Condition Monitoring of Power Transformers using Artificial Intelligence: Addressing Class Imbalance

Elekanyani Siphuma
Institute of Intelligent Systems
University of Johannesburg
Johannesburg, South Africa
elekanyanisiphuma@hotmail.co.za

***Abstract*—Artificial intelligence (AI) plays a pivotal role in improving maintenance strategies for power transformer fault diagnosis. However, many dissolved gas analysis (DGA) datasets are highly imbalance, adversely affecting machine learning model performance in transfoermer fault diagnosis task. This study assesses the influence of various resampling methods on the predictive performance of machine learning classifiers. The research focuses on five common AI models: Support Vector Machine (SVM), Decision Tree, Random Forest (RF), Logistic Regression (LR), and k-Nearest Neighbor (KNN). Classifier performance is assessed for diagnostic fault classification, with a comparison to an imbalanced dataset. The evaluation metrics used to asses classifier performance include accuracy, area under the receiver operating characteristic curve (ROC-AUC), precision, recall, and F-score. The results of this study confirm that applying resampling methods to AI models effectively mitigates challenges arising from imbalanced data distributions, significantly enhancing fault diagnosis performance, particularly for minority classes. Notably, the study identifies Random Forest as the top-performing classifier for transformer fault diagnosis. In the context of fault diagnosis within each class, the SMOTE ENN method emerged as the top-performing resampling method.**

Keywords: Transformer Fault Diagnosis, Imbalanced Datasets, Resampling Methods, Machine Learning.

I. Introduction

Power transformers serve as a critical and capital-intensive asset within the utility sector [1]. Power Transformers are primarily used to step up or down the electrical voltage of alternating current (AC current) while maintaining consistent frequency. By doing so, electricity is delivered safely and efficiently from the power generators to the end users [1]. While power transformers are widely recognized for their robust reliability, a considerable subset of units presently utilized by utility entities has exceeded their initially projected operational lifespan, this circumstance introduces a tangible risk of unanticipated overloads, potentially culminating life-threatening scenarios and necessitating power loss [2]. Consequently, enhancing the operational lifespan of power transformers has emerged as a crucial strategy due to the rising demand for electricity supply [2]. Hence, condition monitoring of the power transformer is a valuable research topic.

Throughout the transformer's lifespan, opportunities for major adjustments or complete overhauls involving dismantling are rare, limiting direct inspection of internal components, especially oil-immersed ones [3]. Thus, transformer fault diagnosis relies heavily on preventive tests, with two main approaches: real-time online monitoring and interruption-based maintenance [4]. The latter offers precise component condition identification but comes with high costs and a risk of missing internal incipient faults during operation [4]. Online monitoring, on the other hand, eliminates the need for shutdown and allows timely detection of early faults, making it the predominant method for transformer fault diagnosis [3], [4]. Dissolved Gas Analysis (DGA) and partial discharge testing are fundamental techniques in online monitoring [5]. While partial discharge testing faces interference challenges, DGA remains the primary method, analyzing gases like hydrogen (H_2), methane (CH_4), ethylene (C_2H_4), ethane (C_2H_6), acetylene (C_2H_2), carbon monoxide (CO), and carbon dioxide (CO_2) in transformer oil [6].

The International Electrotechnical Commission (IEC) established the three-gas ratio method in Dissolved Gas Analysis (DGA), using hydrogen, methane, and ethylene concentrations in transformer oil to assess transformer health accurately [7]. However, this method has limitations, as it only detects specific faults using certain gas ratios, potentially missing other fault types [3]. Additionally, interpreting three-gas ratio method requires expert knowledge, incurring costs [3]. To address these issues, the integration of machine learning algorithms, trained on extensive DGA datasets, offers a promising solution for independent transformer health prediction. Several studies have explored machine learning applications in transformer fault diagnosis. Cheng and Yu's study [3] examined the use of the random forest model, revealing superior prediction accuracy compared to conventional Back Propagation neural networks (BPNN). Valuable insights into influential factors were gained, enhancing the field of transformer fault diagnosis [3]. Kari et al. [8] integrated Genetic Algorithm (GA) and Support Vector Machine (SVM) to optimize parameters and select feature subsets in the DGA dataset, affirming the robustness and practicality of the best feature subset identified by the GA-SVM combination [8]. The study by MehdipourPicha et al. [9] employed a deep neural network (DNN) to identify transformer fault types within Duval triangles. Their research demonstrated the DNN's superior prediction accuracy when compared to the k-nearest neighbor (k-NN) and random forest algorithms across datasets of varying sizes [9]. Maumela [10] explored the

impact of attribute reduction on machine learning algorithms for detecting transformer faults. Techniques like PCA, Rough Set, and incremental granular ranking were employed to reduce the attributes in the DGA dataset. The study centered on two algorithms, SVM and BPNN, with PCA consistently enhancing prediction accuracy compared to other methods [10]

Most DGA datasets are highly imbalanced with the majority of transformers being in good condition. When data is imbalanced, it can cause problems by favouring the majority group and neglecting the minority which eventually results in the poor-performing model when it is fed with new and unseen data [11]. In their study, Tra et al. [12] employed an ADASYN to address the class imbalance in Dissolved Gas Analysis (DGA) datasets for power transformers. The study utilized five different datasets and they found that ADASYN significantly improved the performance of machine learning classifiers as compared to the original imbalanced dataset. Rajesh et al. [13] investigated the impact of data balancing on transformer DGA various fault classifications using five classifiers namely RF, Naive Bayes, LDA, QDA and MLP. They evaluated the effectiveness of five resampling techniques, including random oversampling (ROS), random undersampling (RUS), ROS combined with RUS, SMOTE, and ADASYN. Their results indicated that the ADASYN method outperformed the other techniques in improving classifier performance.

This study seeks to understand how an imbalanced dataset in Dissolved Gas Analysis (DGA) affects the performance of machine learning models in diagnosing transformer faults. It explores various resampling techniques to tackle this issue. Additionally, it aims to identify the best resampling methods and the most suitable classifier for a thorough evaluation, with a specific focus on fault diagnosis considering each transformer health condition class. This study is closely related with Wang et al. study [11], which primarily employed six resampling methods (SMOTE, RUS, ADASYN, B-SMOTE, SMOTE ENN, and cGAN) and evaluated them using three classifiers (SVM, DT, and KNN). However, our research significantly expands upon this foundation by investigating eight resampling techniques. This extension includes introducing RUS, SVM-SMOTE and novel additions to DGA power transformer research, namely, Cluster centroids and SMOTE TOMEK. Furthermore, we broadened the spectrum of classifiers to encompass SVM, Decision tree (DT), KNN, Random Forest (RF), and Logistic Regression (LR). Furthermore, While Wang et al. [11] dataset covered eight fault types, this study classifies transformers into three health conditions (good, normal, and poor), each linked to recommended actions. This research offers novel insights to improve fault detection in power transformers, contributing to enhanced infrastructure stability and operational reliability.

II. Research Methodology

A. Dataset Description

The DGA dataset utilized in this study is obtained from the Kaggle machine learning repository and comprises 470 records containing 15 attributes. To enhance the dataset, we introduced the 'Condition' attribute, derived from Kaggle's 'Health index' attribute. This new attribute categorizes transformer health into 'Good,' 'Fair,' or 'Poor.' The dataset displays class imbalance, with 9 'Good,' 41 'Fair,' and 420 'Poor' instances based on predefined numerical ranges. Table I offers a concise summary of Health Index (HI), transformer health status, and maintenance recommendations.

TABLE I
Transformer Health Index and Corresponding Condition

HI%	Condition	Requirements
70-100	Good	Normal maintenance
50-69	Fair	Increase diagnostic testing, possible remedial work, or replacement needed depending on criticality
0-49	Poor	Immediately assess risk; replace or rebuild based on assessment

B. Resampling Methods

This study investigates how data imbalance in DGA affects the diagnosis of transformer faults. Class imbalance emerges when instances appear unevenly distributed among classes [14]. To address this, resampling methods are employed in imbalanced learning, introducing bias for data balance. These methods ensure equilibrium, while it is feasible to learn from imbalanced data, the pursuit of balanced datasets enhances overall performance [14]. In our experiment, resampling methods were applied only to the training dataset to balance class numbers without introducing data leakage, thereby enabling a fair evaluation of the model's performance on new and unseen data. The training data was adjusted by increasing instances from the rarer classes (oversampling), decreasing instances from the more common ones (undersampling) and a combination of these two techniques. Brief description of each technique:

Random undersampling is a straightforward but somewhat naive method for addressing imbalanced datasets. It involves randomly removing instances of the majority class from the training set until a desired balance is reached [14]. However, it lacks control over which information from the majority class is discarded, potentially leading to the loss of important decision boundary details. Nevertheless, empirical evidence suggests that random undersampling is highly effective, even outperforming more complex methods in practical studies [14]. Random oversampling, similar in simplicity to random undersampling, entails duplicating minority class instances and adding them to the training data [14],[15]. While widely used, this approach can lead to overfitting, where the classifier becomes overly specialized in recognizing training examples, potentially causing poor performance on new data [14].

SMOTE, on the other hand, offers a more sophisticated approach to balancing training data in unbalanced datasets [15]. It generates new examples by blending existing ones, avoiding exact copies and mitigating overfitting to some extent [15]. The technique creates more minority samples by combining

two minority instances and one of their K-nearest neighbors [16]. This process, used in a recent effective study [17], helps increase minority class representation. ADASYN, an extension of SMOTE, overcomes class imbalance challenges by adaptively generating instances based on the density distribution between the minority class data and its k-nearest neighbors [18]. This adaptability increases instance count while reducing duplication in the minority class [19].

SMOTE-ENN combines SMOTE and edited nearest neighbor (ENN) techniques to achieve better class balance [11]. It creates synthetic samples using SMOTE and then filters out instances by checking their K-nearest neighbors [11]. If the observation's class doesn't match the majority class of its K-neighbors, both the observation and the K-neighbors are removed. This process repeats until each class reaches the desired number of instances. SVM-SMOTE focuses on generating new minority class samples near class boundaries through the application of the SVM model, which helps define clear class boundaries [20]. SMOTE-Tomek is an extension of SMOTE that combines it with Tomek's link removal to enhance class balance. It first generates synthetic data using SMOTE and then removes pairs of examples forming Tomek links, which often represent noise or points near the optimal decision boundary [14]. Cluster Centroids identifies clusters of majority class instances and reduces the majority class by keeping only the centroids of these clusters [14]. This approach aims to maintain the majority class's structure and representativeness while reducing its size [14]. These resampling techniques, as indicated by recent studies, offer valuable tools to address class imbalance in machine learning [14].

C. Experimental-Setup

The experimental design for this research study was designed to provide a comprehensive evaluation of how resampling techniques impact classifier performance in transformer fault diagnosis. In this study, Jupyter Notebook served as the platform for running experiments, with Python 3.7 as the programming language. Several critical considerations were taken into account. First, we ensured a fixed random state of 42, a standard practice in machine learning research to ensure the reproducibility of results. This approach allowed us to validate the impact of resampling techniques and assess the model's performance consistently across multiple runs. Additionally, we divided our dataset into two segments, allocating 80% to the training dataset and reserving 20% for testing. From the training set, a stratified five-fold cross-validation strategy was employed. During each fold, one subset was dedicated to validation, while the remaining four were used for training. This approach was instrumental in preventing overfitting and ensuring that our models could generalize effectively.

The resampling methods were only applied to the training sets within each fold, and the test set remained intact to preserve data independence. Maintaining the integrity of the test set prevented data leakage during cross-validation and assured a reliable evaluation of model performance. Our initial steps also involved fine-tuning the hyperparameters of our selected classifiers, including Support Vector Machine (SVM), Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Decision Tree. A systematic hyperparameter optimization process tailored to each classifier was executed. For Logistic Regression, we leveraged gradient descent methods for model optimization, while a grid-search approach was employed for SVM, Random Forest, KNN, and Decision Tree to explore various hyperparameter combinations. This systematic approach allowed us to fine-tune the models, adapting them to the specific characteristics and complexities of our dataset. The primary aim of this hyperparameter optimization was to maximize the performance of each classifier in making predictions.

D. Evaluation Metrics

Confusion matrices were used to visually compare the predicted values from the algorithms with the actual dataset values. These matrices displayed true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). In our experiment, we evaluated the model using the area under the curve-receiver operating characteristic curve (AUC-ROC), precision, recall, and F1 score. These metrics provided a comprehensive assessment of the model's performance in terms of sensitivity, specificity, and the balance between precision and recall. To assess the statistical significance of the obtained results, we conducted a one-way ANOVA test. This involved formulating null and alternative hypotheses statements to determine whether significant differences existed among the experimental groups.

III. Results and Discussion

This research paper aims to investigate the impact of an imbalanced Dissolved Gas Analysis (DGA) dataset on the performance of machine learning classifiers for transformer fault diagnosis. It addresses this issue through the application of diverse resampling techniques. Furthermore, the study aims to identify the most effective resampling methods and select the optimal classifier for conducting a comprehensive assessment of these resampling techniques, with a focus on predictive performance for each transformer condition class. In Table II, we present the results of different resampling methods applied to the task of transformer fault diagnosis, considering overall performance in predictive performance across all classes.

Table II shows that in the imbalanced dataset, all classifiers, except Random Forest, had ROC-AUC values ranging from 0.7401 to 0.7863. Applying resampling techniques generally improved ROC-AUC scores. For instance, SVM's ROC-AUC increased from 0.7401 to 0.8370 with ROS. Random Forest achieved the highest ROC-AUC of 0.9583 with ADASYN. Decision Tree showed improvements in ROC-AUC with various resampling methods, except when using Cluster Centroids, where it dropped to 0.6983 from the original value of 0.7614 in the imbalanced dataset. KNN's ROC-AUC decreased with all resampling methods, except when balanced with SVM-SMOTE, achieving 0.8147, a 0.0352 increase from the imbalanced dataset. The logistic Regression classifier improved with all resampling methods, with ROS achieving the highest ROC-AUC of 0.8072, a 0.052 increase from the imbalanced dataset. Table II indicated that ROS was the most effective in terms of ROC-AUC improvement among all classifiers, except for Random

Forest and KNN, where it ranked as the second-best performer. SVM-SMOTE and ADASYN also displayed good performance, with SVM-SMOTE standing out as the second-best performer. Enhanced ROC-AUC values in transformer fault diagnosis are pivotal, signifying precise classification of "Good," "Fair," and "Poor" transformer health conditions. This accuracy optimizes maintenance decisions, reducing costs and enhancing power infrastructure stability.

Table II highlights the significant impact of dataset imbalance on precision, recall, and F1 scores. In the imbalanced dataset, all classifiers consistently achieved precision, recall and F1 score exceeding 90%, primarily due to their proficiency in identifying the majority class, particularly the 'Poor' class. Post-resampling, precision and recall remained high, with only a subtle decrease. SMOTE ENN, combined with RF, achieved the highest precision and recall at 96.72% and 96.81%, respectively. For the SVM classifier, SMOTE and SMOTE-TOMEK attained the highest precision at 92.78%, while the imbalanced dataset outperformed resampling methods in terms of recall and F1 score, reaching 93.62% and 92.78%, respectively. LR and KNN both achieved their highest precision, recall and F1 scores using the imbalanced dataset, consistently scoring over 90% in all these key metrics. In the case of the Decision Tree classifier, the highest precision (93.10%) was achieved with the SVM-SMOTE method, while the highest recall (92.55%) was obtained with the imbalanced dataset, ROS, SVM-SMOTE, and SMOTE-TOMEK, all yielding identical scores.

RUS significantly decreased recall for all classifiers to below 74%, except for Random Forest. This decrease also impacted the F1 score, dropping it below 80%. For instance, Logistic Regression's recall and F1 score dropped to 63.83% and 74.12%, respectively, after implementing RUS. The cluster centroids method similarly decreased Decision Tree recall to 56.38%. To gain a deeper understanding of each resampling method's impact on a model's ability to correctly identify and capture true positives, we conducted a detailed class-specific analysis of precision and recall, as presented in Table III.

The results in Table II did not identify a single superior resampling method that consistently outperforms others across all classifiers and metrics, it revealed that different resampling techniques excel in varying circumstances. For example, SMOTE-ENN and ADASYN yielded the best performance when combined with the Random Forest classifier across all metrics (Table II). However, SMOTE-ENN and ADASYN performance with the Decision Tree classifier declined, while SVM-SMOTE method enhanced decision tree performance. These results aligned with the findings of Wang et al. [11] which emphasized that the performance of a specific resampling method differs across various classifiers. Researchers and practitioners in transformer fault diagnosis should carefully weigh the trade-offs between metrics and assess resampling's influence on each classifier's performance.

The results presented in Table II consistently indicated that the random forest classifier outperformed other classifiers across various metrics in most cases. Therefore, this study selected the random forest classifier as the primary choice for a more

TABLE II
COMPARISON OF RESAMPLING METHODS FOR DIFFERENT MODELS

Method	AUC	Precision	Recall	F1 Score
SVM				
Imbalanced	0.7401	0.9217	**0.9362**	**0.9278**
ROS	**0.8370**	0.9149	0.8723	0.8865
RUS	0.7553	0.8764	0.6809	0.7644
SMOTE	0.7945	**0.9278**	0.8830	0.8851
ADASYN	0.8061	0.8956	0.9043	0.8965
SVM-SMOTE	0.7607	0.8924	0.9043	0.8965
SMOTE-TOMEK	0.7920	**0.9278**	0.8830	0.8851
SMOTE-ENN	0.8107	0.8689	0.8191	0.8108
Cluster Centroids	0.7475	0.9255	0.8830	0.8972
LR				
Imbalanced	0.7552	**0.9336**	**0.9468**	**0.9373**
ROS	**0.8072**	0.8887	0.7766	0.8287
RUS	0.7896	0.8854	0.6383	0.7412
SMOTE	0.7881	0.9009	0.8085	0.8520
ADASYN	0.8049	0.9003	0.8085	0.8518
SVM-SMOTE	0.7673	0.9274	0.9149	0.9191
SMOTE-TOMEK	0.7898	0.9009	0.8085	0.8520
SMOTE-ENN	0.8011	0.8779	0.7553	0.8103
Cluster Centroids	0.7933	0.9323	0.7979	0.8554
RF				
Imbalanced	0.9572	0.9217	0.9362	0.9278
ROS	0.9211	0.9217	0.9362	0.9278
RUS	0.9115	0.9268	0.8723	0.8934
SMOTE	0.9557	0.9437	0.9468	0.9426
ADASYN	**0.9583**	0.9604	0.9574	0.9570
SVM-SMOTE	0.9574	0.9082	0.9255	0.9144
SMOTE-TOMEK	0.9171	0.9359	0.9362	0.9304
SMOTE-ENN	0.9373	**0.9672**	**0.9681**	**0.9673**
Cluster Centroids	0.8676	0.9190	0.8191	0.8640
KNN				
Imbalanced	0.7863	**0.9086**	**0.9255**	**0.9089**
ROS	0.7795	0.8664	0.8617	0.8639
RUS	0.7108	0.8311	0.6596	0.7285
SMOTE	0.7672	0.8601	0.8404	0.8502
ADASYN	0.7785	0.8915	0.8617	0.8756
SVM-SMOTE	**0.8147**	0.9032	0.9149	0.9050
SMOTE-TOMEK	0.7672	0.8601	0.8404	0.8502
SMOTE-ENN	0.7397	0.8791	0.8298	0.8504
Cluster Centroids	0.7554	0.8745	0.8191	0.8440
DT				
Imbalanced	0.7614	0.9227	**0.9255**	0.9156
ROS	**0.8410**	0.9278	**0.9255**	**0.9245**
RUS	0.7958	0.9041	0.7340	0.8044
SMOTE	0.8356	0.9250	0.9150	0.9140
ADASYN	0.8009	0.9179	0.9255	0.9194
SVM-SMOTE	**0.8410**	**0.9310**	**0.9255**	0.9244
SMOTE-TOMEK	0.8009	0.9227	**0.9255**	0.9211
SMOTE-ENN	0.8141	0.9002	0.8830	0.8890
Cluster Centroids	0.6983	0.9236	0.5638	0.6999

detailed investigation into its performance regarding precision, recall and F1 score for each transformer class. Precision and recall play a pivotal role in understanding the model's ability to identify transformers requiring maintenance while minimizing the risk of false alarms. Table III presents a comprehensive overview of the random forest model's performance under various resampling methods, each tailored to predict three distinct health states of power transformers: "Fair," "Good," and "Poor."

Balancing precision and recall is crucial, especially for transformers in a "Fair" health state. Table III reveals that Random Forest, combined with the cluster centroids method, achieved a precision rate of 67% and a recall of 40% in the fair class. This means the classifier is 67% accurate in identifying trans-

formers in fair health, but 40% recall indicates that the model frequently misses identifying transformers that are actually in a fair health state. These results are poorer than the original imbalanced dataset, which had 75% precision and 60% recall. RUS also resulted in lower precision and recall compared to the original dataset. SMOTE Tomek improved precision to 83% but compromised recall to 50%. This indicates the classifier is good at correctly identifying transformers in the fair class, reducing unnecessary maintenance costs and downtime caused by false alarms, but it sometimes misses those truly in fair health, which might result in overlooking potential issues. The Random Forest classifier excelled when paired with the SMOTE ENN method, achieving 89% precision and an 80% recall rate. This method effectively balances precision and recall for transformers in a "Fair" health state, resulting in an F1 score of 84

In predicting the health status of "Good" transformers, both the SMOTE and SMOTE ENN methods performed remarkably well across all three evaluation metrics, achieving 100% precision, recall, and F1 scores as shown in Table III. This was a remarkable improvement from the original imbalanced dataset, which scored 0% across all metrics. In contrast, methods like ROS, SVM-SMOTE, and Cluster Centroids scored 0% in all three metrics indicating poor performance. The "Good" class was underrepresented in the original dataset, indicating that there was limited testing data available for this class which explains the challenges faced by some of the methods.

When predicting the health status of transformers classified as "Poor", all the resampling methods consistently boosted the performance of the random forest classifier, resulting in high precision, recall, and F1 scores for this class. Among these methods, the lowest precision of 94% was obtained with SVM-SMOTE, and the lowest recall of 88% was obtained when using the cluster centroids method. This indicates that the model excels in identifying transformers in need of maintenance. The 'Poor' health state category encompasses the majority of power transformers, and, as a consequence, all the applied methods achieved exceptional results. ADASYN and SMOTE ENN emerged as the top-performing techniques, boasting precision, recall, and F1 scores above 97%.

Across all three classes, SMOTE ENN consistently outperformed the rest, achieving the best performance across all three metrics. In the realm of predictive maintenance for power transformers, these findings provide valuable insights. Power utilities face the challenge of making strategic decisions tailored to their particular goals and budget constraints. If the primary concern is avoiding the risk of overlooking potential issues, methods that prioritize high recall may be favored. However, if the goal is to optimize resource allocation and prevent unnecessary maintenance actions, methods that strike a balance between precision and recall, like SMOTE ENN, could be a more suitable choice.

The performance of resampling methods, specifically SMOTE ENN, SMOTE, and ADASYN, in predicting the minority class "good", as demonstrated in Table III, aligns with the findings of Wang et al. [11], who aimed to address the class imbalance in transformer fault conditions. Wang et al. [11] primarily focused on predicting minority classes PD (Partial Discharge) and DT (Thermal fault and Discharge). Their results indicated that ADASYN, SMOTE, and SMOTE ENN were the top methods in enhancing precision, recall, and F1 scores for these classes which agrees with our results. While Wang et al. [11] concentrated on a dataset with gas concentrations alone, our study extended diagnostic parameters by incorporating features like power factor, IFT, furan content, and Dielectric Rigidity. These additional features offer a more comprehensive perspective on transformer health, thereby enhancing performance in transformer fault diagnostics. This underscores the significance of diverse features, equipping power utilities with vital insights for well-informed decisions in predictive maintenance and resource allocation.

TABLE III
PERFORMANCE METRICS OF RANDOM FOREST CLASSIFIER BY RESAMPLING METHOD

Resampling Method	Precision	Recall	F1-Score
Fair Class			
Original Scaled Data	0.75	0.60	0.67
Random Oversampling	0.75	0.60	0.67
Random Undersampling	0.71	0.50	0.59
SMOTE	0.86	0.60	0.71
ADASYN	0.88	0.70	0.78
SVM-SMOTE	0.71	0.50	0.59
SMOTE TOMEK	0.83	0.50	0.62
SMOTE ENN	**0.89**	**0.80**	**0.84**
CLUSTER CENTROIDS	0.67	0.40	0.50
Good Class			
Original Scaled Data	0.00	0.00	0.00
Random Oversampling	0.00	0.00	0.00
Random Undersampling	0.12	**1.00**	0.22
SMOTE	**1.00**	**1.00**	**1.00**
ADASYN	0.50	**1.00**	0.67
SVM-SMOTE	0.00	0.00	0.00
SMOTE TOMEK	0.50	**1.00**	0.67
SMOTE ENN	**1.00**	**1.00**	**1.00**
CLUSTER CENTROIDS	0.00	0.00	0.00
Poor Class			
Original Scaled Data	0.95	**0.99**	0.97
Random Oversampling	0.95	**0.99**	0.97
Random Undersampling	0.96	0.92	0.94
SMOTE	0.95	**0.99**	0.97
ADASYN	**0.98**	**0.99**	**0.98**
SVM-SMOTE	0.94	**0.99**	0.96
SMOTE TOMEK	0.95	**0.99**	0.97
SMOTE ENN	**0.98**	**0.99**	**0.98**
CLUSTER CENTROIDS	0.96	0.88	0.92

IV. CONCLUSION

This study investigated the impact of imbalanced DGA datasets on machine-learning classifiers for transformer fault diagnosis. The study aimed to identify effective resampling techniques and the best-performing classifier for evaluating various transformer condition classes. The study employed key evaluation metrics like precision, recall, F1-score, and ROC-AUC. Resampling techniques generally improved all classifiers performance, except for RUS and cluster centroids, which significantly lowered ROC-AUC and overall recall. No single resampling method consistently outperformed others, with effectiveness varying across classifiers and metrics. Random forest outperformed other classifiers and it was chosen for a detailed analysis of precision, recall and F1 score for each transformer class. Pairing

Random Forest with SMOTE ENN for the 'Fair' transformer class achieved the best results, with 89% precision, 80% recall, and an 84% F1-score. For 'Good' transformers class, SMOTE and SMOTE ENN excelled with 100% precision, recall, and F1-scores. 'Poor' transformers benefited from various resampling methods paired with Random Forest, resulting in high precision, recall, and F1-scores above 94%. SMOTE ENN stood out as the top-performing method across all resampling techniques for each transformer class, securing precision and recall rates consistently exceeding 88% and 80%, respectively. Consequently, it achieved F1-scores surpassing 83% for every class. The research findings offer power utilities the knowledge needed for informed decisions on predictive maintenance, resource allocation, and cost reduction while minimizing downtime. However, this study's limitation lies in its focus on a specific dataset, possibly limiting generalizability. Future research could build upon the research findings of this work by delving into hybrid resampling techniques, combining multiple methods to address the challenges of imbalanced datasets. This approach could lead to even more effective strategies for transformer fault diagnosis. By addressing these research directions, we can advance the field and provide more robust solutions for predictive maintenance in the context of transformer health monitoring.

References

[1] Y. Wang, S. Gong, and S. Grzybowski, "Reliability evaluation method for oil–paper insulation in power transformers," *Energies*, vol. 4, no. 9, pp. 1362–1375, 2011.

[2] D. Arvind, S. Khushdeep, and K. Deepak, "Condition monitoring of power transformer: A review," in *2008 IEEE/PES Transmission and Distribution Conference and Exposition*, IEEE, 2008, pp. 1–6.

[3] L. Cheng and T. Yu, "Dissolved gas analysis principle-based intelligent approaches to fault diagnosis and decision making for large oil-immersed power transformers: A survey," *Energies*, vol. 11, no. 4, p. 913, 2018.

[4] X. Chen, H. Cui, and L. Luo, "Fault diagnosis of transformer based on random forest," in *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, IEEE, vol. 1, 2011, pp. 132–134.

[5] Y. Zhang, Y. Tang, Y. Liu, and Z. Liang, "Fault diagnosis of transformer using artificial intelligence: A review," *Frontiers in Energy Research*, vol. 10, p. 1 006 474, 2022.

[6] C. Peng, L.-s. Tong, and G.-n. Wu, "Technical achievements of on-line monitoring of dissolved gas in transformer oil," *Electric Power Automation Equipment*, vol. 24, no. 11, 2004.

[7] N. Lelekakis, D. Martin, W. Guo, and J. Wijaya, "Comparison of dissolved gas-in-oil analysis methods using a dissolved gas-in-oil standard," *IEEE Electrical Insulation Magazine*, vol. 27, no. 5, pp. 29–35, 2011.

[8] T. Kari, W. Gao, D. Zhao, *et al.*, "Hybrid feature selection approach for power transformer fault diagnosis based on support vector machine and genetic algorithm," *IET Generation, Transmission & Distribution*, vol. 12, no. 21, pp. 5672–5680, 2018.

[9] H. MehdipourPicha, R. Bo, H. Chen, M. M. Rana, J. Huang, and F. Hu, "Transformer fault diagnosis using deep neural network," in *2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia)*, IEEE, 2019, pp. 4241–4245.

[10] J. T. Maumela, "Condition monitoring of transformer bushings using computational intelligence," *arXiv preprint arXiv:2204.10193*, 2022.

[11] L. Wang, T. Littler, and X. Liu, "Hybrid ai model for power transformer assessment using imbalanced dga datasets," *IET Renewable Power Generation*, 2023.

[12] V. Tra, B.-P. Duong, and J.-M. Kim, "Improving diagnostic performance of a power transformer using an adaptive over-sampling method for imbalanced data," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 26, no. 4, pp. 1325–1333, 2019.

[13] K. N. Rajesh, U. M. Rao, I. Fofana, P. Rozga, and A. Paramane, "Influence of data balancing on transformer dga fault classification with machine learning algorithms," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 30, no. 1, pp. 385–392, 2022.

[14] T. M. Alam, K. Shaukat, I. A. Hameed, *et al.*, "An investigation of credit card default prediction in the imbalanced datasets," *IEEE Access*, vol. 8, pp. 201 173–201 198, 2020.

[15] M. Hanafy and R. Ming, "Using machine learning models to compare various resampling methods in predicting insurance fraud," *J. Theor. Appl. Inf. Technol*, vol. 99, no. 12, pp. 2819–2833, 2021.

[16] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in *Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13*, Springer, 2009, pp. 475–482.

[17] D. Scrutinio, C. Ricciardi, L. Donisi, *et al.*, "Machine learning to predict mortality after rehabilitation among patients with severe stroke," *Scientific reports*, vol. 10, no. 1, p. 20 127, 2020.

[18] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, Ieee, 2008, pp. 1322–1328.

[19] Y. Pristyanto, A. F. Nugraha, I. Pratama, A. Dahlan, and L. A. Wirasakti, "Dual approach to handling imbalanced class in datasets using oversampling and ensemble learning techniques," in *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, IEEE, 2021, pp. 1–7.

[20] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "Svms modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 281–288, 2008.