

The Improvement of SMOTE-ENN-XGBoost through Yeo-Johnson Strategy on Dissolved Gas Analysis Dataset

Abstract

Power transformers are critical components of the power system, with their reliable operation ensuring grid stability. Dissolved Gas Analysis (DGA) is widely used to diagnose transformer faults by analysing gas concentrations in oil samples. However, overlapping and imbalanced data samples present significant challenges in accurately classifying transformer faults. This study proposes an integrated strategy combining feature transformation and data-level resampling to address these issues. Five transformation techniques, Log, Log1p, Square Root, Box-Cox, and Yeo-Johnson, were explored to stabilise data and reduce class overlap. Resampling techniques, including Synthetic Minority Oversampling Technique (SMOTE), Edited Nearest Neighbors (ENN), and SMOTE-ENN, were employed to balance the dataset and eliminate noisy samples. The XGBoost classifier was used to evaluate performance. Results showed that the combination of Yeo-Johnson transformation and SMOTE-ENN yielded the highest accuracy of 93.20%, outperforming the baseline accuracy of 71.30%. This strategy demonstrates the effectiveness of noise reduction after synthetic balancing for improving classification performance. Future research will focus on scaling the methodology for larger datasets, handling categorical variables, and exploring alternative distance measures to further enhance transformer fault diagnosis.

Keywords

Power Transformer, Dissolved Gas Analysis, Data-Level Technique, Feature Transformation, Fault Classification

1. Introduction

As the main part of the power system, the power transformer helps to convert AC voltage and current so that AC power can be transmitted. During abnormal operation, thermal, electrical, and mechanical stressors can cause an oil-immersed power transformer to produce flammable gasses. This gas output directly impacts the power system's overall stability and safety. When there are hidden faults in the transformer, the state of the dissolved gas in the oil gradually changes (J. Liu et al., 2022). To track and determine the power grid's operating state, a transformer plays an essential role (X. Wang et al., 2022). Due to its ease of use and compatibility with interactive evaluation, Dissolved Gas Analysis (DGA) is the approach for expecting power transformer problems (Taha & Mansour, 2021). There are multiple steps in the DGA process. First, oil samples are taken out of the transformer. These samples are then put into gas chromatography to separate and examine the different gas components contained in the oil sample. It measures the concentration of each element and determines the mixture's substance composition (Flanagan et al., 2008). There are several types of sensors used in gas chromatography, including high-sensitivity and general-purpose sensors. A variety of sensors, including high-sensitivity and general-purpose sensors, are used in gas chromatography. During the analytical process, these sensors play an important part in identifying and measuring the various components contained in a gas mixture.

While high-sensitivity sensors are especially skilled at detecting insignificant levels of particular components, providing a more thorough and accurate analysis, general-purpose sensors offer a wide composition overview. The accuracy and adaptability of gas chromatography in determining the chemical composition of samples are improved by this combination of sensors. Specialised software is installed on computer records and captures the signal the detectors produce inside the gas chromatograph. The results of the separated gas values are shown on a dashboard that the program displays on the screen (Chanchotisation & Vong, 2021). The concentration of potential fault gases, such as hydrogen (H_2), methane (CH_4), ethylene (C_2H_4), ethane (C_2H_6), acetylene (C_2H_2),

carbon dioxide (CO₂), carbon monoxide (CO), and others, is the main information provided by the DGA data (Das et al., 2023a).

In the recent two decades, extensive research and implementation of artificial intelligence (AI) techniques in transformer insulation detection has been done. Several methods, including support vector machine (SVM) (Cui et al., 2014; Xinghui Li et al., 2022), artificial neural network (ANN) (X. Wang et al., 2022), and fuzzy logic (Abu-Siada & Hmood, 2015), have been extensively researched and used for this purpose. The advantage of an AI algorithm is that it uses the oil characteristics from testing on the transformer in issue and a historical dataset of transformers. The AI algorithm may learn from this dataset and detect relationships between oil qualities and transformer insulation state. It enables the AI system to compile a large amount of historical data to diagnose the state of the transformer under consideration accurately. (Cui et al. 2014).

A highly skewed distribution of classes or categories is referred to as an imbalanced dataset (Saurabh Tewari & U.D. Dwivedi, 2019). Adaptable evaluations are severely limited by problems with the DGA dataset, including imbalance, insufficiency and overlap (Taha & Mansour, 2021). A dataset that has more data in one category than another is said to be imbalanced. The DGA data is an example of a dataset that is imbalanced. A class imbalance within the initial sample set may result from the difference in the number of samples for each transformer state. Consequently, the development of techniques for detection suitable for imbalanced and inadequately analyzed DGA datasets is important. The classification model may perform worse as a result of the little research on handling imbalanced data, particularly when it comes to processing imbalanced DGA classification data at the data level (Y. Yuan et al., 2023). As a result, machine learning classifier has been difficult with the imbalanced dataset, which often results in inaccurate error models being used for classification (Ebenezer et al., 2021).

Data transformation is a critical data preprocessing step. It converts the data nonlinearly to make a skewed distribution more symmetric (Kvalheim et al., 1994). It uses a mathematical function that is hardcoded to project the data descriptors to a different feature space. For

example, the usage of kernel functions in SVM classifiers, which seek to enhance the separability property to enable effective hyperplane localization (Mathew, Pang, Luo, & Leong, 2017). Another example of data transformation can be found in the latent representations produced by the hidden layers of Artificial Neural Networks (ANN) (Becker et al., 2020). The aim here is to enhance the compactness of class samples by abstracting the input data away from its specific details. This process highlights a key difference between SVM and ANN. While SVM focuses on increasing the separability of data, ANN focuses on improving the compactness of the data within its feature space.

There are two techniques for handling imbalanced datasets at the data level (Seitanidis et al., 2022). The data level is taken into account using oversampling and undersampling techniques to balance the data (Sabha et al., 2023). The Synthetic Minority Oversampling Technique (SMOTE) is a widely used oversampling technique (Kusdiyanto & Pristyanto, 2022; Sağlam & Cengiz, 2022; L. Wang et al., 2021; X. Zhu et al., 2023). It involves producing synthetic instances of the minority class to balance the class distribution in a dataset. The SMOTE algorithm's main idea is to resample the minority class using synthetic samples. For each instance in the minority class, synthetic instances are generated along the line segments that connect it to its nearest neighbours in the minority class. This technique tries to address the class distribution imbalance by increasing the number of minority class instances (Kovács et al., 2020). The classic SMOTE technique increases fitting performance (Bao & Yang, 2023). Then, SMOTE can generate noise, perhaps causing synthetic data samples from the minority class to be misidentified as part of the majority (Wah et al., 2023).

The Edited Nearest Neighbors (ENN) under-sampling involves deleting samples if their nearest neighbours differ in category (Y. Zhu et al., 2020). The ENN technique can balance the model's performance on the minority class against its performance on all classes, resulting in greater accuracy (Fan et al., 2023). ENN can detect and eliminate data that is out of boundary to make the decision boundary smoother (J. Wu et al., 2021). The ENN eliminates samples that are incorrectly categorized by the k-nearest neighbour method (default three nearest neighbours) and tends to delete more samples than the Tomek (D.

Liu et al., 2023). ENN eliminates samples that are out-of-boundary by the k nearest neighbour technique. This leads to a smoother decision boundary and more accurate classification (J. Wu et al., 2021). However, the number of noise samples deleted by ENN is limited (Xu et al., 2020). This issue occurs because the data resampling process is independent of afterwards classification algorithms and does not fully consider the inherent features of samples (Peng et al., 2019).

Fundamentally, resampling techniques for handling imbalanced datasets such as ENN, and SMOTE, rely on distance functions, particularly through k-nearest Neighbors (KNN) calculations. A key limitation of distance functions is their narrow focus on individual features, which often neglects the interrelationships and combined effects of multiple attributes within the data. This limitation can hinder their ability to accurately reflect the true similarities or differences between samples (Jiao, Geng, & Pan, 2019). Accurate distance metrics are crucial for the optimal performance of many machine learning and data mining algorithms. Techniques like K-means clustering, nearest-neighbour classifiers, and kernel-based methods such as SVMs rely on these metrics to effectively capture the underlying relationships within the data (Xing et al., 2002).

An XGBoost has been developed to evaluate the transformer status and analyse the DGA dataset. This algorithm is a powerful machine-learning technique often used for classification and regression problems (Gautam et al., 2023). This version of the gradient boosting algorithm is improved and optimised. XGBoost is well known for its outstanding performance on a variety of datasets, speed, and scalability (Raichura et al., 2021).

Power transformer fault classification faces significant challenges due to the imbalance and overlapping of samples within the DGA dataset. The dataset suffers from class skewness and overlapping, which undermine the effectiveness of resampling strategies, particularly KNN-based methods. These issues hinder accurate fault classification, necessitating the development of improved techniques to enhance class separability, mitigate skewness, and ensure balanced data distribution for reliable fault diagnosis. Thus, this study aims to address the challenges of overlapping and imbalance samples in power transformer fault

classification by proposing a novel technique that integrates data transformation and resampling. There are three contributions of this study:

1. The implementation of Yeo-Johnson feature transformation to improve class separability and mitigate skewness to enhance the effectiveness of SMOTE-ENN.
2. A thorough comparative analysis of Yeo-Johnson with four other different feature transformation strategies was evaluated and compared using Karl Pearson's Second Coefficient of Skewness formula to demonstrate the improvement in the skewness of DGA dataset.
3. SMOTE was employed to generate synthetic samples for transformer fault classes with scarce data, addressing class imbalance effectively. Following this, the ENN algorithm was applied to refine the dataset by removing marginal and noisy samples to ensure a balanced and high-quality distribution across fault classes.

2. Related Works

Feature transformation is critical in obtaining meaningful insights from oil and gas datasets, which are frequently characterized by high dimensionality (Hernández-Carnerero et al., 2023; Zhu et al., 2020), missing values (Santos et al., 2020, 2022; Su et al., 2022), and skewed distributions (Ragab et al., 2021; Singh Rawat & Kumar Mishra, 2022). Research has addressed these issues by using preprocessing (Pei et al., 2020; Sahraoui et al., 2023; L. Xu et al., 2023), and feature engineering (Chanchotisatien & Vong, 2021; Jahan et al., 2021) strategies to improve model performance.

In addition, feature transformations that preserve the original data space are confined to feature selection, normalization, and scaling (Elmorshedy et al., 2022). This strategy of preprocessing employment enhances the prediction performance of a machine learning model on a dataset by applying a transformation function to the feature space (Zheng & Casari, 2018). Several studies have explored transformations such as Yeo-Johnson, Log, Log1p, Square Root, and Box-Cox, demonstrating their effectiveness across diverse domains. For non-linear standards data, the Box-Cox transformation is recommended. The Box-Cox transformation is a simple and effective statistical strategy for reducing bias in

linear standardization models that employ untransformed data (J. Z. Xu et al., 2006). Similarly, the Yeo-Johnson transformation expands Box-Cox to handle both positive and negative values, providing additional flexibility in datasets with mixed signs (Osborne, 2016). The Yeo-Johnson transformation outperformed the validation set for viscosity prediction, with a mean absolute percentage error of 5.3%, a root-mean-squared error of 0.23, and a coefficient of determination (R^2) of 0.9404 utilizing only 10 latent variables (Caceres-Martinez & Kilaz, 2024). Logarithmic transformations, including Log and Log1p, have been widely employed to reduce skewness in data exhibiting exponential growth patterns. Log transformations are especially good for compressing large values, while Log1p is beneficial for datasets with little or zero values, ensuring numerical stability (Osborne, 2016). The square root transformation is another simple but likely good strategy widely used in counting data or data with non-negative values to reduce skewness and stabilize variance (Osborne, 2016).

In applied contexts like oil and gas, these transformations have proven critical for processing variables such as production rates, reservoir parameters, and seismic data. For example, Yeo-Johnson and Log1p have successfully normalised data with zero or negative values. At the same time, Square Root and Box-Cox transformations have improved the interpretability and stability of prediction models. These strategies not only increase model performance by correcting skewness and outliers, but they also ensure that feature distributions are consistent with machine learning assumptions. The efficiency of these transformations highlights their importance in data preprocessing, especially in areas with complex, skewed, or non-normal datasets.

Resampling techniques play a crucial role in addressing imbalanced datasets, where the disproportionate class distribution can hinder the performance of machine learning algorithms. Oversampling methods like SMOTE generate synthetic samples by interpolating between existing instances of the minority class. This technique effectively enhances the representation of the minority class, mitigating class imbalance issues (Kusdiyanto & Pristyanto, 2022; Sağlam & Cengiz, 2022; Kovács et al., 2020). Designed to address class imbalance concerns in tiny datasets (Xi et al., 2025). SMOTE can reduce

class imbalance in low-dimensional feature space (Patil et al., 2020) and the active learning algorithm, which incorporates SVM for class-imbalance learning, produced exceptional results (Patil et al., 2020).

In terms of boosting techniques, XGBoost is a scalable machine learning system. XGBoost has a significant impact on machine learning and data mining difficulties (Wang et al., 2021). Researchers used the XGBoost algorithm to investigate the correlation between steel performance, composition, and manufacturing characteristics and compared it to other machine-learning models (Sheng & Yu, 2022). Research indicates that certain machine-learning models perform better than others (Sheng & Yu, 2022). XGBoost is a massively parallel-boosted tree tool. XGBoost is the quickest and best open-source boosted tree toolkit available, with speeds more than ten times quicker than popular toolkits. This makes it ideal for solving industrial-scale challenges. Scalability is key to XGBoost's performance in all scenarios (Wang et al., 2021). The XGBoost algorithm is advantageous for its speed and performance (Li et al., 2020).

3. Proposed Methodology

The dataset for this article was collected from National Grid U.K. and DGALab consisted of 790 samples. The dataset has 6 features, including 5 numerical columns and 1 categorical column. Figure 1 shows the flowchart for the proposed methodology.

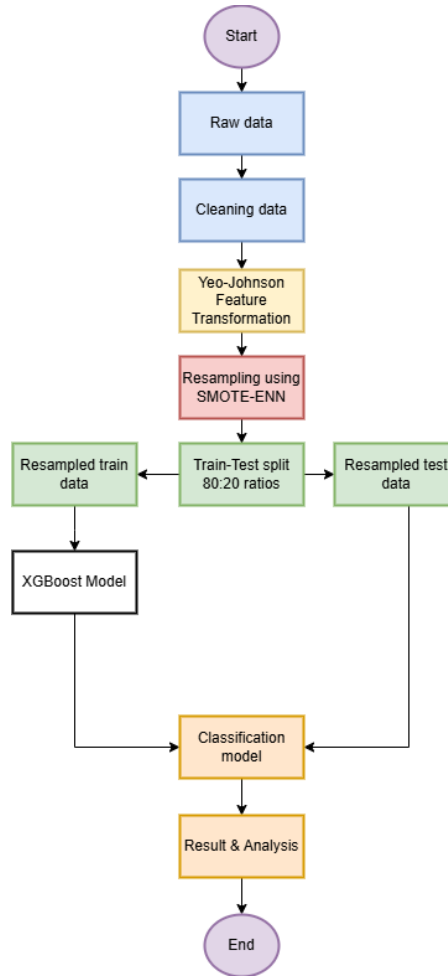


Figure 1. Flowchart for the proposed methodology.

A. Cleaning the data

The dataset contained four duplicate rows. Features with duplicate values were dropped. The dataset reduced the total number from 790 to 781 samples. The dataset includes six gas properties: hydrogen (H₂), methane (CH₄), ethane (C₂H₆), ethylene (C₂H₄), acetylene (C₂H₂), and a category variable 'act' for fault type classification. The presence and concentration of these gases are important indicators for power transformers, providing insights into their health and spotting potential defects. In an imbalanced data collection, the sample size varies significantly across categories. Transformer faults are rare from initial use to decommissioning, and the frequency of different fault types varies significantly. As a result, monitoring equipment detects only a small amount of data and

distinguishes between fault types. The transformer oil fault dataset contains 203 high-energy discharge fault samples (D2), more than any other type of fault (Table 2). This suggests an imbalanced dataset. However, most transformer fault diagnostic techniques require balanced input data.

Power transformer fault classes data description shown in Table 1. Partial Discharges (PD) are low-energy discharges within a transformer that are frequently associated with localized ionization of gas or oil and indicate insulating problems. High-energy discharges are divided into two categories includes D1 and D2. D1 refers to high-energy partial discharges that produce significant electrical stress on the insulating system, whereas D2 denotes even higher energy levels, frequently resulting in arcing and extensive damage. Thermal faults are characterized according to temperature ranges. T1 refers to mild thermal faults that occur at temperatures below 300 °C. They are often produced by slight overheating via inadequate connections or localized hot spots. T2 medium thermal faults occur between 300 °C and 700 °C and are commonly related to sustained high loads or insulation degradation. T3 thermal faults occur at temperatures above 700 °C, resulting in severe effects including insulation charring, carbonization, or material collapse. These classifications aid in determining maintenance priorities and ensuring transformer reliability.

Table 1. Classes of Faults from the DGA Dataset

| Fault Type | Description |
|------------|--|
| PD | Low-energy partial discharge |
| D1 | High-energy partial discharge |
| D2 | High-energy discharge |
| T1 | Low thermal fault (<300 °C) |
| T2 | Medium thermal fault (300 °C - 700 °C) |
| T3 | High thermal fault (>700 °C) |

Table 2 presents the class distribution of the DGA dataset before and after removing duplicate samples. It includes six fault classes of power transformers fault classes (PD, D1, D2, T1, T2, and T3) and shows the total number of samples in each category. Initially, the dataset consisted of 790 samples, but after duplicate removal, it was reduced to 781. Notably, the class distribution remained unchanged for D1, D2, T1, and T2, indicating no duplicates in these categories. However, the counts for PD and T3 decreased slightly, from 90 to 86 and 142 to 137, respectively, reflecting duplicates in these classes. This reduction ensures a cleaner dataset, minimizing redundancy and enhancing the quality of subsequent analysis while reducing the risk of bias in machine learning models. This refined dataset, containing 781 samples, is used as the basis for the experiments conducted in this study, ensuring a cleaner and more reliable foundation for analysis and machine learning model evaluation.

Table 2. The Class Count Before And After Remove Duplicate

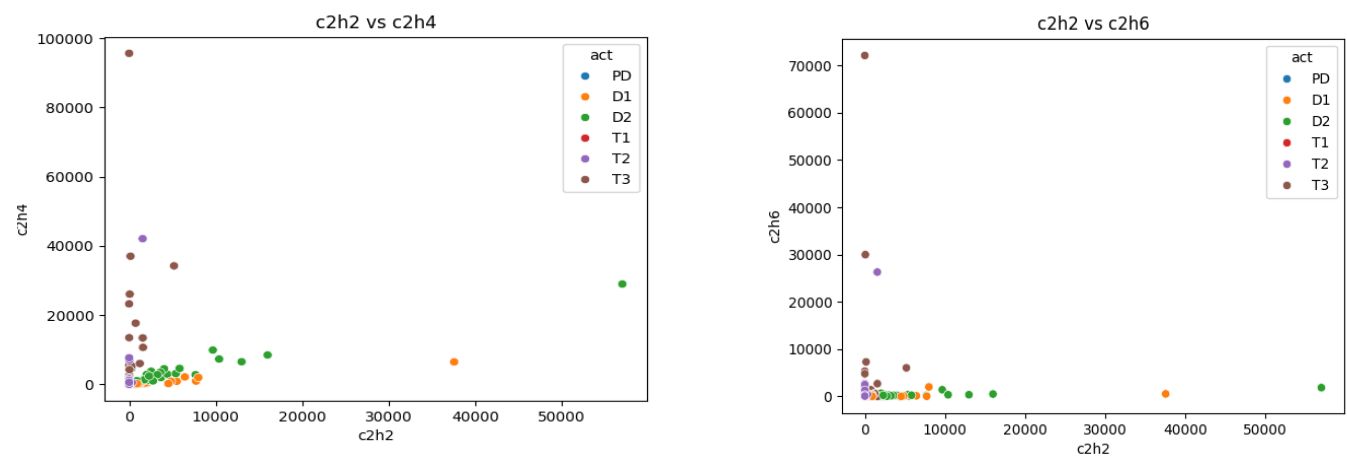
| Class | PD | D1 | D2 | T1 | T2 | T3 | Total |
|-------------------------|----|-----|-----|-----|----|-----|-------|
| Before Remove duplicate | 90 | 140 | 203 | 130 | 85 | 142 | 790 |
| After Remove duplicate | 86 | 140 | 203 | 130 | 85 | 137 | 781 |

B. Feature Transformation of Transformer Oil Gases

Figure 3 shows the pairwise interactions and distributions of transformer oil gases (H₂, CH₄, C₂H₆, C₂H₄, C₂H₂) across six fault kinds (represented by the act variable). Most gases have a highly skewed distribution, with the majority of data points near lower values and a few outliers at higher ones. This suggests that most samples had modest gas concentrations; however, some fault types (e.g., D2 and T3) contribute to higher amounts of specific gases. Pairwise graphs show weak or nonlinear connections between gas characteristics. Some combinations, like CH₄ and C₂H₆, show separate clusters corresponding to fault classes, implying that specific gas ratios can differentiate between fault kinds. However, overlap across

fault classes is visible in several plots, showing difficulties in recognizing faults based simply on these characteristics. This unpredictability emphasizes the necessity for improved machine learning models or data pretreatment procedures to improve fault classification.

Figure 3. Scattered plot of each gasses with each transformer fault.



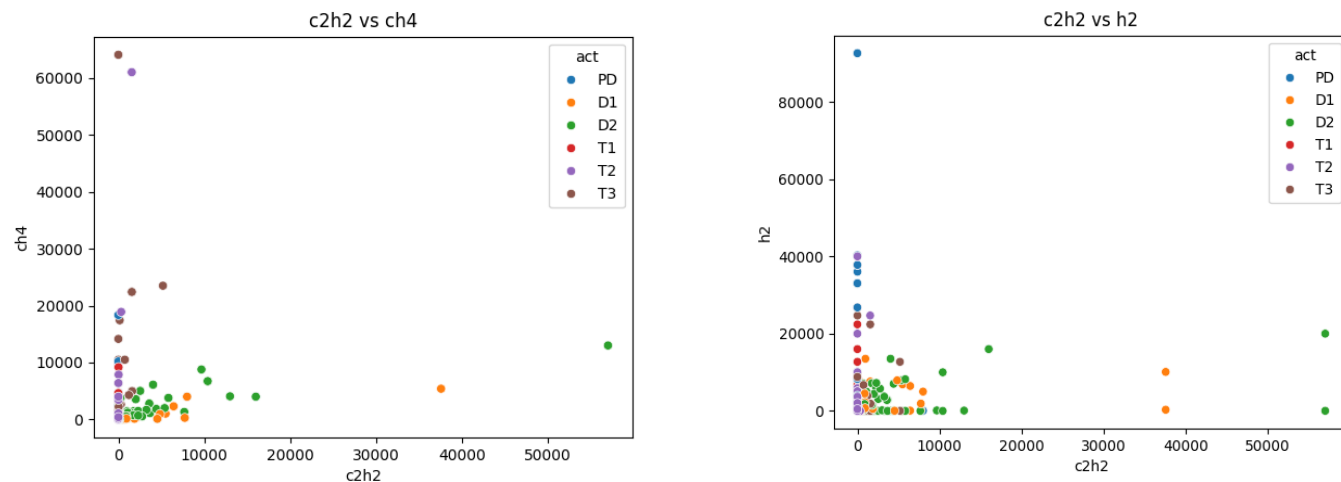
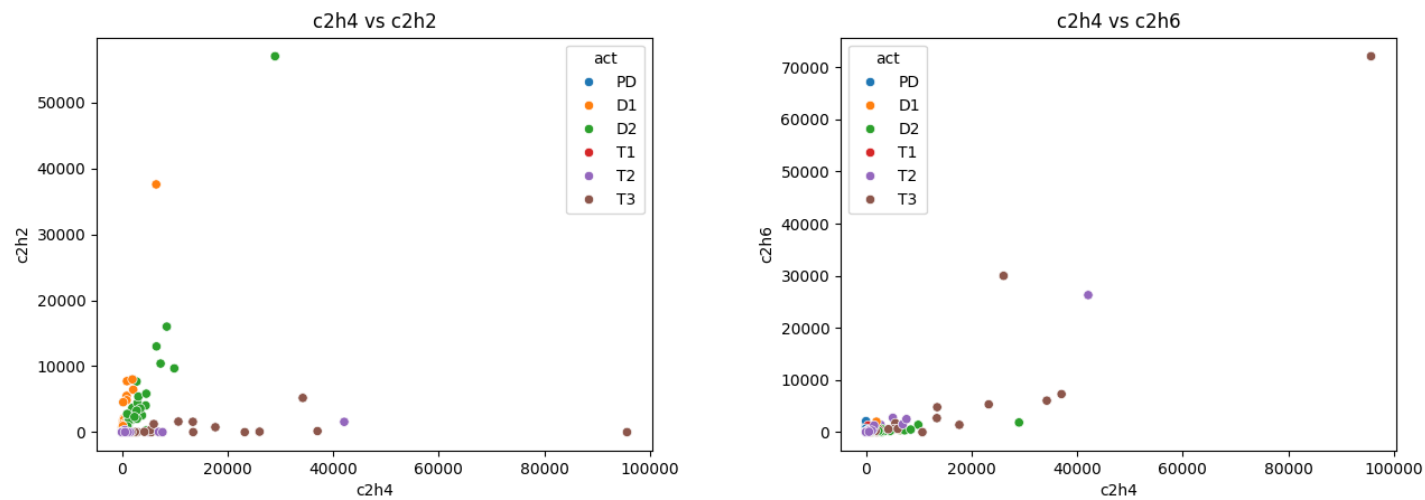


Figure 3. Scatter Plot of Gas C2H2 with other gases: (a) C2H4. (b) C2H6. (c) CH4. (d) H2



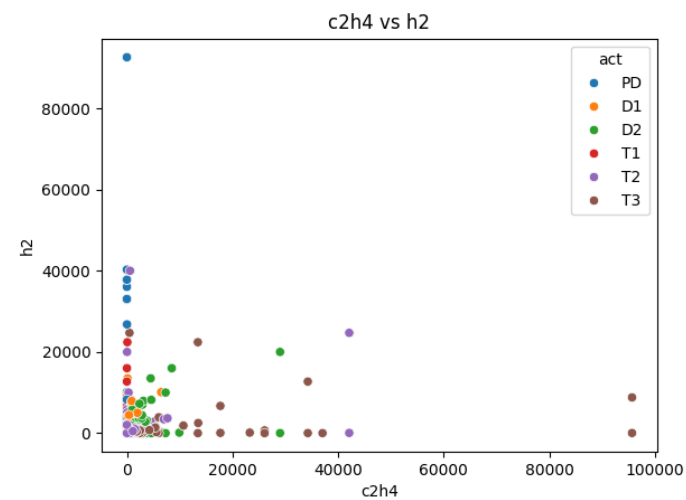
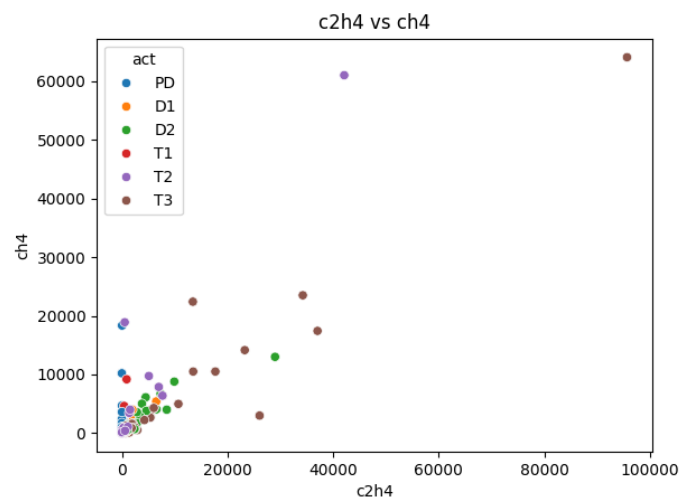
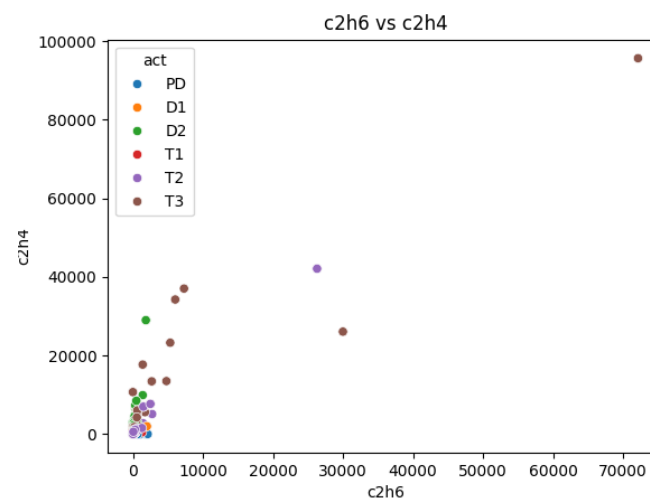
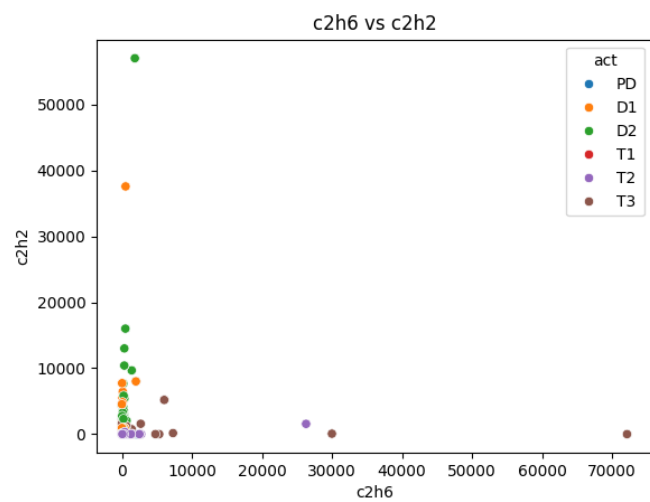


Figure 3. Scatter Plot of Gas C2H4 with other gasses: (a) C2H2. (b) C2H6. (c) CH4. (d) H2



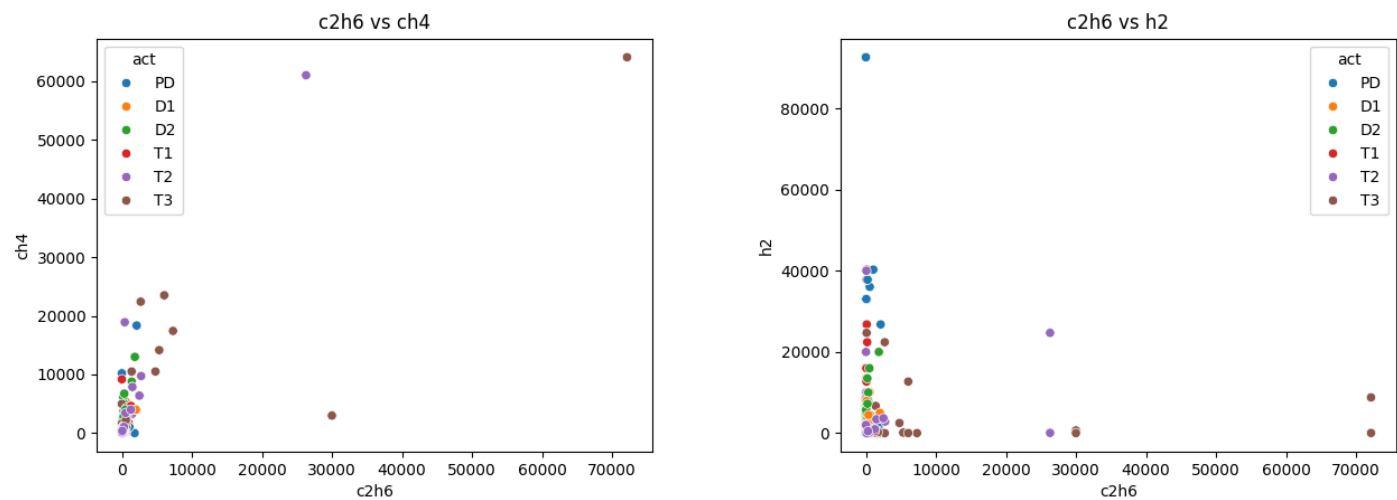
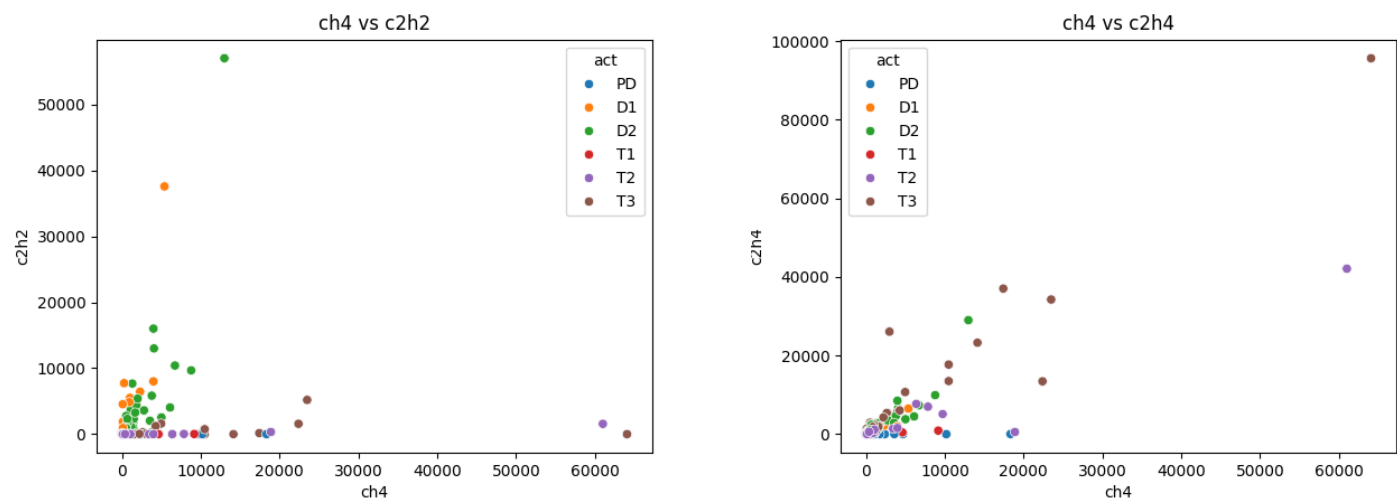


Figure 3. Scatter Plot of Gas C2H6 with other gasses: (a) C2H2. (b) C2H4. (c) CH4. (d) H2



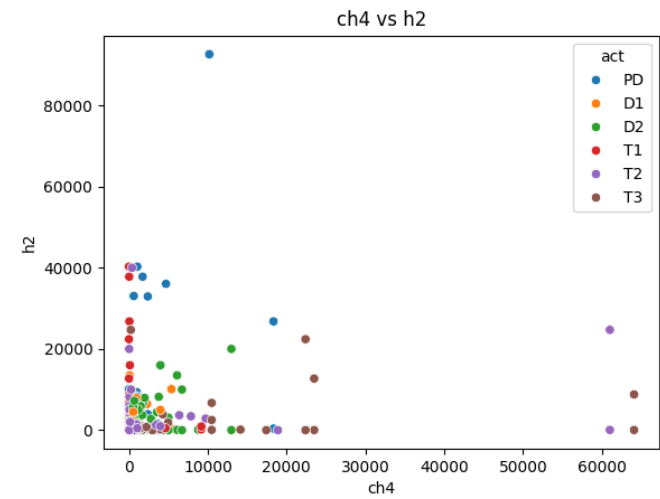
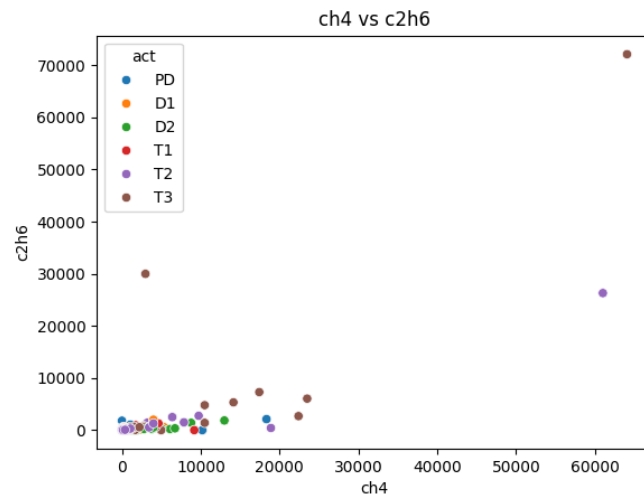
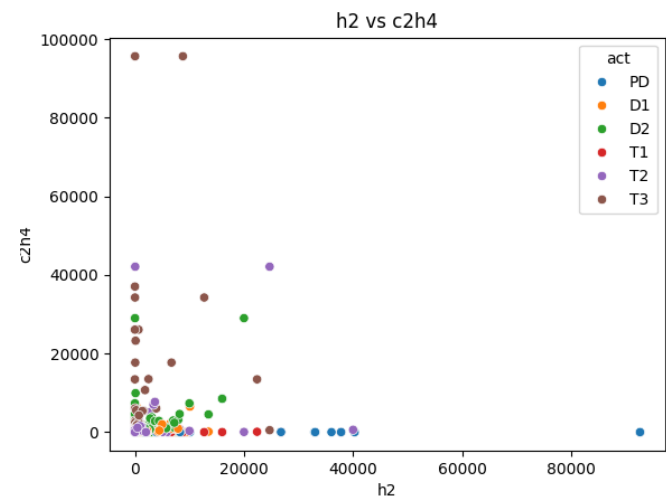
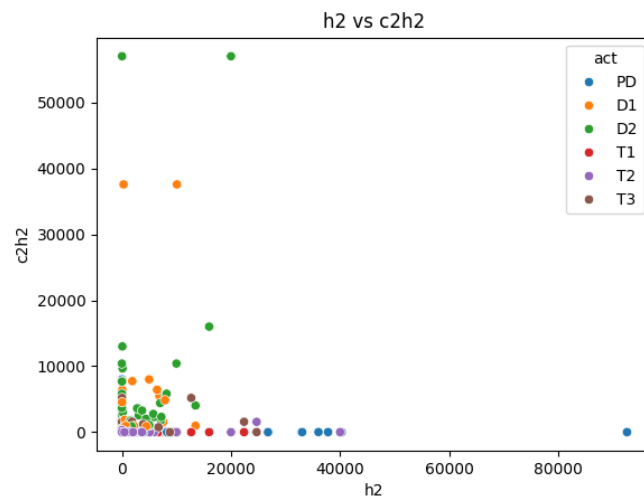


Figure 3. Scatter Plot of Gas CH₄ with other gasses: (a) C₂H₂. (b) C₂H₄. (c) C₂H₆. (d) H₂



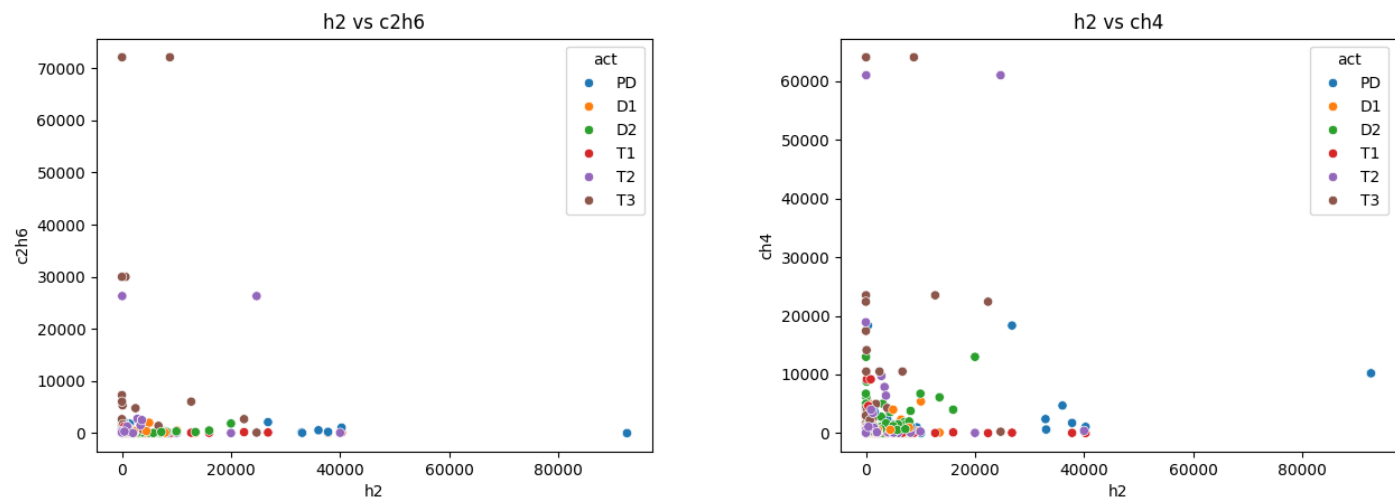
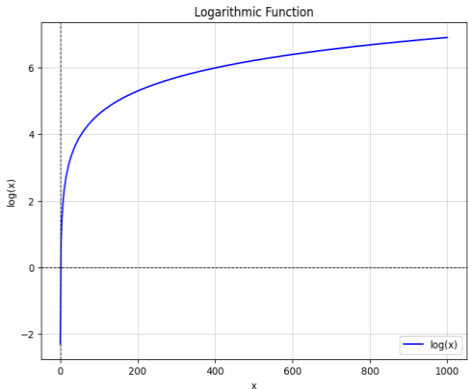
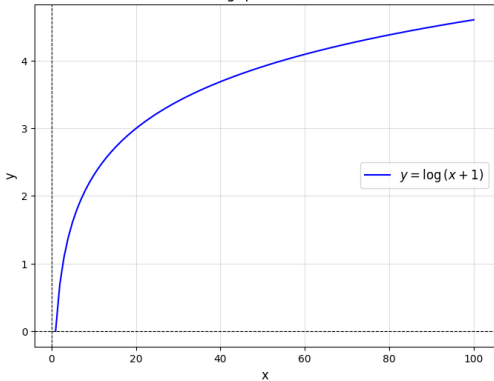
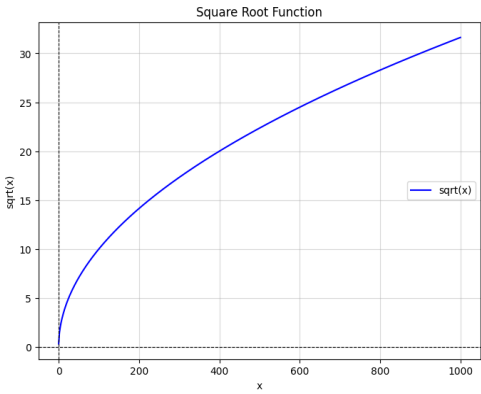
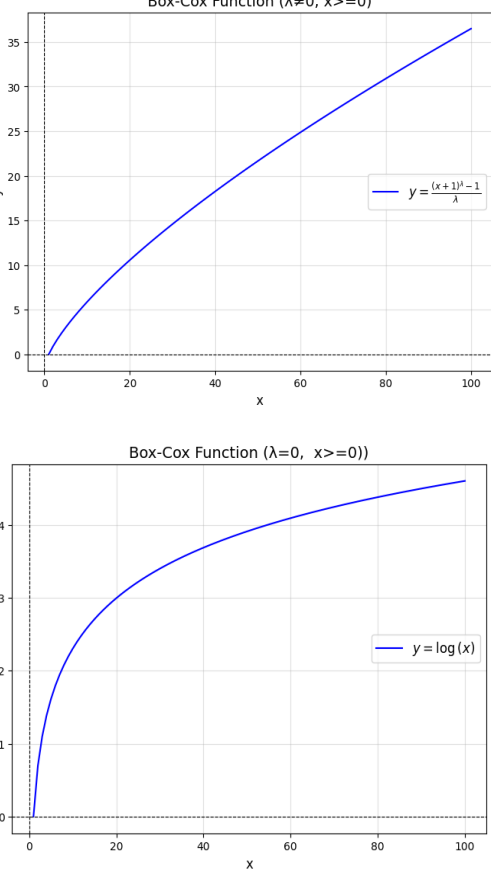


Figure 3. Scatter Plot of Gas H2 with other gasses: (a) C2H2. (b) C2H4. (c) C2H6. (d) CH4

The following lists the formulas (equations 1-7) in Table 3 for each of the five feature transformations that were taken into consideration. X represents the magnitude of each of the five major gases (H2, CH4, C2H6, C2H4 or C2H2) in each sample, while Y represents the transformation of the dissolved gas concentrations ratio. When performing feature transformations using the Log and Box-Cox methods, a small constant (2.2×10^{-16}) is added to the data before applying the transformation to handle zero values, as the logarithm of zero is undefined (Thin & Van Dua, 2024).

Table 3. Feature Transformation Formula with the Graph of a Function

| No. | Feature transformation | Graph of a Function |
|-----|--|--|
| 1. | <div>Log transformation</div> <div>$y = \log(x)$</div> <div>(1)</div> |  |
| 2. | <div>Log1p transformation</div> <div>$y = \log(x + 1)$</div> <div>(2)</div> |  |

| | | |
|----|---|---|
| 3. | <p><i>Square root transformation</i></p> $y = \sqrt{x} \quad (3)$ |  <p>The graph shows the square root function, $y = \sqrt{x}$, plotted against x. The x-axis ranges from 0 to 1000, and the y-axis ranges from 0 to 30. The curve starts at the origin (0,0) and increases at a decreasing rate, passing through points like (100, 10) and (400, 20). A legend indicates the blue line represents $\text{sqrt}(x)$.</p> |
| 4. | <p><i>Box-Cox transformation</i></p> $y = \begin{cases} \frac{(x+1)^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, x \geq 0, \\ \log(x), & \text{if } \lambda = 0, x \geq 0. \end{cases} \quad (4)$ <p>Where λ is the optimal parameter for input data, using maximum likelihood estimation. MLE equation for Box-Cox is as below:</p> $\text{LLF} = (\lambda - 1) \sum_i \log(x_i) - \frac{N}{2} \log \left(\frac{1}{N} \sum_i (y_i - \bar{y})^2 \right) \quad (5)$ |  <p>The top graph, titled "Box-Cox Function ($\lambda \neq 0, x \geq 0$)", shows the function $y = \frac{(x+1)^\lambda - 1}{\lambda}$ for $\lambda \neq 0$. The x-axis ranges from 0 to 100, and the y-axis ranges from 0 to 35. The curve starts at (0,0) and increases monotonically. A legend indicates the blue line represents $y = \frac{(x+1)^\lambda - 1}{\lambda}$.</p> <p>The bottom graph, titled "Box-Cox Function ($\lambda = 0, x \geq 0$)", shows the function $y = \log(x)$ for $\lambda = 0$. The x-axis ranges from 0 to 100, and the y-axis ranges from 0 to 4. The curve starts at (1,0) and increases monotonically. A legend indicates the blue line represents $y = \log(x)$.</p> |

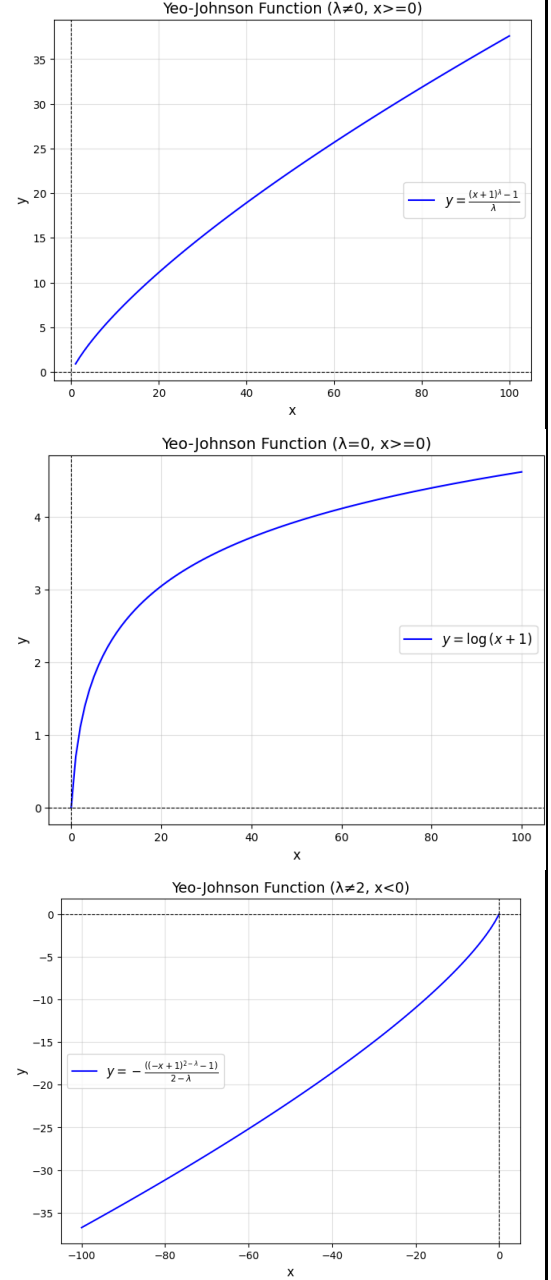
5.

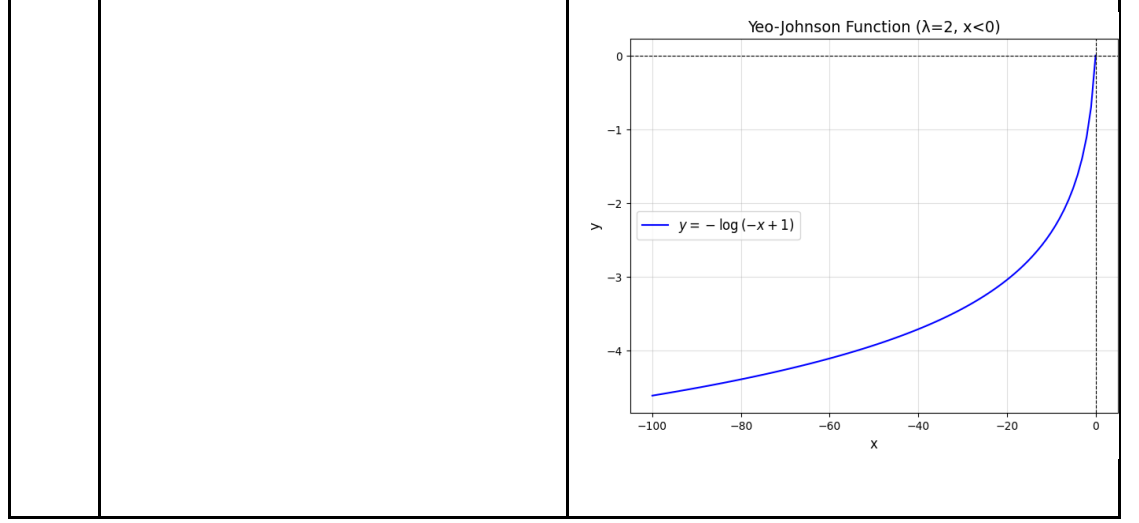
Yeo-johnson transformation

$$y = \begin{cases} \frac{(x+1)^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, x \geq 0, \\ \log(x+1), & \text{if } \lambda = 0, x \geq 0, \\ -\frac{((-x+1)^{2-\lambda} - 1)}{2-\lambda}, & \text{if } \lambda \neq 2, x < 0, \\ -\log(-x+1), & \text{if } \lambda = 2, x < 0. \end{cases} \quad (6)$$

Where λ is the optimal parameter for input data, using maximum likelihood estimation. MLE equation for Yeo-johnson is as below:

$$LLF = -\frac{N}{2} \log(\hat{\sigma}^2) + (\lambda - 1) \sum_i \text{sign}(x_i) \log(|x_i| + 1) \quad (7)$$





The asymmetry in the distribution of each gas feature is assessed using Karl Pearson’s Second Coefficient of Skewness, as defined in Equation 6, as the mode is not defined. This evaluation aims to compare feature transformations and identify those that result in values closest to zero.

$$S_k = \frac{3(\text{Mean} - \text{Mode})}{\sigma} \quad (8)$$

The coefficient equals zero for a perfectly symmetrical distribution. When the mean exceeds the mode, the coefficient of skewness is positive; otherwise, it is negative. For moderately skewed distributions, the value of Karl Pearson’s coefficient of skewness typically falls within ± 1 (Trivedi, 2017).

C. Resampling with SMOTE-ENN

Figure 4 shows the distribution of the act variable, which represents six classes of transformer faults. The distribution is imbalanced, as the counts differ significantly between classes. Class 3 (D2) has the most samples in the dataset (203), followed by Classes 6 (T3), 2 (D1), and 4 (T1), which have similar counts (137, 140, and 130). Classes 1 (PD) and 5 (T2) have the minority samples, with 86 and 89, respectively. This skewness suggests a difference in representation, with Class 3

(D2) overrepresented and Classes 1 (PD) and 5 (T2) underrepresented. Such an imbalanced distribution may induce bias in machine learning models, favouring the majority class and reducing predicted performance for minority classes.

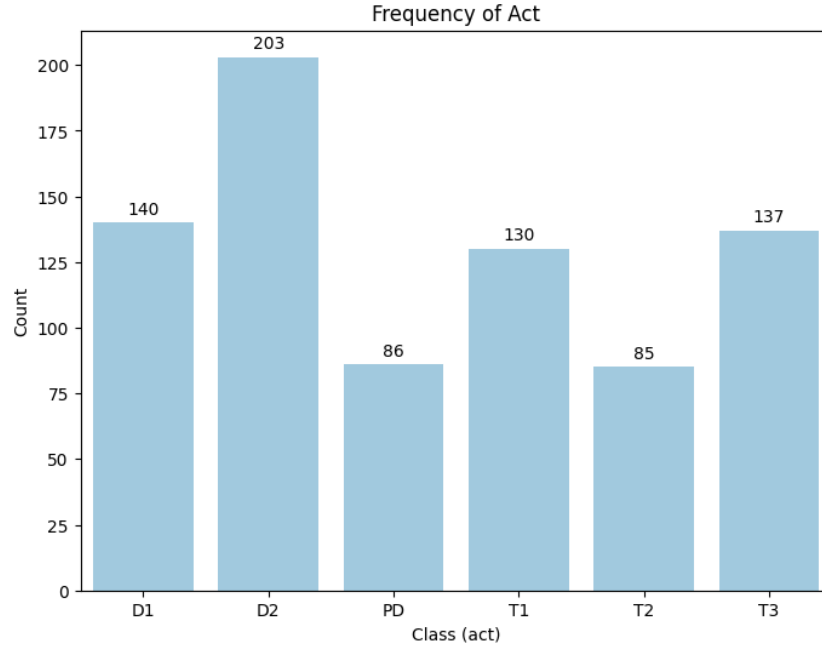


Figure 4. The Count of each DGA Classes

The dataset included in class 3 (D2) has the majority of instances of the transformer fault classes. This suggested that the dataset was skewed. The model was more efficient in classifying the majority class but did not reliably classify the minority class. SMOTE (Wu et al., 2022) was employed to resolve this issue. This oversampled the minority class to create an equal percentage of the six classes which eliminated the imbalance. To create new samples in an imbalanced dataset, a minority sample (xi) is selected as the base sample using Equation 7. To generate k -neighboring samples for each minority class, the distance measure Euclidean in Equation 9 is applied to calculate the distance between the base sample and all other samples within the corresponding minority class set.

Calculate the degree of imbalance between the majority and minority classes in the transformer oil chromatography fault data set. Then, use Equation 10 to randomly

select n samples from k -neighboring samples of x_i as auxiliary samples (y_1, y_2, \dots, y_n). In Equation 11, random interpolation is performed between the root sample x_i of the transformer minority fault type and a randomly selected auxiliary sample y_i . Then, n corresponding minority transformer fault samples p_i are synthesised to achieve data balance (Wu et al., 2022).

$$C_{\text{minority}} = \arg \min_{i \in \{1, 2, \dots, k\}} |C_i| \quad (9)$$

$$C_{\text{majority}} = \arg \max_{i \in \{1, 2, \dots, k\}} |C_i| \quad (10)$$

$$d(x, y) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2} \quad (11)$$

$$n = \text{round}(IL) \quad (12)$$

round means rounding the calculated IL.

$$p_i = x_i + \text{rand}(0, 1) \star (y_i - x_i), i = 1, 2, 3, \dots, n \quad (13)$$

$\text{rand}(0, 1)$ represents a random number in the production interval (0,1).

The under-sampling technique enhances minority sample classification performance by minimizing the number of majority samples. The closest neighbour rule identifies the nearest neighbour samples of each majority sample based on the distance between two samples. Consistent labelling is used to assess whether the majority of samples are noise.

$$KNN(X_i, k) = \{y \in X | \text{dist}(X_j, X_i) \leq \text{dist}(X_{i'}', X_i)\} \quad (14)$$

The k -nearest neighbour of the sample X_i is the sum of samples from the dataset X whose distance X is smaller than the distance between X_i 's k -th nearest neighbour sample and X . It is expressed as in equation 12. Where X_i' is the k -th nearest neighbour sample of X_i in the dataset X , and dist is the distance between sample X_i and its nearest neighbour, which is often Euclidean (Xu et al., 2020).

The principle of ENN is to delete samples whose class differs from the majority class of their k -nearest neighbours (Xu et al., 2020). The algorithm's primary goal is to remove the vast majority of noise samples. The steps for SMOTE-ENN are as follows:

| Proposed Technique: Yeo Johnson-SMOTE-ENN | |
|--|---|
| Input: | x (features), y (target), k (number of neighbours for SMOTE and ENN) |
| Output: | x_smote_enn (resampled and cleaned features), y_smote_enn (resampled and cleaned target), λ |
| Step 1 | Apply Yeo-Johnson transformation |
| | Find the most optimal λ using maximum likelihood estimation (MLE) $LLF = -\frac{N}{2} \log(\hat{\sigma}^2) + (\lambda - 1) \sum_i \text{sign}(x_i) \log(x_i + 1)$ |
| | Apply x to the with defined λ $y = \begin{cases} \frac{(x+1)^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, x \geq 0, \\ \log(x + 1), & \text{if } \lambda = 0, x \geq 0, \\ -\frac{((-x+1)^{2-\lambda} - 1)}{2-\lambda}, & \text{if } \lambda \neq 2, x < 0, \\ -\log(-x + 1), & \text{if } \lambda = 2, x < 0. \end{cases}$ |
| Step 2 | Perform SMOTE |
| | Identify the minority classes in y $C_{\text{minority}} = \arg \min_{i \in \{1, 2, \dots, k\}} C_i $ |
| | For each minority class: |

| | |
|--------|---|
| | Repeat the following until the number of samples in the minority class equals the number in the majority class: |
| | For each minority sample, find the x k-nearest neighbours $KNN(X_i, k) = \{y \in X dist(X_j, X_i) \leq dist(X_i', X_i)\}$ |
| | Randomly select one of the k-neighbors |
| | Generate a synthetic sample along the line between the sample and its selected neighbour: |
| | Add a random number between 0 and 1 to the synthetic sample $p_i = x_i + rand(0,1) * (y_i - x_i), i = 1, 2, 3, \dots, n$ |
| | Append the synthetic sample to x and assign the corresponding label to y |
| Step 3 | Combine synthetic samples with the original dataset |
| | Update x and y with the newly generated synthetic samples and their labels |
| Step 4 | Apply ENN (Edited Nearest Neighbors) to the updated dataset |
| | Identify the majority and minority classes in y $C_{majority} = \arg \max_{i \in \{1, 2, \dots, k\}} C_i $ $C_{minority} = \arg \min_{i \in \{1, 2, \dots, k\}} C_i $ |
| | For each sample in the dataset: |
| | Find the k-nearest neighbours $KNN(X_i, k) = \{y \in X dist(X_j, X_i) \leq dist(X_i', X_i)\}$ |
| | Check the class consistency of the sample: |

| | |
|--------|---|
| | Compare the class of the sample with the classes of its k neighbours |
| | If the majority of the k neighbours belong to a different class, mark the sample as inconsistent or noisy |
| | Remove inconsistent or noisy samples from x and y |
| Step 5 | Return the cleaned dataset |
| | $x_{\text{smote_enn}}$: Features with noisy points removed |
| | $y_{\text{smote_enn}}$: Target labels with noisy points removed |

The proposed technique, YeoJohnson-SMOTE-ENN, integrates feature transformation, synthetic oversampling, and noise removal for enhanced dataset balance and quality. In Step 1, the Yeo-Johnson transformation is applied to the features x . This involves determining the most optimal transformation parameter λ using maximum likelihood estimation (MLE). The identified λ is used to transform x , making the data more Gaussian-like and suitable for subsequent modeling. In Step 2, SMOTE is employed. Minority classes in the target variable y are identified, and synthetic samples are iteratively generated to balance class distributions. For each minority sample, k -nearest neighbors are computed, a neighbor is randomly selected, and a synthetic sample is created along the line between the sample and its selected neighbor, with a random weight applied to add variability. These synthetic samples, along with their corresponding labels, are then combined with the original dataset in Step 3, creating a balanced dataset for further processing.

In Step 4, ENN is used to clean the dataset by removing noisy or inconsistent samples. This involves identifying k -nearest neighbors for each sample in the dataset and assessing class consistency. If the majority of neighbors belong to a class different from the sample's class, the sample is marked as noisy or inconsistent and removed. This step ensures the removal of mislabeled or overlapping samples, enhancing the quality of the dataset. Finally, in Step 5, the cleaned and balanced

dataset is returned as x_{smote_enn} (resampled and cleaned features) and y_{smote_enn} (resampled and cleaned target labels), providing a robust dataset for subsequent modeling.

D. Train-Test Split

The dataset was divided into training and testing subsets at an 80:20 ratio to ensure a thorough evaluation of the model's performance. Stratified sampling was used to preserve the target variable's original distribution across the train and test sets. This method retains the proportion of each class in the target variable, ensuring that both subsets reflect the whole dataset. Stratification is especially crucial for imbalanced datasets because it eliminates over-representation or under-representation of specific classes, which could result in biased or misleading model evaluations. This split serves as the foundation for training the model on various data and validating its ability to generalize well.

E. XGBoost Classifier

XGBoost uses a weighted quantile sketch and a sparsity-aware algorithm. Sparsity refers to zero or missing values, whereas a weighted quantile sketch employs approximation tree learning for merging and pruning (Salekshahrezaee et al., 2022). XGBoost employs gradient boosting to improve performance while minimizing overfitting. XGBoost leverages level 1 baseline models to handle categorical variables, reduce overfitting, manage non-linear decision boundaries, and solve imbalanced classes (Aslam et al., 2022). The parameter definitions for XGBoost are set in Table 4.

Table 4. Hyperparameter used in for the XGBoost classifier

| Hyperparameter | Value |
|----------------|-------|
| n_estimators | 100 |
| learning_rate | 0.1 |
| max_depth | 3 |

F. Performance Metrics

This study evaluates accuracy, weighted average precision (WAP), weighted average recall (WAR), weighted average F1-score (WAF), and model processing time. These measures were used to evaluate the performance of the created models. Weighted averages for recall, accuracy, and F1-score were computed using a particular mathematical formula.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (15)$$

$$Precision = \frac{TP}{TP+FP} \quad (16)$$

$$Recall = \frac{TP}{TP+FN} \quad (17)$$

The ratio of accurately predicted occurrences to total instances in a dataset is known as accuracy (Equation 15). Equation 16 is the precision calculated by adding the number of true positives and false positives. Lastly, recall as in Equation 17 calculated by adding the number of true positives and false negatives.

4. Experimental results

The measurement of skewness for each gas feature in transformer data was conducted shown in Table 5 to assess the distributional symmetry of the features and evaluate the impact of various transformation techniques. Skewness quantifies the asymmetry of a distribution, and values close to zero indicate a symmetrical distribution. Using the formula from Equation 6, the skewness of the raw features (H2, CH4, C2H6, C2H4, and C2H2) was calculated before and after applying five transformations: Log, Log1p, Square Root,

Box-Cox, and Yeo-Johnson. Each transformation was evaluated to determine its ability to reduce skewness and bring the values closer to zero.

The Yeo-Johnson transformation is the best transformer for handling skewness, especially because it can be applied to both positive and zero values, making it more versatile than Box-Cox. The λ parameter plays a similar role in both transformations by determining the degree of transformation, but Yeo-Johnson's ability to handle zero and negative values makes it more adaptable.

For example, after applying Yeo-Johnson, the skewness of H2 is reduced to 0.2443, which is the closest to zero among all transformations for this feature, indicating a nearly symmetrical distribution. Similarly, it performs well for CH4, reducing skewness to 0.0017, which is significantly closer to symmetry compared to other transformations. This flexibility and effectiveness in normalizing features with varying distributions make Yeo-Johnson a robust choice for preprocessing, particularly when working with datasets containing a mix of positive and zero values.

The Yeo-Johnson transformation pulls the data closer to normality and reduces the variance by applying different functions based on the data point value and the λ parameter (Equation 5). Outliers can negatively influence distance calculations because of bias introduced by their disproportionately large distance between points. The transformation effectively mitigates the impact of outliers by compressing extreme values and reducing their scalar. By reducing the range of extreme values, the Yeo-Johnson transformation ensures that no single point dominates the computation of distances. This normalization process enhances the robustness and reliability of algorithms that rely on k -nearest neighbors distance measure.

Table 5. Measurement of Skewness for Each Transformer Gasses Features

| Features | Transformation | | | | | |
|----------|----------------|-----|-------|-------------|---------|-------------|
| | Before | Log | Loglp | Square Root | Box-cox | Yeo-Johnson |

| | | | | | | |
|------|--------|---------|---------|--------|---------|----------------|
| H2 | 0.7213 | -0.0917 | 0.3544 | 1.0836 | 0.4784 | 0.2443 |
| CH4 | 0.6297 | 0.0054 | 0.2425 | 0.9738 | 0.335 | 0.0017 |
| C2H6 | 0.3363 | -0.5797 | -0.0821 | 0.6364 | -0.0751 | -0.2066 |
| C2H4 | 0.5769 | -0.3768 | 0.1381 | 0.9794 | 0.2037 | -0.0363 |
| C2H2 | 0.5533 | -1.1614 | 0.4479 | 1.0717 | -0.7095 | -0.0563 |

Figure 5 shows different feature transformations that affect the feature distributions of the power transformer gases. Initially, the raw data in Figure 5(a) exhibits high skewness and lengthy tails, indicating non-Gaussian distributions, which can harm machine learning models. The log transformation in Figure 5(b) compresses the scale of high values, reducing skewness while preserving some residual asymmetry. Similarly, the log1p transformation Figure 5(c), which handles zero values, normalizes the distributions beyond the raw and log-transformed data. The square root modification Figure 5(d) considerably improves skewness while keeping longer tails. In contrast, the Box-Cox treatment Figure 5(e) is extremely effective at normalizing data, resulting in smoother and more symmetric distributions when the data is strictly positive. Finally, the Yeo-Johnson transformation Figure 5(f) works similarly to Box-Cox while allowing for both positive and negative values, resulting in well-normalized distributions. Overall, the transformations effectively reduce skewness and improve symmetry, with Box-Cox and Yeo-Johnson being the most robust strategies for preparing the data for machine learning and preprocessed steps that incorporate distance measure.

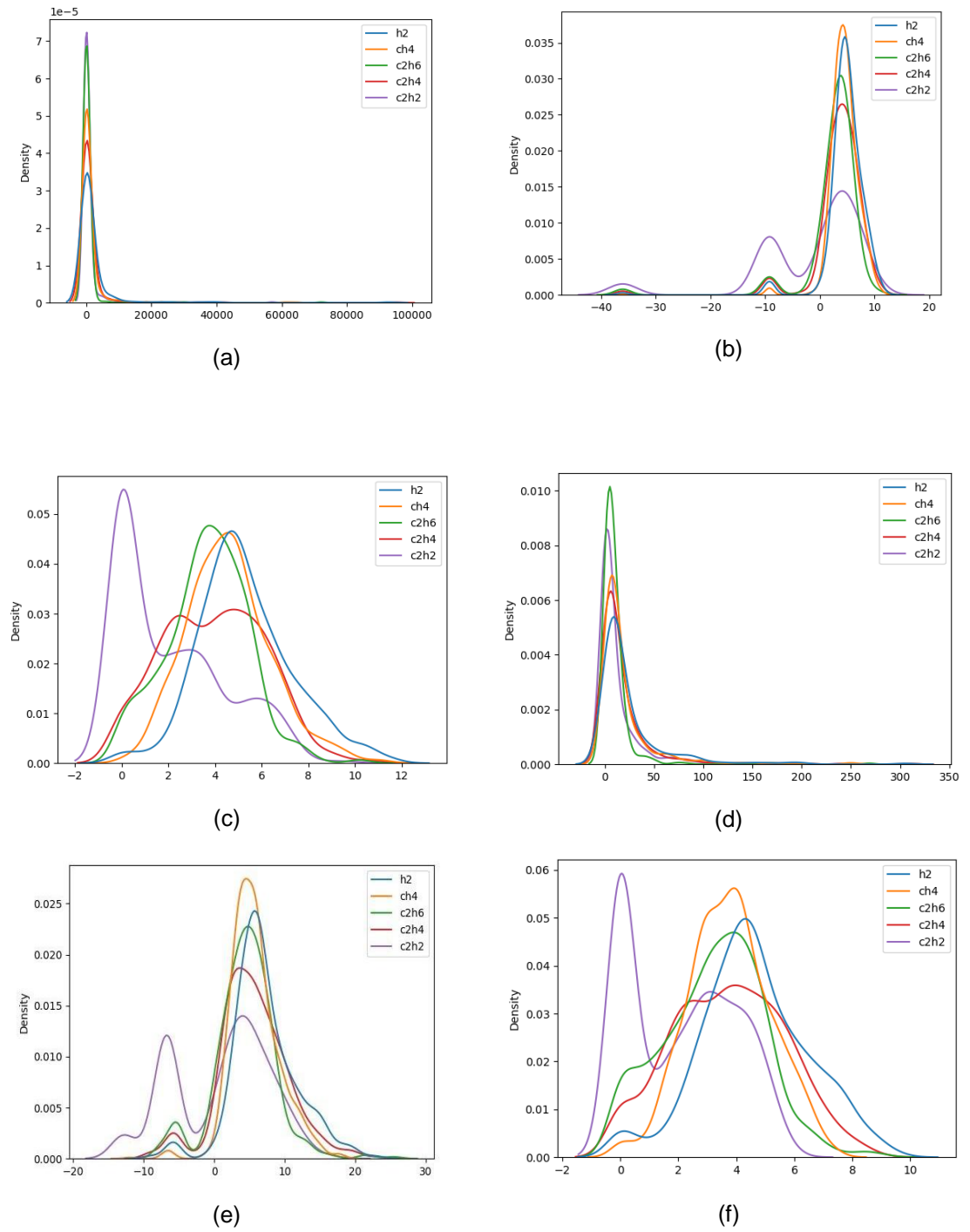


Figure 5. Kernel Density Estimate Plot: (a) Before Implementing Feature Transformation. (b) Log Transformation. (c) Log1p Transformation. (d) Square Root Transformation. (e) Box-cox Transformation. (f) Yeo-Johnson Transformation

Figure 6 demonstrates the impact of applying the Yeo-Johnson transformation on the features of a DGA dataset. On the left, the original feature distributions exhibit significant skewness and non-Gaussian behavior, with values heavily concentrated near zero and scattered unevenly across the range. Such skewed distributions can negatively affect performance that rely on distance measures as many algorithms assume features to follow a Gaussian-like distribution. On the right, after applying the Yeo-Johnson transformation, the bar plot shows that the features have become more symmetrical and normalized, reducing skewness. Additionally, the scatter plots show improved linear relationships between the features, with better class separability as indicated by the distribution of colored points. The transformation ensures the data is better structured and balanced, which enhances its suitability to calculate distance measure without bias.

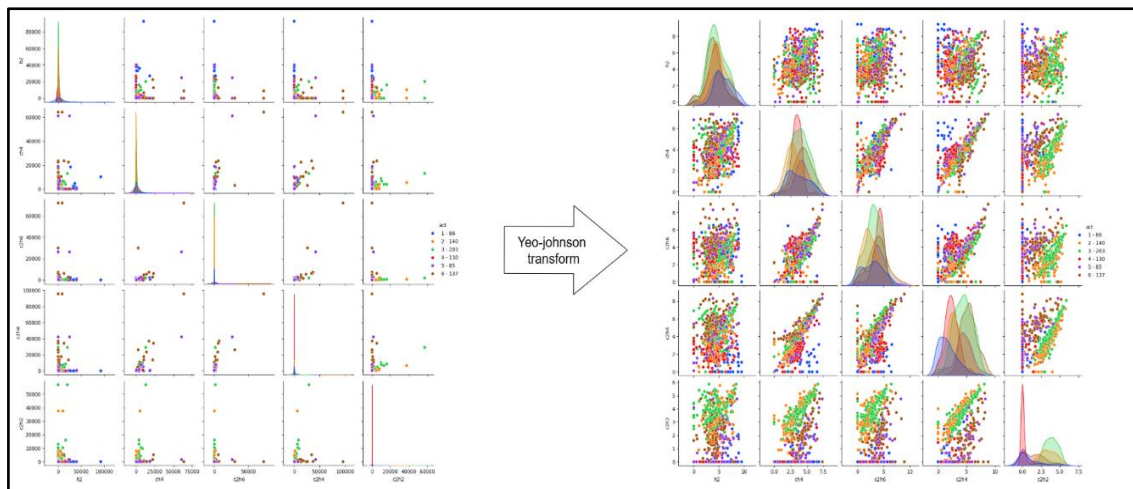


Figure 6. Pairplot Before Transform and After Transform using Yeo-Johnson Strategy

Figure 7 shows the class distributions of the DGA dataset before and after applying resampling techniques, emphasizing the impact of ENN, SMOTE, and the combined SMOTE-ENN techniques. Initially, the dataset is highly skewed, with significant class imbalances where D1 has the highest representation (203 samples) and D2 and T2 have the lowest (86 and 85 samples, respectively). ENN reduces noise by removing misclassified samples but worsens the class imbalance, particularly for smaller classes like PD (16 samples) and D2 (19 samples), increasing skewness. In contrast, SMOTE effectively balances the dataset by oversampling minority classes to ensure all classes reach an equal

count of 203 samples, creating a perfectly balanced distribution. However, the combined SMOTE-ENN technique provides the most practical balance by generating synthetic samples first and removing out-of-boundary data using ENN. This results in a cleaner, less noisy dataset with improved class distributions, where most classes range between 135 and 203 samples. SMOTE-ENN reduces skewness, improves data quality, and avoids overfitting, making it particularly useful for enhancing the performance of machine learning models on imbalanced datasets.

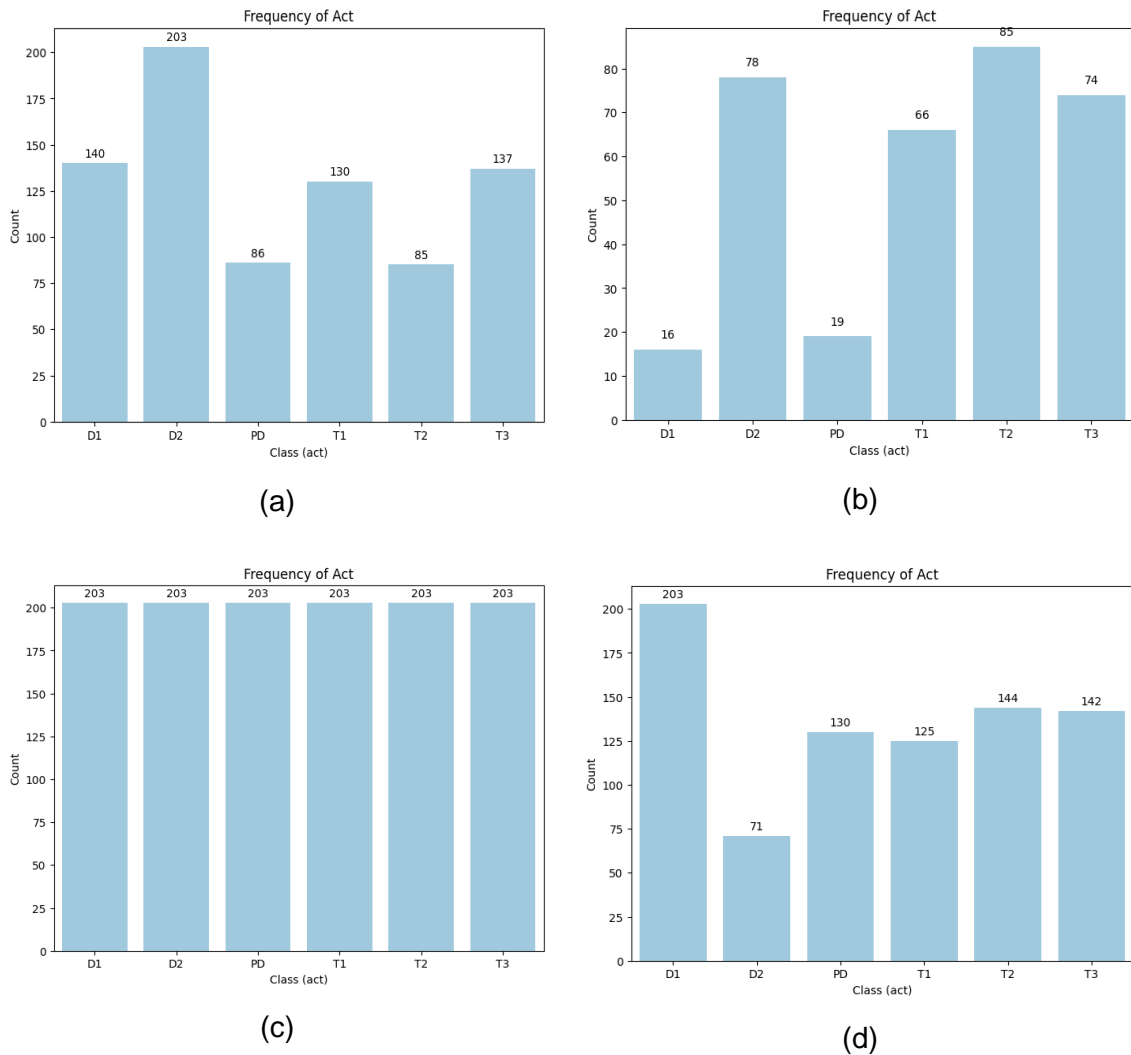


Figure 7. Distribution of Target Classes: (a) Before Implementing Resampling Technique. (b) ENN Technique. (c) SMOTE Technique. (d) SMOTE-ENN Technique.

Table 5 evaluates the accuracy of various resampling techniques combined with feature transformations. The resampling techniques examined are ENN, SMOTE, and SMOTE-ENN, with accuracy measured before and after applying transformations such as Log, Log1p, Square Root, Box-cox, and Yeo-Johnson. All of these techniques are being evaluated using an XGBoost machine learning classifier with 80:20 ratios of training and testing subsets.

ENN-XGBoost shows significant improvement with feature transformations, increasing from 78.00% to a maximum of 92.60% using Yeo-Johnson. This method demonstrates the highest accuracy and consistency across all models due to its deterministic nature, with no reliance on randomized factors. SMOTE-XGBoost, on the other hand, displays less consistent gains, with some transformations slightly reducing accuracy. SMOTE-ENN-XGBoost achieves the highest baseline accuracy (89.90%) and further improves to 93.20% with Yeo-Johnson or Log1p transformations, demonstrating its robustness and effectiveness. Overall, the results highlight the importance of selecting appropriate feature transformations alongside resampling techniques, with SMOTE-ENN-XGBoost combined with Yeo-Johnson being the most effective approach for improving classification performance on imbalanced datasets.

Table 5. Performance Evaluation of Each Resampling Technique and Feature Transformation

| Resampling Techniques | Accuracy (%) | | | | | |
|--------------------------|--------------|------------------------|-------|-------------|--------|--------------------|
| | Before | Feature Transformation | | | | |
| XGBoost | 71.30 | Log | Log1p | Square Root | Boxcox | Yeo-Johnson |
| ENN-XGBoost | 78.00 | 88.20 | 89.70 | 82.40 | 85.90 | 92.6 |
| SMOTE-XGBoost | 79.50 | 75.40 | 80.70 | 76.20 | 78.30 | 77.00 |
| SMOTE-ENN-XGBoost | 89.90 | 87.70 | 93.00 | 85.20 | 83.00 | 93.20 |

Table 6 provides precision and recall for power transformer classes PD (1), D1 (2), D2 (3), T1 (4), T2 (5), and T3 (6) using different resampling techniques and feature transformations combined with the XGBoost model. Among the resampling techniques, ENN-XGBoost, SMOTE-XGBoost, and SMOTE-ENN-XGBoost were evaluated across transformations including Log, Log1p, Square Root, Box-Cox, and Yeo-Johnson. ENN-XGBoost with Yeo-Johnson transformation achieves perfect precision and recall (100%) for PD (1), D1 (2), and D2 (3) classes, and strong performance for other classes. Similarly, SMOTE-ENN-XGBoost with Log1p transformation provides perfect precision and recall for T1 (4), T2 (5), and T3 (6), alongside consistently high scores across other classes. Notably, SMOTE-XGBoost shows moderate but less consistent performance, with the highest precision and recall for T2 (5) achieved with Log1p transformation. Overall, ENN-XGBoost with Yeo-Johnson is the best combination for precise classification of early faults (PD, D1, D2), while SMOTE-ENN-XGBoost with Log1p excels in diagnosing severe faults (T1, T2, T3). Upon reviewing the results, SMOTE-ENN-XGBoost with Yeo-Johnson transformation emerges as the best overall model. It achieves consistently high precision and recall across all power transformer classes, including high precision and recall for minority classes with 90.70 and 95.10 for PD (1) and 100, 100 for T2 (5), making it the most effective choice for accurate fault diagnosis for imbalanced data.

Table 6. Precision and Recall of Each Class, Resampling Technique, and Feature Transformation

| Resampling Technique | Feature Transformation | PD (1) | | D1 (2) | | D2 (3) | | T1 (4) | | T2 (5) | | T3 (6) | |
|----------------------|------------------------|---------------|------------|---------------|------------|---------------|------------|---------------|------------|---------------|------------|---------------|------------|
| | | Precision (%) | Recall (%) | Precision (%) | Recall (%) | Precision (%) | Recall (%) | Precision (%) | Recall (%) | Precision (%) | Recall (%) | Precision (%) | Recall (%) |
| XGBoost | Before | 52.90 | 52.90 | 60.00 | 64.30 | 81.60 | 75.60 | 74.10 | 76.90 | 60.00 | 70.60 | 88.00 | 78.60 |
| ENN-XGBoost | Log | 100 | 66.70 | 60.00 | 100 | 94.10 | 88.90 | 86.70 | 92.90 | 87.50 | 82.40 | 92.30 | 92.30 |
| | Log1p | 100 | 100 | 75 | 100 | 94.10 | 94.10 | 85.70 | 92.30 | 82.40 | 82.40 | 100 | 86.70 |
| | Square Root | 75.00 | 75.00 | 0 | 0 | 75.00 | 93.80 | 80.00 | 72.70 | 86.70 | 76.50 | 89.50 | 100 |
| | Box-cox | 100 | 100 | 60.00 | 75.00 | 93.80 | 88.20 | 86.70 | 92.90 | 77.80 | 82.40 | 92.90 | 81.20 |
| | Yeo-Johnson | 100 | 100 | 100 | 100 | 100 | 100 | 85.70 | 92.30 | 87.50 | 82.40 | 93.30 | 93.30 |
| SMOTE-XGBoost | Log | 80.80 | 52.50 | 68.80 | 82.50 | 74.40 | 70.70 | 73.80 | 75.60 | 77.80 | 85.40 | 79.50 | 85.40 |
| | Log1p | 84.80 | 70.00 | 67.30 | 87.50 | 78.40 | 70.70 | 90.60 | 70.70 | 80.90 | 92.70 | 88.40 | 92.70 |
| | Square Root | 71.80 | 70.00 | 65.10 | 70.00 | 73.20 | 73.20 | 90.90 | 73.20 | 81.00 | 82.90 | 78.30 | 87.80 |
| | Box-cox | 82.40 | 70.00 | 65.90 | 72.50 | 73.20 | 73.20 | 82.90 | 70.70 | 82.60 | 92.70 | 84.10 | 90.20 |
| | Yeo-Johnson | 82.90 | 72.50 | 63.60 | 70.00 | 67.40 | 70.70 | 90.30 | 68.30 | 78.00 | 95.10 | 85.40 | 85.40 |
| SMOTE-ENN- | Log | 83.30 | 97.60 | 66.70 | 76.90 | 91.70 | 68.80 | 95.80 | 88.50 | 90.00 | 78.30 | 96.30 | 96.30 |

| | | | | | | | | | | | | | |
|----------------|--------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|---------------|------------|--------------|------------|
| XGBoost | Log1p | 90.50 | 92.70 | 88.20 | 88.20 | 78.90 | 93.80 | 100 | 89.30 | 100 | 96.30 | 96.40 | 96.40 |
| | Square Root | 85.70 | 87.80 | 71.40 | 76.90 | 87.50 | 93.30 | 100 | 91.70 | 80.00 | 76.90 | 83.30 | 83.30 |
| | Box-cox | 87.20 | 82.90 | 66.70 | 66.70 | 64.70 | 68.80 | 95.80 | 85.20 | 80.00 | 92.30 | 89.30 | 89.30 |
| | Yeo-Johnson | 90.70 | 95.10 | 90.00 | 75.00 | 75.00 | 69.20 | 96.20 | 96.20 | 100.00 | 100 | 96.70 | 100 |

Table 7 presents a comparative analysis of the proposed Yeo-Johnson-SMOTE-ENN method with other methods applied to DGA datasets, highlighting their respective model performances. The referenced studies used various methods to address the issue of imbalanced data, achieving different levels of classification accuracy. Vuttipittayamongkol et al. employed Random Forest with SMOTE and attained a performance of 80.66% (Vuttipittayamongkol et al., 2021), while Tra et al. utilized a Multi-Layer Perceptron combined with SMOTE, achieving 85.10% (Tra et al., 2019). Chanchotisatien and Vong applied the LightGBM model, obtaining a performance of 87.63% (Chanchotisatien & Vong, 2021), and Taha and Mansour implemented an ensemble method that yielded a higher accuracy of 90.61% (Taha & Mansour, 2021).

In comparison, the proposed Yeo-Johnson-SMOTE-ENN technique significantly outperformed these methods, achieving the highest accuracy of 93.20%. This demonstrates the effectiveness of the proposed technique in addressing class imbalanced and improving classification performance, likely due to its integrated method of feature transformation, synthetic oversampling, and noise removal. This comparison underscores the superiority of the proposed method for imbalanced DGA datasets in transformer fault classification tasks.

Table 7. Comparison Results Between The Proposed Method and Other Methods

| Reference | Model | Model Performance (%) |
|------------------------------------|--------------------------------|-----------------------|
| (Vuttipittayamongkol et al., 2021) | RF+SMOTE | 80.66 |
| (Tra et al., 2019) | MLP+SMOTE | 85.10 |
| (Chanchotisatien & Vong, 2021) | LightGBM | 87.63 |
| (Taha & Mansour, 2021) | Ensemble method | 90.61 |
| Proposed Method | Yeo-Johnson-SMOTE-ENN- XGBoost | 93.20 |

5. Discussion

The results demonstrate the combined impact of feature transformation strategy and data-level resampling techniques (ENN, SMOTE, and SMOTE-ENN) on the accuracy of the

XGBoost model for the DGA dataset. Before any resampling or feature transformation strategy, XGBoost achieved a baseline accuracy of 71.30%, highlighting the limitations of working with skewed and imbalanced raw data.

When implementing feature transformation strategies such as Log, Log1p, Square Root, Box-cox, and Yeo-Johnson, the performance of XGBoost improved significantly, particularly when paired with data-level resampling techniques. Particularly, the Yeo-Johnson transformation consistently outperformed other transformations, achieving the highest accuracy across all data-level resampling techniques. This indicates its effectiveness in stabilizing the data distribution and reducing skewness, which enhances the data distribution to resample.

Among the data-level resampling techniques, ENN-XGBoost improved the baseline accuracy by removing out-of-boundary samples, achieving accuracies ranging from 78.00% (Log) to 92.60% (Yeo-Johnson). This result shows that ENN is effective in cleaning the data particularly when paired with Yeo-Johnson, where the accuracy reached its peak.

SMOTE-XGBoost, which focuses on resampling the dataset through synthetic samples, showed moderate improvements. While it improved the baseline accuracy, its performance ranged between 75.40% (Log1p) and 80.70% (Square Root), demonstrating that while resampling improves class representation, the introduction of synthetic samples may add noise, limiting its overall effectiveness,

The SMOTE-ENN-XGBoost technique, which combines noise removal (ENN) and class balancing (SMOTE), produced the highest overall accuracies. When paired with Yeo-Johnson and Square Root transformation strategies, the accuracy reached 93.20% and 93.00%, respectively, outperforming all other techniques. This outcome highlights the advantage of integrating both data-level resampling techniques to ensure data quality while addressing class imbalance.

6. Conclusion

In conclusion, the research highlights the effectiveness of combining feature transformation strategy and data-level resampling techniques to address skewness and class imbalance to improve the performance of the XGBoost model on the DGA dataset. Initially, the baseline accuracy of 71.30% demonstrated the limitations of raw, imbalance data. Feature transformation strategies such as Yeo-Johnson played a critical role in changing data distribution to normal, minimizing the overlapping of each data, and reducing skewness, leading to significant improvements in accuracy.

Among the data-level resampling techniques, SMOTE-ENN consistently outperformed other techniques by effectively balancing the dataset while removing noisy and out-of-boundary samples, achieving the highest accuracy of 93.20% when combined with the Yeo-Johnson transformation strategy. This result demonstrates the benefits of applying noise reduction (ENN) on synthetic balancing data (SMOTE) to ensure cleaner and more balanced data for model training.

Overall, the combination of Yeo-Johnson transformation and SMOTE-ENN is the most effective strategy, producing optimal accuracy and generalization. This technique provides a robust solution for addressing imbalance datasets, enhancing the predictive performance of machine learning models, and ensuring reliable transformer fault classification.

For future research, efforts will be directed towards scaling the technique to handle larger dataset sizes, which will present new challenges and opportunities for further model optimization. Additionally, addressing the complexities of categorical data and incorporating more advanced techniques to handle it will improve the adaptability of the models. Exploring alternative distance measures beyond Euclidean distance could lead to better performance by improving on the model calculations of similarities, especially in high-dimensional or non-linearly separable data. These directions will contribute to advancing the field of transformer fault diagnosis and improve the overall robustness and reliability of machine learning models in real-world applications.

7. Acknowledgment

The authors gratefully acknowledge the support from the Universiti Teknologi MARA (UiTM), the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, Malaysia.

8. Reference

- Bao, Y., & Yang, S. (2023). Two Novel SMOTE Methods for Solving Imbalanced Classification Problems. *IEEE Access*, 11, 5816–5823.
<https://doi.org/10.1109/ACCESS.2023.3236794>
- Becker, M., Lippel, J., Stuhlsatz, A., & Zielke, T. (2020). Robust dimensionality reduction for data visualization with deep neural networks. *Graphical Models*, 108, 101060.
- Caceres-Martinez, L. E., & Kilaz, G. (2024). Kinematic viscosity prediction of jet fuels and alternative blending components via comprehensive two-dimensional gas chromatography, partial least squares, and Yeo-Johnson transformation. *Journal of Separation Science*, 47(5). <https://doi.org/10.1002/jssc.202300816>
- Chanchotisation, P., & Vong, C. (2021). Feature engineering and feature selection for fault type classification from dissolved gas values in transformer oil. *ICSEC 2021 - 25th International Computer Science and Engineering Conference*, 75–80.
<https://doi.org/10.1109/ICSEC53205.2021.9684595>
- Gautam, A. K., Kour, J., & Shukla, A. (2023). Machine Learning Model for Transformer Health Monitoring and Fault Detection. *2023 7th International Conference on Computer Applications in Electrical Engineering-Recent Advances: Sustainable Transportation Systems, CERA 2023*.
<https://doi.org/10.1109/CERA59325.2023.10455442>
- Hernández-Carnerero, À., Sánchez-Marrè, M., Mora-Jiménez, I., Soguero-Ruiz, C., Martínez-Agüero, S., & Álvarez-Rodríguez, J. (2023). Dimensionality reduction and ensemble of LSTMs for antimicrobial resistance prediction. *Artificial Intelligence in Medicine*, 138. <https://doi.org/10.1016/j.artmed.2023.102508>
- Jahan, L. N., Munshi, T. A., Sutradhor, S. S., & Hashan, M. (2021). A comparative study of empirical, statistical, and soft computing methods coupled with feature ranking for the prediction of water saturation in a heterogeneous oil reservoir. *Acta Geophysica*, 69(5), 1697–1715. <https://doi.org/10.1007/s11600-021-00647-w>

- J. Mathew, C. K. Pang, M. Luo and W. H. Leong, "Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 9, pp. 4065-4076, Sept. 2018, doi: 10.1109/TNNLS.2017.2751612
- Jiao, L., Geng, X., & Pan, Q. (2019). BP \$ k \$ NN: \$ k \$ \$-nearest neighbor classifier with pairwise distance metrics and belief function theory. *IEEE Access*, 7, 48935-48947.
- Kvalheim, O. M., Brakstad, F., & Liang, Y. (1994). Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. *Analytical Chemistry*, 66(1), 43-51
- Li, W., Peng, X., Cheng, K., Wang, H., Xu, Q., Wang, B., & Che, J. (2020). A Short-Term Regional Wind Power Prediction Method Based on XGBoost and Multi-stage Features Selection. *2020 IEEE Student Conference on Electric Machines and Systems, SCEMS 2020*, 614–618.
<https://doi.org/10.1109/SCEMS48876.2020.9352249>
- Osborne, J. W. (2016). *Notes on the Use of Data Transformations*.
<https://www.researchgate.net/publication/200152356>
- Patil, A., Framewala, A., & Kazi, F. (2020). Explainability of SMOTE based oversampling for imbalanced dataset problems. *Proceedings - 3rd International Conference on Information and Computer Technologies, ICICT 2020*, 41–45.
<https://doi.org/10.1109/ICICT50521.2020.00015>
- Pei, H., Ren, J., Qingcai, Z., Minghui, L., Linggao, L., & Yadi, Y. (2020). Shear Velocity Prediction in the Tight Oil Formation with Deep Learning. *Proceedings - 2020 7th International Conference on Information Science and Control Engineering, ICISCE 2020*, 1274–1277. <https://doi.org/10.1109/ICISCE50968.2020.00257>
- Peng, M., Zhang, Q., Xing, X., Gui, T., Huang, X., Jiang, Y.-G., Ding, K., & Chen, Z. (2019). Trainable undersampling for class-imbalance learning. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 4707–4714.
- Ragab, A. M. S., Yakoot, M. S., & Mahmoud, O. (2021). Application of machine learning algorithms for managing well integrity in gas lift wells. *Society of Petroleum Engineers - SPE/IATMI Asia Pacific Oil and Gas Conference and Exhibition 2021, APOG 2021*. <https://doi.org/10.2118/205736-MS>

- Raichura, M., Chothani, N., & Patel, D. (2021). Efficient CNN-XGBoost technique for classification of power transformer internal faults against various abnormal conditions. *IET Generation, Transmission and Distribution*, 15(5), 972–985. <https://doi.org/10.1049/gtd2.12073>
- Sahraoui, M. A., Rahmoune, C., Zair, M., Gougam, F., & Damou, A. (n.d.). Enhancing fault diagnosis of undesirable events in oil & gas systems: A machine learning approach with new criteria for stability analysis and classification accuracy. *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering*, 0(0), 09544089231213778. <https://doi.org/10.1177/09544089231213778>
- Santos, M. S., Abreu, P. H., Fernández, A., Luengo, J., & Santos, J. (2022). The impact of heterogeneous distance functions on missing data imputation and classification performance. *Engineering Applications of Artificial Intelligence*, 111. <https://doi.org/10.1016/j.engappai.2022.104791>
- Santos, M. S., Abreu, P. H., Wilk, S., & Santos, J. (2020). How distance metrics influence missing data imputation with k-nearest neighbours. *Pattern Recognition Letters*, 136, 111–119. <https://doi.org/10.1016/j.patrec.2020.05.032>
- Sheng, C., & Yu, H. (2022). An optimized prediction algorithm based on XGBoost. *Proceedings - 2022 International Conference on Networking and Network Applications, NaNA 2022*, 442–447. <https://doi.org/10.1109/NaNA56854.2022.00082>
- Singh Rawat, S., & Kumar Mishra, A. (2022). *Review of Methods for Handling Class-Imbalanced in Classification Problems*.
- Su, Y., Xia, Y., & Zhang, R. (2022). A Missing Data Tolerance Data-driven Method for Open-Circuit Fault Diagnosis of Three-phase Inverters Based on Random Forest and Resampling Scheme. *Proceedings of the 11th International Conference on Innovative Smart Grid Technologies - Asia, ISGT-Asia 2022*, 359–363. <https://doi.org/10.1109/ISGTAsia54193.2022.10003581>
- Thinh, H. X., & Van Dua, T. (2024). Optimal Surface Grinding Regression Model Determination with the SRP Method. *Engineering, Technology and Applied Science Research*, 14(3), 14713–14718. <https://doi.org/10.48084/etasr.7573>
- Trivedi, M. (2017). Unit-4 Skewness and Kurtosis. *Descriptive Statistics*, 67–72.
- Wah, Y. B., Ismail, A., Azid, N. N. N., Jaafar, J., Aziz, I. A., Hasan, M. H., & Zain, J. M. (2023). Machine Learning and Synthetic Minority Oversampling Techniques for

- Imbalanced Data: Improving Machine Failure Prediction. *Computers, Materials and Continua*, 75(3), 4821–4841. <https://doi.org/10.32604/cmc.2023.034470>
- Wang, B., Li, T., Xu, N., Zhou, H., Xiong, Z., & Long, W. (2021). A Novel Reservoir Modeling Method based on Improved Hierarchical XGBoost. *IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 1918–1923. <https://doi.org/10.1109/IAEAC50856.2021.9390679>
- Wu, M., Wang, G., & Liu, H. (2022). Research on Transformer Fault Diagnosis Based on SMOTE and Random Forest. *Proceedings - 2022 4th International Conference on Electrical Engineering and Control Technologies, CEECT 2022*, 359–363. <https://doi.org/10.1109/CEECT55960.2022.10030548>
- Xi, H., Luo, Z., & Guo, Y. (2025). Reservoir evaluation method based on explainable machine learning with small samples. *Unconventional Resources*, 5. <https://doi.org/10.1016/j.uncres.2024.100128>
- Xing, E., Jordan, M., Russell, S. J., & Ng, A. (2002). Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 15.
- Xu, J. Z., Guo, Z., Zhang, M., Li, X., Li, Y. J., & Rao, S. Q. (2006). Peeling off the hidden genetic heterogeneities of cancers based on disease-relevant functional modules. *Molecular Medicine*, 12(1–3), 25–33. <https://doi.org/10.1016/j.chroma.2005.09.076>
- Xu, L., Wang, Y., Mo, L., Tang, Y., Wang, F., & Li, C. (2023). The research progress and prospect of data mining methods on corrosion prediction of oil and gas pipelines. In *Engineering Failure Analysis* (Vol. 144). Elsevier Ltd. <https://doi.org/10.1016/j.engfailanal.2022.106951>
- Xu, Z., Shen, D., Nie, T., & Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *Journal of Biomedical Informatics*, 107. <https://doi.org/10.1016/j.jbi.2020.103465>
- Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists* (1st ed.). O'Reilly Media, Inc.
- Zhu, R., Guo, Y., & Xue, J. H. (2020). Adjusting the imbalance ratio by the dimensionality of imbalanced data. *Pattern Recognition Letters*, 133, 217–223. <https://doi.org/10.1016/j.patrec.2020.03.004>