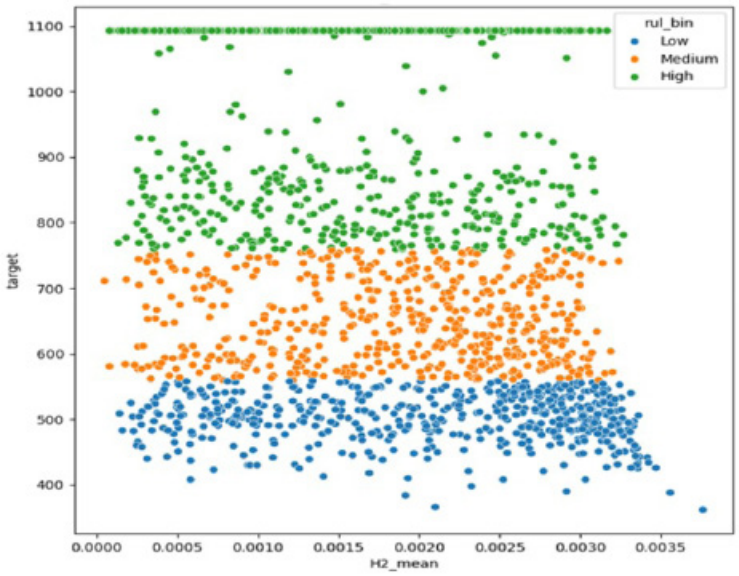

Figure 15: Scatter Plot Depicting Mean H_2 Concentration in RUL Classes

Figure 16 depicts the association between mean Acetylene concentration (C_2H_2_mean) and RUL, with data points distinguished by colour coding for Low (blue), Medium (orange), and High (green) classifications. The plot demonstrates a pronounced inverse correlation, showing

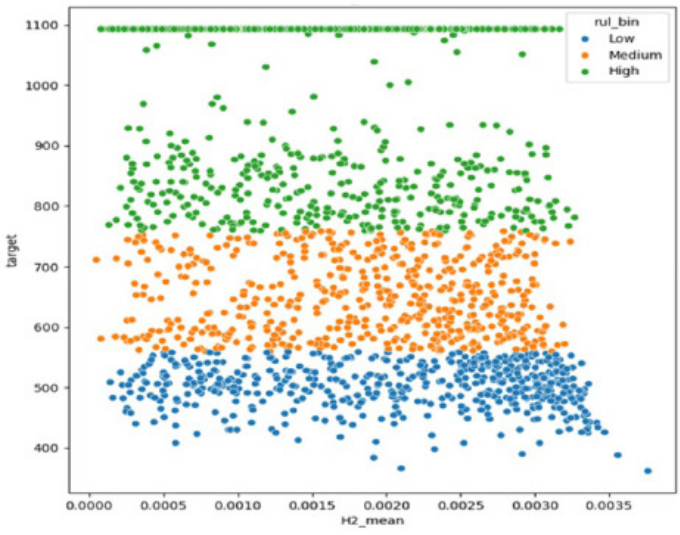

Figure 16: Scatter Plot Depicting Mean C_2H_2 Concentration in RUL Classes

that RUL declines as C_2H_2_mean rises. The three RUL categories form clearly separated horizontal clusters across the range, highlighting the strong stratification achieved. This separation confirms that C_2H_2_mean provides substantial discriminatory capability for RUL classification. The High RUL group appears as a compact horizontal band at the upper end, establishing a reference baseline for healthy operation, while the Low and Medium groups, despite exhibiting greater vertical dispersion, remain distinctly apart. This clustering pattern reinforces the role of Acetylene concentration as a reliable prognostic indicator of transformer condition and validates its inclusion as a significant feature within the predictive model.

Assessment of Model Performance in RUL Prediction

Figure 17 compares the actual RUL values from the test dataset with those predicted by the model, where the dashed red line represents the condition of perfect accuracy. The concentration of data points around the 45° reference line demonstrates the model's strong effectiveness in RUL estimation. Its most consistent performance is evident for healthy transformers (High RUL), with predictions forming a compact vertical cluster near the actual RUL of 1100, confirming outstanding accuracy in this range. Predictions for degraded transformers (Low RUL) display greater spread, yet the overall proximity to the reference line shows that the model continues to capture fault progression with notable reliability. The distribution also reveals systematic

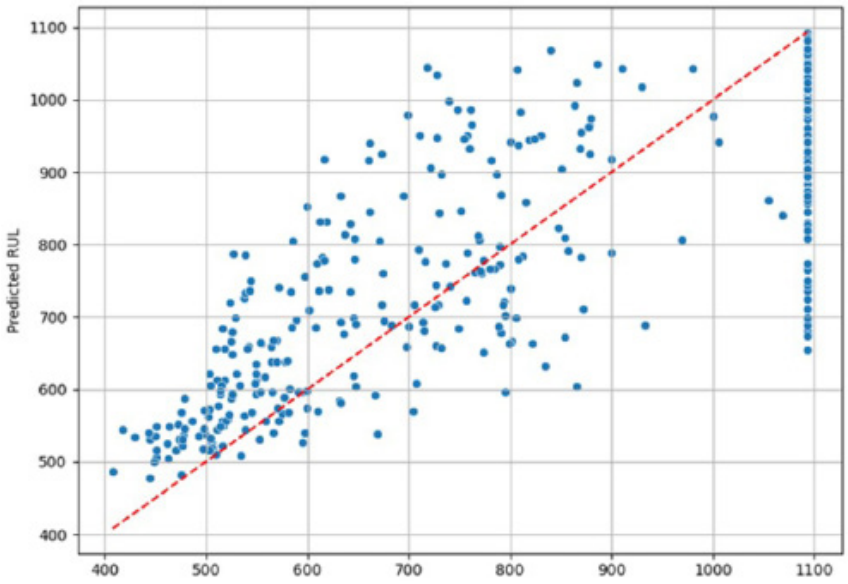

Figure 17: Evaluation of RUL Prediction Accuracy

tendencies: points below the line indicate conservative overestimation of RUL in degraded units, while those above reflect underestimation.

CONCLUSION

The purpose of this study was to create a data-driven framework that integrates Random Forest regression with DGA to forecast the remaining useful life of power transformers. The main goal was to develop a continuous, accurate, and comprehensible estimate of transformer life expectancy, moving away from threshold-based diagnostic interpretation and conventional fault-classification techniques. The findings show that the proposed framework addresses major issues related to class imbalance, multicollinearity, and redundant feature representation while offering dependable RUL prediction performance. The most significant predictors of transformer degradation were found to be the concentrations of acetylene and ethylene, namely their mean values and temporal fluctuation. These results strengthen condition-based evaluation by confirming the diagnostic value of fault gases and offering improved understanding of fault progression processes. The advancement of transformer diagnostics from reactive maintenance to proactive lifecycle management is the study's larger relevance. Utilities can prioritise interventions, reduce unplanned outages, manage maintenance schedules, and enhance asset-replacement planning with accurate RUL estimation. The methodology promotes better risk control, cost-effective asset management, and more resilient grid operation for industry, especially in areas where transformer failure has significant operational and financial ramifications. This work uses an interpretable ensemble learning approach to directly link dissolved gas behaviour to continuous RUL estimate, thereby filling a significant research gap. This study expands on the present awareness of machine learning applications in transformer prognostics and advances data-driven asset-health analysis by offering an organised approach that combines feature selection and model optimisation. There are certain drawbacks to this study. Ageing dynamics may be further influenced by complementary sources like loading conditions, ambient temperature, or past maintenance records, which are not included in the dataset and are limited to DGA features alone. Additionally, a lack of severe-fault case

representation may have an impact on model performance for severely degraded units. Therefore, hybrid or deep learning architectures should be investigated in future study, operational and environmental variables should be added to feature integration, and the framework should be validated using a variety of transformer populations and field-acquired datasets. The proposed approach shows that combining Random Forest regression with dissolved gas monitoring offers a reliable route to precise transformer life prediction and intelligent maintenance scheduling. This study facilitates the shift to predictive asset-management paradigms and makes a significant contribution to the creation of more intelligent, secure, and sustainable power-system infrastructures by permitting accurate prediction of remaining usable life.

Data Availability Statement

Data for this study are available at: https://www.kaggle.com/datasets/yuriykatser/power-transformers-fdd-and-rul?resource=download&select=labels_fdd_test.csv

REFERENCES

Aciu, A. M., Enache, S., & Niţu, M. C. (2024). A Reviewed Turn at of Methods for Determining the Type of Fault in Power Transformers Based On Dissolved Gas Analysis. *Energies, 17*(10), 1-26. https://doi.org/10.3390/en17102331.

Adekunle, A. A., Fofana, I., Picher, P., Rodriguez-Celis, E. M., Arroyo-Fernandez, O. H., Simard, H., & Lavoie, M. A. (2025). Multiclass Fault Diagnosis in Power Transformers Using Dissolved Gas Analysis and Grid Search-optimised Machine Learning. *Energies, 18*(13), 1-30. https://doi.org/10.3390/en18133535.

Bajwa, A., Tonoy, A. A. R., & Khan, M. A. M. (2025). IoT Enabled Condition Monitoring in Power Transformers: A Proposed Model. *Review of Applied Science and Technology, 4*(2), 118-144. https://doi.org/10.63125/3me7hy81.

Biradar, V., Kakeri, D., & Agasti, A. (2024). Machine Learning Based Predictive Maintenance in Distribution Transformers. In *8th International Conference on Computing, Communication, Control and Automation* (pp. 1-5). https://doi:10.1109/ICCUBEA61740.2024.10774993.

David, O, Ebenezer, A., Bridget, M., Isaac, K. O., Andrew, Q., & Kwabena, A. K. (2023). Predictive Health Monitoring of a Power Transformer Using Machine Learning. *Electrical and Electronic Engineering, 13*(1), 12-18. http://doi:10.5923/j.eee.20231301.03.

Dladla, V. M., & Thango, B. A. (2025). Fault Classification in Power Transformers via Dissolved Gas Analysis and Machine Learning Algorithms: A Systematic Literature Review. *Applied Sciences, 15*(5), 1-43. https://doi.org/10.3390/app15052395.

Dubey, P. K., Singh, B, Patel, D. K., & Kumar, R. (2025). Review of Transformer Health Monitoring System and Fault Detection. *International Journal of Sciences and Innovation Engineering, 2*(6), 209-223. https://doi.org/10.70849/IJSCI.

El-Rashidy, N., Sultan, Y. A., & Ali, Z. H. (2025). Predicting Power Transformer Health Index And Life Expectation Based On Digital Twins and Multitask LSTM-GRU Model. *Scientific Reports, 15*(1), 1-29. https://doi.org/10.1038/s41598-024-83220-x.

Emme, S. S., & Moola, P. R. (2025). Application of Data Science Techniques and Machine Learning Based Classifiers for Transformer Health Assessment. *Iraqi Journal of Science, 66*(6), 2581-2592. https://doi.org/10.24996/ijs.2025.66.6.31.

Gao, Z., Yu, B., Guang, J., Jiang, S., Cong, X., Zhang, M., & Yu, L. (2025). Predicting the Remaining Service Life of Power Transformers Using Machine Learning. *Processes, 13*(11), 1-19. https://doi.org/10.3390/pr13113459.

Ibrahim, R. A., & Hebala, A. (2025). A Feature-Enhanced Approach to Dissolved Gas Analysis for Power Transformer Health Prediction Through Interpretable Ensemble Learning and Multi-model Evaluation. *Technologies, 14*(1), 1-30. https://doi.org/10.3390/technologies14010006.

Khan, M. A. M. (2025). AI and Machine Learning in Transformer Fault Diagnosis: A Systematic Review. *American Journal of Advanced Technology and Engineering Solutions. 1*(1), 290-318. http://dx.doi.org/10.2139/ssrn.5190585.

Liu, C., & Yang, W. (2025). Transformer Fault Diagnosis Using Machine Learning: A Method Combining SHAP Feature Selection and Intelligent Optimisation of LGBM. *Energy Informatics, 8*(52), 1-19. https://doi.org/10.1186/s42162-025-00519-3.

Liu, S., Xie, Z., & Hu, Z. (2025). DGA-based Fault Diagnosis Using Self-organising Neural Networks with Incremental Learning. *Electronics, 14*(3), 1-17. https://doi.org/10.3390/electronics14030424

Mashifane, L. D., Mendu, B., & Monchusi, B. B. (2025). State-of-the-Art Fault Detection and Diagnosis in Power Transformers: A Review of Machine Learning and Hybrid Methods. *IEEE Access, 13*(2), 48156-48172. https://doi: 10.1109/ACCESS.2025.3546861.

Radu, D. G., & Năvrăpescu, V. (2025). Optimising Energy Efficiency and Monitoring Methods for Power Transformers." In *International Aegean Conference on Electrical Machines and Power Electronics and International Conference on Optimisation of Electrical and Electronic Equipment* (pp. 1-8). https://ieeexplore.ieee.org/document/11075283.

Shaqaq, S. A., & Alghadeer, A. A. (2025). Enhancing Power Transformer Reliability Through Real-Time Machine Learning Monitoring. In *15th International Conference on Power, Energy, and Electrical Engineering* (pp. 22-25). https://ieeexplore.ieee.org/document/10987327.

Sintiya, E. S., Prasojo, A. R., & Hidayah, H. K. (2025). Application of Machine Learning for Predictive Maintenance in Power Transformer Health Assessment: A Comparative Study of Support Vector Machine, Artificial Neural Network, and Random

Forest. In *Proceeding International Seminar of Science and Technology* (pp. 70-83). https://doi.org/10.33830/isst.v4i1.5233.

Velásquez, R. M. A. (2024). A Comprehensive Analysis for Wind Turbine Transformer and Its Limits in The Dissolved Gas Evaluation. *Heliyon, 10*(20), 1-24. https://doi.org/10.1016/j.heliyon.2024.e39449.

Yang, Y., & Wang, H. (2025). Random Forest-based Machine Failure Prediction: A Performance Comparison. *Applied Sciences, 15*(16), 1-22. https://doi.org/10.3390/app15168841.