P-ISSN: 2656-1743 | E-ISSN: 2656-1735 **JURNAL RISET INFORMATIKA**
DOI: https://doi.org/10.34288/jri.v8i3.539 Vol. 8, No. 3. Juni 2026
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

MACHINE LEARNING APPROACH FOR TRANSFORMER CONDITION ASSESSMENT USING K-MEANS CLUSTERING AND MULTI-CLASSIFIER MODELS

Zulfiana Safitri Majid[-1], Andarini Asri[-2], Musfirah Putri Lukman[-3], Wisna Saputri Alfira WS[-4], Auliya Nabila[-5]

Electrical Engineering Department
State Polytechnic of Ujung Pandang
zulfianasafitri@poliupg.ac.id[-1], andariniasri@poliupg.ac.id[-2], musfirahputrilukman@poliupg.ac.id[-3], alfirasaputri@poliupg.ac.id[-4], auliyanabila@poliupg.ac.id[-5]

Abstract

Transformers play a critical role in power systems, yet their degradation is often difficult to detect due to complex influencing factors. Conventional diagnostic methods, such as Dissolved Gas Analysis (DGA), are time-consuming and rely heavily on expert interpretation. This study proposes a machine learning approach for transformer condition assessment by combining clustering and classification techniques. K-Means clustering is first applied to identify patterns in transformer condition data without prior labeling, with the optimal number of clusters determined as three using the Elbow Method. The resulting clusters are then used as pseudo-labels to train multiple classification models, including KNN, Decision Tree, SVM, Gradient Boosting, Extra Trees, and Voting Classifier. The results show that all models achieve high performance, with accuracy above 94%. Ensemble methods, particularly Gradient Boosting and Voting Classifier, achieve the best performance with an accuracy of 98.30%. These findings demonstrate that the proposed approach effectively improves transformer condition assessment and supports faster and more reliable maintenance decision-making.

Keywords: Transformer Fault; Machine Learning; Clustering; Classification; DGA

Abstrak

Transformator memegang peranan penting dalam sistem tenaga listrik, namun proses degradasinya sering kali sulit dideteksi karena dipengaruhi oleh berbagai faktor yang kompleks. Metode diagnostik konvensional, seperti Dissolved Gas Analysis (DGA), memerlukan waktu yang relatif lama serta sangat bergantung pada interpretasi para ahli. Penelitian ini mengusulkan pendekatan machine learning untuk penilaian kondisi transformator dengan mengombinasikan teknik klasterisasi dan klasifikasi. Metode K-Means terlebih dahulu diterapkan untuk mengidentifikasi pola pada data kondisi transformator tanpa pelabelan awal, dengan jumlah klaster optimal sebanyak tiga yang ditentukan menggunakan metode Elbow. Hasil klasterisasi tersebut kemudian digunakan sebagai pseudo-label untuk melatih beberapa model klasifikasi, meliputi KNN, Decision Tree, SVM, Gradient Boosting, Extra Trees, dan Voting Classifier. Hasil penelitian menunjukkan bahwa seluruh model memiliki kinerja yang tinggi dengan tingkat akurasi di atas 94%. Metode ensemble, khususnya Gradient Boosting dan Voting Classifier, memberikan performa terbaik dengan akurasi mencapai 98,30%. Temuan ini menunjukkan bahwa pendekatan yang diusulkan mampu meningkatkan efektivitas penilaian kondisi transformator serta mendukung pengambilan keputusan pemeliharaan yang lebih cepat dan andal.

Kata kunci: Kegagalan Transformator; Machine Learning; Klasterisasi; Klasifikasi; DGA

INTRODUCTION

Transformers perform a critical role in electrical power systems (Majid, Ariwibowo, et al., 2025); however, in practice, they often experience degradation or failures whose root causes are not immediately identified (Arumugam, 2021). This uncertainty arises because multiple factors, including thermal stress, moisture contamination, and chemical reactions within the insulating oil, influence transformer degradation (Medina et al., 2017). As a result, failures may develop gradually and remain undetected until they reach a critical stage.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.539
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Conventional maintenance practices, such as periodic inspections and laboratory-based oil testing (e.g., Dissolved Gas Analysis and oil quality measurements), remain widely used. Although these methods provide valuable diagnostic information, they typically require significant time for data collection, analysis, and expert interpretation. This time-consuming process can delay fault identification, reducing the effectiveness of preventive maintenance and increasing the risk of unexpected transformer failure. Therefore, there is a need for a more efficient, data-driven approach to transformer condition assessment (Majid, Bachtiar, et al., 2025).

In recent years, Machine learning has emerged as a promising solution by enabling automatic pattern recognition from historical and operational data (Kurniawan et al., 2019; Ridho Tri Putra Nanda, 2023). However, a critical but often overlooked challenge is that transformers experiencing early-stage degradation frequently go undetected during routine maintenance or conventional assessment procedures, because traditional diagnostic methods are not designed to discover hidden patterns embedded in multivariate transformer data. As a result, fault conditions may persist unrecognized until they reach a critical stage. Most existing machine learning studies address this only partially, relying on individual supervised learning methods that require pre-labeled datasets and thus fail to uncover latent data structures. Furthermore, the combined use of unsupervised clustering and multiple supervised classifiers within a single integrated framework—which enables both pattern discovery and condition prediction without prior labeling—has not been previously explored in a comprehensive and systematic manner.

To address these gaps, this study proposes a hybrid approach combining unsupervised clustering and supervised classification techniques. In the first stage, K-Means clustering is applied to group transformer data based on similarities in gas composition and oil properties, enabling the identification of hidden degradation patterns without requiring prior labeling. In the second stage, multiple classification algorithms—K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), Gradient Boosting Classifier, Extra Trees Classifier, and Voting Classifier—are trained on the cluster-derived labels to learn fault characteristics and predict transformer conditions. This two-stage framework represents, to the best of the authors' knowledge, the first systematic comparison of these classification models within a unified clustering-based pipeline for transformer condition assessment.

Although some studies have applied similar methods individually, a comprehensive comparison of these classification algorithms within a single framework is still limited. Therefore, this research aims to evaluate and identify the most effective model for transformer condition assessment.

By integrating K-Means clustering with multiple classification models, the proposed approach provides a robust framework that not only uncovers hidden degradation patterns but also determines the most accurate model for condition prediction. This is expected to improve diagnostic efficiency, support better maintenance decision-making, and reduce operational risks and costs associated with transformer failures. Specifically, the primary contribution of this study is to provide a machine learning-based condition assessment method that enables earlier detection of critical transformer conditions—such as active faults, moisture-dominated degradation, and chemical contamination—so that maintenance actions can be taken before these conditions escalate. Early detection directly translates to reduced maintenance costs, prevention of unplanned outages, and a significant reduction in the risk of catastrophic failure, including transformer oil degradation, short circuits, and in the worst case, explosions.

RESEARCH METHODS

This study proposes a hybrid machine learning framework that integrates clustering and classification techniques for transformer condition assessment. The overall process consists of several stages: data collection, data preprocessing, clustering using K-Means, and classification using multiple machine learning algorithms. The workflow is designed first to identify hidden patterns in the data and then to evaluate the performance of different classifiers in predicting transformer conditions.

Data Collection

The dataset used in this study consists of transformer condition data obtained from oil analysis, particularly Dissolved Gas Analysis (DGA). A total of 470 data samples were collected from transformer oil testing results. Each sample includes concentrations of key gases, namely hydrogen (H_2), methane (CH_4), ethane (C_2H_6), ethylene (C_2H_4), and acetylene (C_2H_2), which are widely used as indicators of transformer faults.

The selection of these gases is based on international standards, specifically IEC 60599 and IEEE C57.104, which relate gas concentration patterns to specific fault types such as thermal faults, partial discharge, and arcing. In addition to DGA parameters, the dataset also includes Dibenzyl Disulfide (DBDS), an important chemical indicator associated with sulfur contamination in transformer oil. High DBDS concentration may lead to copper sulfide formation and insulation deterioration, which can negatively affect transformer reliability and lifetime.

These parameters are widely recognized as reliable indicators for identifying fault characteristics and supporting machine learning-based analysis for transformer condition assessment.

Data Preprocessing

Before applying machine learning models, several preprocessing steps are performed to ensure data quality and improve model performance.

Data Cleaning: Missing values are handled appropriately, either by removal or imputation, while outliers are identified and treated to reduce their impact on model accuracy.

Each numerical feature was subjected to the Interquartile Range (IQR) approach in order to identify and eliminate outlier data points. The difference between the dataset's third quartile (Q_3) and first quartile (Q_1), expressed as a difference, was used to determine the IQR (Mazarei et al., 2025).

$$IQR = Q_3 - Q_1 \quad (1)$$

If data points met the following criteria, they were deemed outliers:

$$(x < Q_1 - 1.5\,x\,IQR)\ or\ (x > Q_3 + 1.5\,x\,IQR) \quad (2)$$

The Interquartile Range (IQR) approach, where Q_1 and Q_3 stand for the 25th and 75th percentiles of each feature, was used to detect outliers. To lower noise and increase data dependability, values that fell below Q_1 − 1.5× IQR or above Q_3 + 1.5× IQR were considered outliers and eliminated. This process maintains the underlying distribution necessary for precise clustering and classification while removing extreme values that are probably the result of measurement errors or temporary disruptions.

Normalization: Feature scaling is applied to transform all variables into a comparable range, preventing features with larger numerical values from dominating the learning process. Equation (3) illustrates how all numerical variables were standardized using Min–Max scaling (Sinsomboonthong, 2022) to a range of 0–1 after outlier removal. In the absence of scaling, distance-based algorithms like K-Means and KNN may be dominated by variables with greater magnitudes. By putting all features into a consistent range, min-max scaling guarantees that each feature contributes proportionately, enhancing model stability and avoiding bias in clustering and classification.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

where X stands for the initial value, X_{min} and X_{max} are the feature's minimum and maximum observed values, and X_{norm} is the normalized value generated via Min–Max scaling. In order to avoid high-magnitude features controlling smaller-scale variables during clustering and classification, this normalization was required.

Feature Selection: Relevant parameters, particularly key gas concentrations from DGA, are selected based on their significance in representing transformer fault characteristics, while redundant or less informative features are excluded.

These preprocessing steps are crucial to enhance data consistency, reduce noise, and ensure that the machine learning models can effectively learn underlying patterns for accurate classification.

K-Means Clustering

In the first stage, an unsupervised learning approach is applied using the K-Means clustering algorithm to group transformer data into distinct clusters based on similarities in dissolved gas concentrations and oil characteristics. This step aims to uncover hidden patterns in the dataset without requiring prior labeling.

K-Means partitions the dataset into k clusters by minimizing the distance between each data point and the centroid of its assigned cluster. The objective function is defined as:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (4)$$

where J represents the total within-cluster variance, C_i denotes the set of data points in the i -th cluster, x is an individual data point, and μ_i is the centroid of cluster i. The centroid is computed as

the mean of all data points within a cluster (MacQueen, 1967):

$$\mu_i = \frac{1}{|C_i|}\sum_{x \in C_i} x \quad (5)$$

The similarity between data points and centroids is commonly measured using the Euclidean distance (Bishop & Nasrabadi, 2006; Park & Lee, 2011).

$$d(x, \mu_i) = \sqrt{\sum_{j=1}^{n}(x_j - \mu_{ij})^2} \quad (6)$$

where n is the number of features. The algorithm iteratively performs two main steps: (1) assigning each data point to the nearest centroid, and (2) updating centroids based on the mean of assigned data points, until convergence is achieved (Chakraborty et al., 2024; Hastie et al., 2009; Ikotun et al., 2023; Suyal & Sharma, 2024).

Before clustering, the preprocessed data are used to ensure that all features contribute equally to the distance calculation. The optimal number of clusters (k) is determined using evaluation methods such as the Elbow Method and Silhouette Score, which assess cluster compactness and separation.

The resulting clusters represent groups of transformer conditions with similar characteristics and are subsequently used as pseudo-labels for the classification stage.

Classification Methods

After clustering, the labeled clusters are used as target classes for the classification stage. Several supervised learning algorithms are applied to evaluate their performance in predicting transformer conditions.

1. K-nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm that classifies a data point based on the majority class of its k nearest neighbors in the feature space. Unlike model-based approaches, KNN does not require an explicit training phase, as it directly uses the entire dataset during the classification process. The similarity between data points is typically measured using the Euclidean distance, defined as:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n}(x_j - x_{ij})^2} \quad (7)$$

where x represents the query data point, x_i is a training data point, and n is the number of features. The algorithm computes the distance between the query point and all training samples, selects the k closest neighbors, and assigns the class label based on majority voting among these neighbors.

The choice of the parameter k plays a crucial role in model performance. A small value of k may lead to overfitting and sensitivity to noise, while a large value can smooth decision boundaries but may reduce classification accuracy. Therefore, an optimal k is selected through empirical evaluation. In this study, KNN is utilized due to its simplicity and effectiveness in capturing local data structures (Singh et al., 2017; Zhang et al., 2017; Zhao et al., 2013). It is particularly suitable for transformer condition data, where similar gas composition patterns are expected to belong to the same fault category.

2. Decision Tree

Decision Tree is a supervised learning algorithm that classifies data by recursively partitioning the feature space into subsets based on feature values. The model is represented as a tree structure consisting of decision nodes and leaf nodes, where each internal node corresponds to a test on a feature, and each leaf node represents a class label. The splitting process is typically guided by impurity measures such as Gini Index or Entropy. For example, the Gini Index is defined as (Anwar et al., 2020; Dicoding Indonesia, 2024):

$$Gini = 1 - \sum_{i=1}^{c} p_i^2 \quad (8)$$

where p_i is the probability of class i in a node and c is the number of classes. The algorithm selects the feature that results in the highest reduction in impurity.

Decision Tree models are widely used due to their interpretability and ability to handle nonlinear relationships. Recent studies highlight their continued relevance and integration into advanced ensemble methods due to their simplicity and strong performance on tabular data (Balcan & Sharma, 2024).

3. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm that constructs an optimal hyperplane to separate data with maximum margin. It is based on statistical learning theory and the maximum margin principle, which aims to improve generalization performance (Du et al., 2024). The decision boundary is defined as:

$$w.x + b = 0 \quad (9)$$

where w is the weight vector and b is the bias term. The optimal hyperplane maximizes the margin between support vectors, which are the closest data points to the boundary.

For non-linearly separable data, SVM utilizes kernel functions such as the radial basis function (RBF) to project data into a higher-dimensional space, enabling effective classification.

SVM is particularly effective for high-dimensional data and is included in this study to evaluate its capability in distinguishing complex transformer fault patterns. Recent research demonstrates that SVM remains a competitive classification method, especially when combined with optimization and ensemble techniques to improve accuracy(Khan et al., 2024).

4. Gradient Boosting Classifier

Gradient Boosting Classifier is an ensemble learning method that builds a strong predictive model by combining multiple weak learners, typically decision trees, in a sequential manner. Each new model is trained to correct the errors of the previous models by minimizing a loss function using gradient descent.

The general formulation can be expressed as:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \qquad (10)$$

Where $F_m(x)$is the updated model, $h_m(x)$ is the new weak learner, and γ_m is the learning rate(Sable et al., 2024).

This method is known for its high predictive accuracy and robustness. Recent studies confirm that Gradient Boosting remains one of the most powerful techniques for predictive modeling due to its ability to capture complex nonlinear relationships and improve performance iteratively.

5. Extra Trees Classifier

Extra Trees (Extremely Randomized Trees) is an ensemble learning method that constructs multiple decision trees using random subsets of features and random split points. Unlike Random Forest, which searches for optimal splits, Extra Trees introduces additional randomness by selecting split thresholds randomly. This randomness reduces variance and helps prevent overfitting, while maintaining computational efficiency. The final prediction is obtained by aggregating the outputs of all trees, typically using majority voting.

Tree-based ensemble methods, including Extra Trees and Random Forest variants, remain highly effective for tabular datasets and are widely adopted due to their robustness and computational efficiency in real-world applications (Pensa et al., 2025).

6. Voting Classifier

Voting Classifier is an ensemble technique that combines predictions from multiple models to improve overall classification performance. It aggregates outputs using either majority voting (hard voting) or probability averaging (soft voting).

Recent studies emphasize that ensemble strategies, including voting-based approaches, can significantly enhance classification accuracy by leveraging the strengths of different models and reducing individual model bias and variance(Khan et al., 2024)

RESULTS AND DISCUSSION

Before conducting the clustering and classification processes, the dataset underwent several preprocessing steps to ensure data quality and reliability. These steps included data cleaning to remove missing values and anomalous data, followed by outlier handling to minimize the influence of extreme values. Subsequently, normalization was applied to scale all features into a comparable range, ensuring that each parameter contributed equally during the modeling process.

Clustering Result

After preprocessing, the optimal number of clusters was determined using the Elbow Method. This method evaluates the distortion score for different numbers of clusters (k) to identify the point at which adding more clusters does not significantly improve clustering performance. Based on the visualization results (as shown in Fig. 1), an elbow point is observed at k=3, with a distortion score of 72.045.

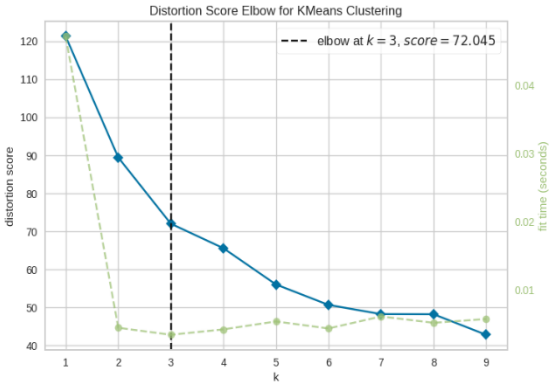

Figure 1. Elbow for K-Means Clustering

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.539
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

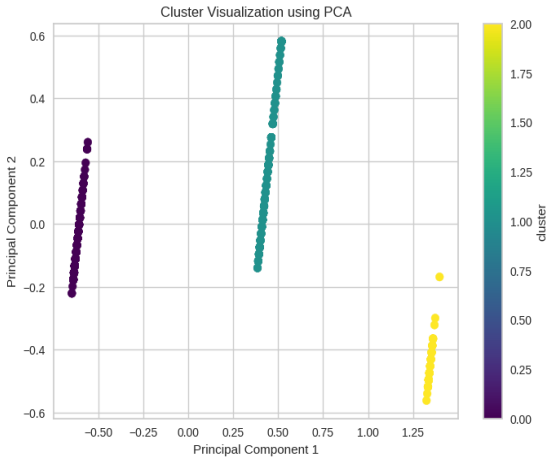

Figure 2. Clustering Visualization

The PCA-based visualization demonstrates that the dataset is clearly divided into three distinct clusters. Each cluster is well separated in the two-dimensional space, with minimal overlap, indicating strong clustering performance. The separation is primarily driven by the first principal component (PC1), which effectively differentiates the groups. These results confirm that the chosen number of clusters (k = 3) is appropriate and that the clustering structure is robust. Therefore, it can be concluded that the optimal number of clusters for this dataset is three, and this value is used in the subsequent clustering analysis.

Table 1. Clustering Result

Feature	Cluster 1	Cluster 2	Cluster 3
Fault Gases	High	Low	Low
Water Content	Medium	Very High	Low
DBDS	Low	Low	Very High
Interfacial Voltage	High	Low	Medium
Condition	Active Fault	Moisture	Chemical Contamination

Based on the selected optimal number of clusters, the K-Means algorithm was applied to group the transformer condition data. The statistical characteristics of each cluster are presented in Table 1, and the interpretation of each cluster is discussed as follows.

Cluster 1: Active Fault Condition

Cluster 1 contains the largest portion of the dataset, consisting of 237 samples. This cluster is characterized by high concentrations of fault gases, including hydrogen (H_2), methane (CH_4), ethylene (C_2H_4), and acetylene (C_2H_2), which are key indicators of active fault conditions. Additionally, carbon monoxide (CO) and carbon dioxide (CO_2), associated with insulation aging, are also present at elevated levels.

In addition, the interfacial voltage is relatively high, and the water content is moderate, indicating that the oil quality is still within acceptable limits. These characteristics suggest that this cluster represents transformers operating under normal conditions or experiencing moderate aging without significant fault development.

Cluster 2: Moisture (Moisture-Dominated Degradation)

Cluster 2 consists of 186 samples and is primarily distinguished by a significantly higher water content compared to other clusters. This cluster also exhibits low interfacial voltage values and increased power factor, indicating a degradation in oil quality.

Although the concentrations of fault gases remain relatively low, the high moisture level suggests a deterioration of the insulating medium, which can reduce dielectric strength and accelerate aging processes. Therefore, this cluster represents transformers with degraded oil conditions, where moisture contamination is the dominant factor and poses a high risk for future failure if not properly managed.

Cluster 3: Chemical Contamination (DBDS-Related)

Cluster 3 contains 47 samples and is uniquely characterized by a very high concentration of DBDS (Dibenzyl Disulfide), while other gas concentrations remain relatively low. This pattern indicates that the dominant issue in this cluster is not electrical or thermal fault, but chemical contamination.

High DBDS levels are associated with sulfur-related corrosion, which can lead to the formation of copper sulfide deposits on transformer windings. This type of degradation is particularly critical as it can compromise insulation performance and lead to long-term reliability issues. Therefore, this cluster represents transformers affected by chemical degradation mechanisms.

Furthermore, key parameters such as water content and DBDS are shown to have strong discriminative power in separating different clusters, highlighting their importance in transformer condition assessment. The clustering

results are subsequently used as pseudo-labels for the classification stage, enabling further evaluation of machine learning models.

Classification Result

The clustering results obtained from the K-Means algorithm were subsequently used as pseudo-labels for the classification stage. In the confusion matrices presented in Figure 3, class indices 0, 1, and 2 correspond to Cluster 1 (Active Fault Condition), Cluster 2 (Moisture-Dominated Degradation), and Cluster 3 (Chemical Contamination), respectively.

Table 2. Classification Result

Model	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors (KNN)	94.04	0.9451	0.9404	0.9408
Decision Tree (DT)	96.17	0.9628	0.9617	0.9617
Support Vector Machine (SVM)	95.74	0.9599	0.9574	0.9575
Gradient Boosting Classifier (GBC)	98.30	0.9838	0.9830	0.9830
Extra Trees Classifier (ETC)	96.60	0.9666	0.9660	0.9660
Voting Classifier (VC)	98.30	0.9832	0.9830	0.9830

Several machine learning models were evaluated, including K-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting Classifier (GBC), Extra Trees Classifier (ETC), and Voting Classifier (VC). The performance of each model was assessed using accuracy, precision, recall, F1-score, and confusion matrix. As shown in Table 2, all classification models achieved high performance, with accuracy values exceeding 94%, indicating that the clustering-based pseudo-labeling approach provides a reliable foundation for supervised learning. However, notable differences can be observed among the models in terms of classification performance.

The best-performing models are the Gradient Boosting Classifier (GBC) and Voting Classifier (VC), both achieving the highest accuracy of 98.30%, along with superior precision, recall, and F1-score values. The confusion matrices of these models indicate that almost all samples were successfully grouped into their corresponding clusters, particularly for Cluster 3, which shows highly consistent classification results. The strong performance of these models can be attributed to their ensemble learning mechanisms. Gradient Boosting improves performance by sequentially refining prediction errors, allowing it to capture complex and nonlinear relationships in the data. Similarly, the Voting Classifier combines multiple models to leverage their complementary strengths, resulting in more robust predictions.

Extra Trees Classifier (ETC) and Decision Tree (DT) also demonstrate strong performance, with accuracy values of 96.60% and 96.17%, respectively. The confusion matrix patterns suggest that Cluster 1 and Cluster 2 exhibit relatively similar characteristics, particularly in terms of gas composition and oil properties, which explains the slightly lower separation capability compared to ensemble methods. Nevertheless, Cluster 3 remains consistently identified with high accuracy across both models. These tree-based approaches are effective in handling nonlinear feature interactions and heterogeneous data distributions.

The Support Vector Machine (SVM) achieves a competitive accuracy of 95.74%. The classification results indicate that SVM performs effectively in separating well-defined transformer condition patterns, especially for Cluster 3. However, the similarity between Cluster 1 and Cluster 2 may reduce the distinctiveness of the decision boundary, slightly affecting overall performance.

K-Nearest Neighbors (KNN) produces the lowest performance among the evaluated models, with an accuracy of 94.04%. The classification pattern indicates that KNN is more influenced by local similarities among samples, particularly between Cluster 1 and Cluster 2. Since KNN relies heavily on distance-based measurements, its performance can be affected when different transformer conditions exhibit closely related DGA characteristics.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.539
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

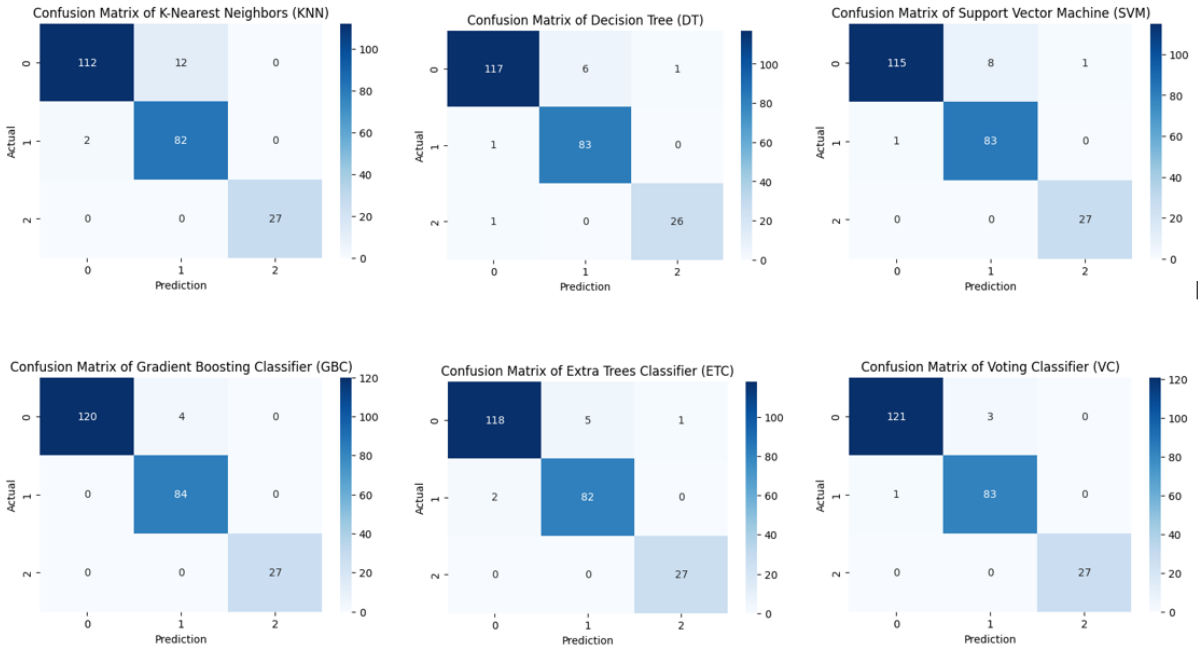

Figure 3. Confusion Matrix of Models

Overall, the confusion matrix analysis shows that Cluster 3 is the most distinguishable condition category across all models, likely due to its unique DBDS-related characteristics. In contrast, Cluster 1 and Cluster 2 demonstrate more similar DGA and oil property patterns, reflecting the complexity of distinguishing aging-related and moisture-related degradation mechanisms. These findings confirm that ensemble-based methods provide the most robust and reliable performance for transformer condition classification.

From a practical and interpretability standpoint, each identified cluster carries direct implications for transformer health management. Cluster 1 (Active Fault Condition) indicates that fault gases have reached elevated levels, signaling active thermal or electrical stress within the transformer. If left unaddressed, such conditions accelerate the breakdown of insulating oil and solid insulation, potentially leading to inter-winding short circuits. Cluster 2 (Moisture-Dominated Degradation) represents a condition where high water content has significantly reduced the dielectric strength of the transformer oil, impairing its ability to insulate high-voltage components. Prolonged moisture contamination weakens insulation barriers and raises the risk of partial discharge escalating into a full internal arc. Cluster 3 (Chemical Contamination) reflects elevated DBDS concentration, which drives sulfur corrosion of copper windings and generates conductive copper sulfide deposits on insulation surfaces—a progressive failure mechanism that may ultimately result in insulation puncture and short circuit. In the most severe scenarios across all three conditions, the combination of internal arcing, flammable gas accumulation, and oil degradation can culminate in a transformer explosion. The ability of the proposed machine learning framework to classify these conditions early—before they reach a catastrophic stage—therefore represents a meaningful contribution to transformer safety and cost-effective maintenance planning.

CONCLUSION

This study proposes a hybrid approach combining K-Means clustering and classification models for transformer condition assessment using DGA data. Three distinct condition patterns were identified, representing active fault conditions, moisture-dominated degradation, and chemical contamination—each carrying direct implications for transformer insulating oil performance and long-term reliability. All models achieved accuracy above 94%, with Gradient Boosting and Voting Classifier performing best at 98.30%. Misclassifications mainly occurred between clusters with similar DGA characteristics. The key contribution of this work lies in its ability to detect critical transformer conditions at an early stage, enabling timely maintenance intervention before degradation escalates to severe failures such as short circuits or, in worst-case scenarios,

explosions. This early detection capability is expected to reduce unplanned outages, lower maintenance costs, and extend transformer service life.

Overall, the proposed method provides an effective and reliable approach for transformer condition monitoring.

REFERENCES

Anwar, M. T., Hadikurniawati, W., Winarno, E., & Widiyatmoko, W. (2020). Performance Comparison of Data Mining Techniques for Rain Prediction Models in Indonesia. *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 83–88.

Arumugam, S. (2021). Failure diagnosis and root-cause analysis of in-service and defective distribution transformers. *Energies*, *14*(16), 4997.

Balcan, M.-F., & Sharma, D. (2024). Learning accurate and interpretable tree-based models. *ArXiv Preprint ArXiv:2405.15911*.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, Number 4). Springer.

Chakraborty, J., Pradhan, D. K., & Nandi, S. (2024). A multiple k-means cluster ensemble framework for clustering citation trajectories. *Journal of Informetrics*, *18*(2), 101507.

Dicoding Indonesia. (2024, October 25). *Belajar Machine Learning untuk Pemula*. Dicoding Academy.

Du, K.-L., Jiang, B., Lu, J., Hua, J., & Swamy, M. N. S. (2024). Exploring kernel machines and support vector machines: Principles, techniques, and future directions. *Mathematics*, *12*(24), 3935.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer series in statistics New-York.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210.

Khan, M., Hooda, B. K., Gaur, A., Singh, V., Jindal, Y., Tanwar, H., Sharma, S., Sheoran, S., Vishwakarma, D. K., & Khalid, M. (2024). Ensemble and optimization algorithm in support vector machines for classification of wheat genotypes. *Scientific Reports*, *14*(1), 22728.

Kurniawan, A., Rahmawati, Y., & Putranto, H. (2019). Studi Performa Transformator Daya Menggunakan Metode Health Index di Gardu Induk Waru Sidoarjo. *SinarFe7*, *2*(1), 33–38.

MacQueen, J. (1967). Multivariate observations. *Proceedings Ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability*, *1*, 281–297.

Majid, Z. S., Bachtiar, M. I., Faraby, M. D., Widyaningsih, D., Nurfadhilah, A., & Ridhwan, M. (2025). Machine Learning-Based Analysis of Transformer Insulation Failure. *2025 8th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 245–250.

Mazarei, A., Sousa, R., Mendes-Moreira, J., Molchanov, S., & Ferreira, H. M. (2025). Online boxplot derived outlier detection. *International Journal of Data Science and Analytics*, *19*(1), 83–97.

Medina, R. D., Romero, A. A., Mombello, E. E., & Rattá, G. (2017). Assessing degradation of power transformer solid insulation considering thermal stress and moisture variation. *Electric Power Systems Research*, *151*, 1–11.

Park, Y. S., & Lee, Y.-S. (2011). Diagnostic cluster analysis of mathematics skills. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *4*, 75–108.

Pensa, R. G., Crombach, A., Peignier, S., & Rigotti, C. (2025). Explaining Random Forest and XGBoost with Shallow Decision Trees by Co-clustering Feature Importance. *Machine Learning*, *114*(12), 287.

Ridho Tri Putra Nanda, M. (2023). *Perhitungan Health Index Untuk Menentukan Umur dan Kondisi Transformator Daya Kapasitas 80 MVA*.

Sable, N. P., Patil, R. V, Deore, M., Bhimanpallewar, R., & Mahalle, P. N. (2024). Machine Learning Based Agricultural Profitability Recommendation Systems: A Paradigm Shift in Crop Cultivation. *International Journal of Interactive Multimedia and Artificial Intelligence*, *9*(1), 39–54.

Singh, A., Halgamuge, M. N., & Lakshmiganthan, R. (2017). Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms. *International Journal of Advanced Computer Science and Applications*, *8*(12).

Sinsomboonthong, S. (2022). Performance comparison of new adjusted min-max with decimal scaling and statistical column normalization methods for artificial neural network classification. *International Journal*

of Mathematics and Mathematical Sciences, *2022*(1), 3584406.

Suyal, M., & Sharma, S. (2024). A review on analysis of k-means clustering machine learning algorithm based on unsupervised learning. *Journal of Artificial Intelligence and Systems*, *6*(1), 85–95.

Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2017). Efficient kNN classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(5), 1774–1785.

Zhao, S., Rui, C., & Zhang, Y. (2013). MICkNN: multi-instance covering kNN algorithm. *Tsinghua Science and Technology*, *18*(4), 360–368.