ISSN: 0067-2904

Analysis of Transformer Health Index Using Statistical and Machine Learning Techniques

Sushma Sagar Emme[1*], Pratapa Raju Moola[2]
[1] *Computer Science and Multi Media, Lincoln University College, Petaling Jaya, Malaysia*
[2] *Engineering Department, University of Technology and Applied Sciences, Ibra, Oman*

Received: 1/1/2024 Accepted: 2/2/2025 Published: 30/1/2026

Abstract

Data Science and Machine Learning have been playing a major role in assessing, predicting, and maintaining the health of power transformers using data analysis. This paper focuses on leveraging data science techniques to analyze and interpret Dissolved Gas Analysis (DGA) datasets associated with power transformers to predict Health Index (HI). The Exploratory Data Analysis (EDA) involving the correlation matrix and heat maps showed the correlation among all the features and indicated that the dataset considered is not balanced hence, the data balancing technique of oversampling is employed to balance the data. Principal Component Analysis (PCA) is used to estimate the principal components of the data, helping in selecting the features which are most useful in the prediction. Classifiers, namely Support Vector Machine (SVM), Random Forest (RF), XGBoost, and k Nearest Neighbors (KNN) are employed on both the balanced data as well as the imbalanced data and the results are compared. RF classifier outperformed all the other classifiers with an accuracy of 96.9%.

Keywords: Dissolved Gas Analysis, Exploratory Data Analysis, Support Vector Machine, Random Forest, XGBoost, k-Nearest Neighbors.

1. Introduction

Power Transformers (PTs) are critical components of the Power System Network (PSN), responsible for regulating voltage levels from generation to the load center [1]. Effective maintenance of PTs is essential for ensuring reliable operations. Maintenance strategies fall into three main categories. The first, run-to-failure, involves acting only after a failure occurs, often resulting in costly repairs and significant downtime. The second category, preventive maintenance, is based on scheduled interventions to address potential issues before they arise. The third and most cost-effective approach is predictive maintenance, which focuses on assessing equipment health and detecting failures in advance. Predictive maintenance offers numerous advantages, including improved system reliability, reduced unexpected outages, and eliminating unnecessary maintenance operations, thereby lowering overall costs. While both predictive and preventive maintenance strategies enhance reliability, predictive maintenance stands out for its superior cost efficiency. Current research and practical applications primarily emphasize predictive maintenance for large PTs.

*Email: sushma.phdscholar@lincoln.edu.my

Transformer failures often result in extensive outages and blackouts, which eventually have an effect on transmission [2] and distribution systems [3]. Hence, it is quite important to design and implement predictive maintenance mechanisms for the uninterrupted operation of the PSN. As weather conditions have a significant role in transformer failures, it is required to instigate weather monitoring systems in transmission and distributed systems to track the conditions [4]. The other reasons could be cooling system failures, overloading, over voltages, and over currents [5]. Therefore, power suppliers prioritize health assessment of transformers for effective and healthy operation.

2. Related Work

The study outlined in reference [6] introduces an expert system designed for conducting insulation diagnostics, while researchers in [7] delve into the current state and recent advancements in various approaches to diagnosing PTs. The objective of [8] is to detail, examine, and elucidate existing Physicochemical diagnostic methods employed for assessing the insulation condition in aging transformers. Developing fault prediction models often involves data mining, a multifaceted approach combining computer science and statistics to extract concealed, previously unknown, and potentially significant information from extensive databases [9]. However, the prediction of the transformer's health in the context of predictive maintenance is significantly contributed to by DGA analysis.

Transformer oil DGA presented in [10] is a quite useful aspect in transformer health assessment/ index. The techniques employed to evaluate transformer health rely on DGA. DGA examines the concentration of particular gases about the insulation oil of transformers. The concentration levels of dissolved gases serve as indicators of the insulation's decomposition. Gases commonly analyzed in DGA encompass hydrogen, carbon monoxide, methane, ethane, acetylene, ethylene, and carbon dioxide [11].

Technologies based on Artificial Intelligence (AI) are also employed to study extensive data and extract knowledge from the available data [12]. The primary approaches emphasized in [13] for constructing predictive models are classification and regression. In classification, each item in a dataset is assigned to predefined classes or groups [14]. Machine Learning (ML) facilitates computers learning from experience, analogous to natural human learning processes. ML techniques do not use any mathematical model to analyze the data but utilize computational methods to glean information.

In the study presented in [15], data is obtained from liquid insulation parameters, including DGA data, water content, furan, and Interfacial Tension (IFT). The goal is to analyze the transformer's health conditions and evaluate the transformer's remaining life-span based on operating temperature. The predominant focus in current research is on evaluating the HI of transformers by scrutinizing the deterioration of oil using DGA. A deep generative model-based framework integrates an MLP and logistic regression to classify transformer health into eight categories. Tested on 18,848 samples from 608 transformers, it achieved 99% accuracy, surpassing existing methods. The framework compresses data and incorporates expert insights, enabling precise and reliable diagnostics for grid operations[16].

In 2023, a Multimodal Mutual Neural Network (MMNN) was introduced for power transformer health assessment by combining dissolved gas data (DGD) and infrared imaging. It uses a 1D CNN for DGD analysis, a DRSE network for infrared features, and a ProbSparse self-attention mechanism to integrate multimodal data. This reliable, accurate approach is ideal for real-time transformer health monitoring in substations[17].

The paper presents a transformer asset management model using online DGA data and CNNs for fault diagnostics and life assessment. It classifies faults like partial discharge and thermal issues with 87% accuracy based on 1,083 samples. Detecting multi-label faults and

estimating insulation degradation through CO_2 and CO levels enables real-time condition monitoring, eliminating the need for offline tests[18]. Additionally, support vector machines, and deep belief networks have also been utilized, including the extreme learning machine (ELM) [19-21].

A Deep Learning Neural Network (DLNN) was developed for transformer health evaluation, combining an Echo State Network (ESN) for data augmentation and an Improved Deep Residual Shrinkage Network (IDRSN) with 1D CNN for feature extraction. Using a Concat and Softmax layer, the ESN-DRSN-CW model achieved 0.82% better accuracy than DRSN-CW, offering reliable diagnostics despite higher computational time. Validation on real DGA datasets confirmed its effectiveness[22]. As presented in [23], online condition monitoring for High Impedance (HI) using DGA interpretation, utilizing the C4.5 algorithm employing the decision tree model for transformers is done. The algorithm leverages ML techniques such as WEKA, and Orange to achieve optimal learning outcomes. The results obtained through this approach are compared with those derived from other models.
Another study by Sarajcev et al. 2018 [24], presented a Bayesian multinomial logistic regression model for estimating transformer HI. However, it neglects effects related to the inherent ordering of categories, a consideration applicable to applications of Artificial Neural Networks (ANNs) and certain other Machine Learning (ML) models utilized for classification tasks. Furthermore, the proposed model allows for implementing online learning/monitoring, offering potential benefits for health assessment. The study, which Leauprasert discussed in 2020 [25], presented the utilization of regression models to assess the %HI in terms of percentage for estimating the condition of the transformer. It highlights limitations such as the lack of a proper elucidation of model parameters and the study notes that HI values obtained from regression models may fall outside the intended range.

The study established by Patil et al. in 2020 [26] presented an online health monitoring strategy specifically in the case of 33 kV steel-mill transformers. It utilizes fuzzy models for computing the HI and estimating the life left for the transformer in a fuzzy mode. The research suggests the potential for further exploration, particularly extending the approach to higher voltage transformers. The proposal includes developing a generalized fuzzy model applicable to various transformers, including generation and transmission. Although there is considerable ML based HI predictions based health assessments performed with higher accuracy, application of data science techniques such as ‘EDA’ and ‘IMB learn’ is not emphasized. The outcomes of using the above techniques are clearly mentioned. Towards the end, HI based Health Assessment is presented in terms of encoded categorical indicators such as 0, 1, 2, 3, 4, which stand for very poor, poor, fair, good, and very good, respectively.

3. Methodology

The work presented in this article follows a specific methodology, and the different stages involved are detailed in this section. The entire implementation is based on the DGA carried out in specific methods. Renowned methods of DGA are presented in this section.

a. Dissolved Gas Analysis (DGA)

DGA is a diagnostic technique employed to evaluate the condition of PTs. This technique involves analyzing gases of the insulating oil pertaining to transformers. Identifying and measuring distinct gas concentrations in oil offers valuable insights into internal problems like overheating, electrical discharges, and insulation deterioration. DGA serves as an effective tool for assessing the health of PTs to prevent potential issues and ensure optimal performance. There are several methods to run DGA for the Transformers. Three of them are described in this section.

i. Key Gas Method

The Key Gas method [27-28] is a diagnostic technique that measures gases emitted from the insulating oil following faults, particularly when elevated temperatures occur in PTs. Unlike traditional methods, this approach focuses on individual gas measurements rather than gas ratios and identifies important gases known as key gases. The primary cause of faults lies in the stress- induced breakdown of oil or cellulose molecules, leading to the formation of gases. These gases, including H_2, CH_4, C_2H_2, C_2H_4, C_2H_6, CO, and CO_2, dissolve either fully or partially in the oil under various thermal as well as electrical stress conditions due to faulty currents in transformers. The key gas approach categorizes Hydrocarbon and hydrogen, Carbon oxides, and non-fault gases into three groups.

ii. Dornenburg Ratio Method

Thermal faults, corona discharge, and arcing are identified by the Dornenburg Ratio method [29] by using gas concentration proportions like C_2H_2/C_2H_4, C_2H_2/CH_4, C_2H_4/C_2H_6, CH4/H2, and principles of thermal degradation. While specified in IEEE Standard of C57.104-2008 [27], this method might yield too many no-interpretation results.

iii. Rogers Ratio Method

The Dornenberg ratio technique is outperformed by the Rogers ratio approach [30] for diagnosing thermal faults in oil-insulated transformers. It examines gas ratios such as CH4/H2, C2H6/CH4, C2H4/C2H6, and C2H2/C2H4 by a straightforward coding scheme rooted in predefined ratio ranges. Integrated into IEEE Standard C57.104-2008 [27], the method effectively identifies conditions like normal aging, partial discharge, and various electrical and thermal faults of transformers. Limitations, such as inconsistencies between ratio values and diagnostic codes and excluding dissolved gases below normal concentrations, often lead to data misinterpretation.

b. Flow chart of the implementation

Different stages of methodology, as described in the flow chart presented in Figure 1, are discussed in the section below. There are five stages in this implementation.

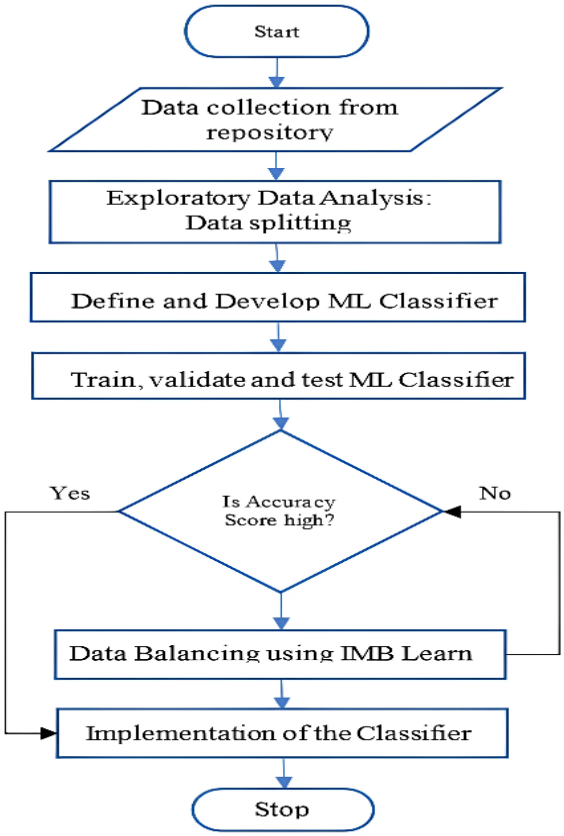

Figure 1: Flow chart of ML classifier Implementation using DGA data

Stage 1: Data acquisition

Data collection is the first stage of the ML implementations. In this context, DGA data of transformer oil of several transformers considered from [31] is used for the implementation.

Table 1: Transformer oil DGA data

S. No	Hydrogen	Oxygen	Nitrogen	Methane	CO	CO_2	Ethylene	Ethane	Acetylene	DBDS	Power factor	Interfacial V	Dielectric rigidity	Water content	HI
0	2845	5860	27842	7406	32	1344	16684	5467	7	19.0	1.00	45	55	0	95.2
1	12886	61	25041	877	83	864	4	305	0	45.0	1.00	45	55	0	85.5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
468	15	169	50600	5	77	532	0	72	0	0.0	1.21	33	54	11	13.4
469	15	308	39700	3	64	581	5	27	0	0.0	1.00	32	60	18	13.4

The data has 470 samples of 15 features, and they are Hydrogen, Oxygen, Nitrogen, Methane, CO, CO_2, Ethylene, Ethane, Acetylene, DBDS, Power Factor, Interfacial V, Die Electric Rigidity, Water Content and HI, these are listed in Figure 2.

Stage 2: Application of Data Science Techniques

Data science is a multidisciplinary field that employs scientific methods and algorithms to gain insights and knowledge from structured and unstructured data. In electrical engineering, data science is used to decipher and enhance various facets of the field. EDA is an effective tool in Data Science to preprocess the data and make it compatible with training. Preprocessing includes replacing null values, dropping the duplicate values, checking and replacing missing values, finding correlations between predictors and response variables (Heat maps), plotting the scatter plots for checking data distribution, and encoding categorical variables. As part of EDA, the correlation between the predictors and response variable is checked to quantify the variables for better training of the ML. The results are presented in the results section.

Stage 3: Defining the ML classifiers to assess Transformers' health using DGA data

Applying data science techniques empowers electrical engineers to make well-informed decisions, address challenges, and contribute to the evolution of smart grids and sustainable energy solutions. This paper explains the training and testing of four different classifiers to ascertain the condition of the transformer concerning HI predictions. The theoretical background of these classifiers is presented in this section.

a. Random Forest ML algorithm

Ho (1995) initially proposed the concept of the random-subspace method, but Breiman (2001) expanded this concept to include random forest (RF). This model represents an algorithm is based on ensemble tree-based learning, where predictions are averaged across multiple individual trees. These trees are constructed on bootstrap samples of the original dataset, a technique known as bootstrap aggregating or bagging, which effectively mitigates overfitting. While individual decision trees offer ease of interpretability. RF consistently delivers an accurate estimation of the error rate compared to decision trees. Notably, mathematical proof by Breiman (2001) demonstrates that the error rate always converges as

the number of trees in the random forest increases [32].

b. KNN ML algorithm

The k Nearest-Neighbors (kNN) method is a classification approach that is distribution free. It is simple and effective. To classify a data sample x, the algorithm selects its k nearest neighbors and forms a neighborhood around x. The classification for x is established by majority voting among the data samples in this neighborhood. The successful application of kNN is based on an appropriate value for k. Numerous methods are used for selecting the k, one being running the algorithm multiple times using a different k value every time and picking up the k value that gives optimal performance [33].

c. SVM ML algorithm

To understand SVM, two components called the hypothesis spaces and the corresponding loss functions must be understood. The conventional perspective on SVM is to identify an optimal hyperplane for solving the issue. The basic SVM structuring is linear, and the hyperplane resides in the space of the input data t. In their more general crafting, SVM identifies a hyperplane distinct from the input data t. This hyperplane exists in a feature space generated by a kernel K. Through the kernel K, the hypothesis space is characterized as a collection of hyperplanes in the feature space induced by K. This perspective is interpreted as a collection of functions in Reproducing Kernel Hilbert Space (RKHS) proposed by Wahba, (1990), Vapnik (1998) [34].

d. XGBoost ML algorithm

XGBOOST (Extreme Gradient Boosting) represents a highly efficient and scalable implementation of the Gradient Boosting Machine (GBM), which has emerged as a formidable tool in the realm of artificial intelligence [35]. XGBoost stands out as a competitive choice because of its superior prediction accuracy. There are several advantages: Firstly, in XGBoost, multithreading parallel computing is invoked automatically, which helps predict transient stability in the actual power grid. Subsequently, adding a regularization term to XGBoost enhances its generalization ability and addresses the problem of decision trees prone to overfitting. Lastly, XGBoost, a tree structure model, eliminates the need to normalize data collected by Phasor Measurement Units (PMU) in power systems.

Stage 4: Define and develop ML based Classifiers

As described in the paragraph above, four different classifiers are considered in this article, where RF, KNN and SVM classifiers are called from 'SCIKIT Learn' python library as functions, and the data is fit. Secondly, the XGBoost classifier is imported from the XGBoost python package, and the data is fit.

Stage 5: Training and Testing of Classifiers

The training is executed under two categories: one is with balancing the data, and the other is without balancing. Balancing is a process carried out to distribute the data more evenly so that each class has enough sample points. In this article, the 'IMB learn' library is used to balance the data.

4. Results: EDA, Training, Testing performed

First, the outcomes of the EDA process, such as Correlation Matrix, Heat map, and Statistics of the data are presented to examine the nature of the data and how well it suits the application considered. Correlation Matrix Statistics are presented in Tables 2 and 3. The data undergoes initial scrutiny through a heat map, enabling the analysis of relationships between different predictors and the target variable.

Table 2: Correlation Matrix of DGA data

S. No	Hydrogen	Oxygen	Nitrogen	Methane	CO	CO_2	Ethylene	Ethane	Acetylene	DBDS	Power factor	Interfacial V	Dielectric rigidity	Water content	HI
Hydrogen	1.00	-0.05	-0.11	0.63	-0.02	0.01	0.44	0.48	0.35	-0.04	0.22	0.10	0.05	-0.07	0.37
Oxygen	-0.05	1.00	0.09	-0.03	-0.04	-0.04	-0.01	-0.07	0.20	-0.03	-0.07	0.21	-0.12	-0.14	0.12
Nitrogen	-0.11	0.09	1.00	-0.10	0.02	0.13	-0.06	-0.05	-0.01	0.15	0.08	-0.07	-0.19	0.00	0.08
Methane	0.63	-0.03	-0.10	1.00	-0.00	0.06	0.80	0.91	0.23	-0.02	0.07	0.11	0.02	-0.03	0.36
CO	-0.02	-0.04	0.02	-0.00	1.00	0.55	-0.02	-0.09	-0.01	0.05	0.10	0.14	-0.04	-0.03	0.11
CO_2	0.02	-0.04	0.13	0.06	0.55	1.00	0.03	-0.00	-0.01	0.08	0.30	0.03	-0.07	0.07	0.16
Ethylene	0.44	-0.01	-0.06	0.80	-0.02	0.03	1.00	0.75	0.25	-0.02	0.01	0.10	0.02	-0.00	0.27
Ethane	0.48	-0.07	-0.05	0.91	-0.09	-0.01	0.75	1.00	0.20	-0.04	0.04	0.00	0.02	0.02	0.23
Acetylene	0.35	0.20	-0.01	0.23	-0.01	-0.01	0.25	0.20	1.00	-0.05	-0.01	0.13	-0.01	-0.0	0.24
DBDS	-0.04	-0.03	0.15	-0.02	0.05	0.08	-0.02	-0.04	-0.05	1.00	-0.06	0.18	-0.06	-0.20	0.46
Power factor	0.22	-0.01	0.08	0.07	0.10	0.30	0.01	0.04	-0.01	-0.06	1.00	-0.20	-0.01	0.08	0.09

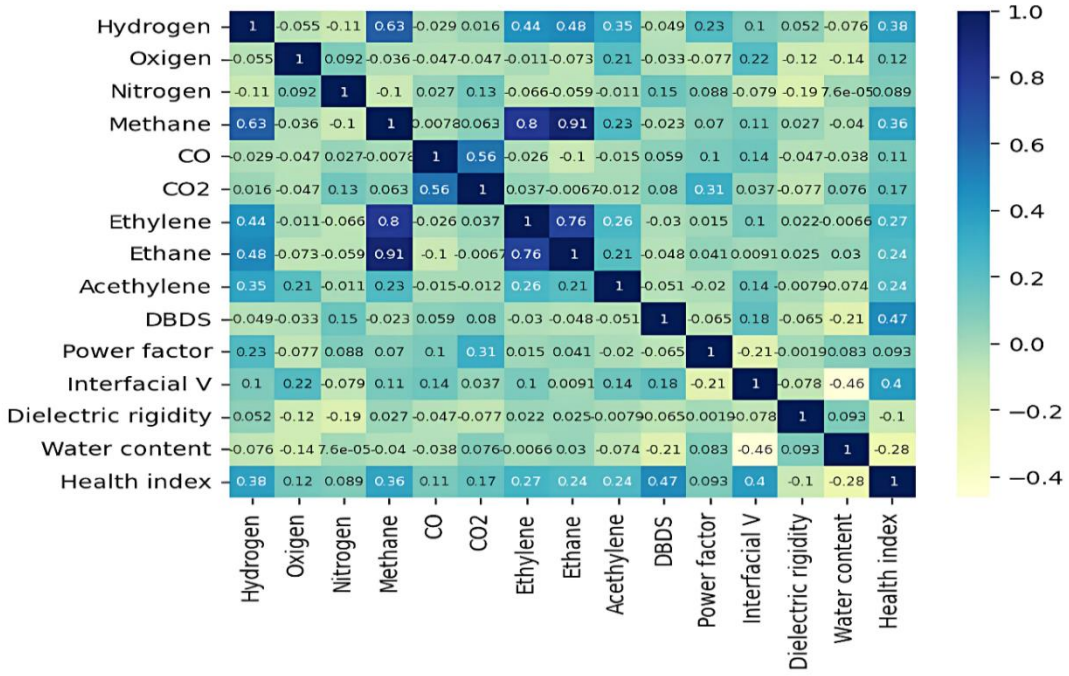

Figure 2: Heat map of DGA data

The heat map presented in Figure 2 provides insight into the correlation among all the predictors. The heat map is generated by coding in Python. Dibenzyl Disulfide (DBDS), Hydrogen, Methane, Interfacial V have a positive correlation with HI whereas Dielectric rigidity and water content have negative correlation with the HI. HI is encoded into five categories such as very poor, poor, fair, good, and very good represented with 0, 1,2,3,4 respectively for assessing the health of the transformer. Table 3 describes the detailed correlation between the variables used. Statistics such as mean, standard deviation, min and max are calculated to comprehend the distribution of the data samples. PCA is a dimensionality reduction technique used to get information from a high-dimensional space.

Table 3: Statistics of the DGA data

Statistics	Hydrogen	Oxygen	Nitrogen	Methane	CO	CO_2	Ethylene	Ethane	Acetylene	DBDS	Power factor	Interfacial V	Dielectric rigidity	Water content	HI
count	470.0	470.0	470.0	470.0	470.0	470.0	470.0	470.0	470.0	470.0	470.0	470.0	470.0	470.0	470.0
mean	404.26	8357.37	47759.5	79.69	244.0	18161.41	162.92	81.940	91.49	17.03	1.84	38.43	53.49	16.28	27.50
std	2002.14	14164.2	13760.4	489.32	237.26	22567.79	1323.81	342.57	644.36	46.735	6.14	6.17	6.45	17.11	17.74
min	0.00	57.0	3600.00	0.00	10.00	48.00	0.00	0.00	0.00	0.00	0.05	21.00	27.00	0.00	13.4
25%	4.00	496.00	41700	2.00	66.0	641.75	0.00	0.00	0.00	0.00	0.57	32.0	51.0	5.00	13.4
50%	9.00	3810.00	49100.0	3.00	150.50	11250.00	3.00	4.00	0.00	0.00	1.00	39.0	54.0	12.0	13.4
75%	34.0	14875.0	55875.0	7.00	361.75	22575.50	6.00	69.75	0.00	2.00	1.00	44.0	56.0	21.0	38.55
max	23349.0	249900.	85300.0	7406.00	1730.00	24900.0	16684.0	5467.00	9740.00	227.00	73.20	57.00	75.00	183.00	95.20

This is done by projecting it into a lower-dimensional sub-space. However, it should be ensured that the essential components with higher data variation are retained while eliminating non-essential components with a lower variation. In this context, dimensions refer to features that characterize the data. The results of PCA on the considered transformer data

do not yield good results as there is no clear elbow in the Variance plot of the PCA components. So, instead of eliminating any feature, the total data and all the features are classified. PCA curve and Scree plot are presented in Figure 3 (a) and Figure 3 (b), respectively. RF, kNN, SVM, and XGBoost Classifiers are trained with 470 samples (352 Training data + 118 Testing data) of 15 features of DGA data of transformer oil, which estimates the HI of the PT. Predictor variables considered are Hydrogen, Oxygen, Nitrogen, Methane, CO, CO2, Ethylene, Ethane, Acetylene, DBDS, Power Factor, Interfacial V, Die Electric Rigidity, Water Content, and the Target or Response variable considered is HI. As already mentioned, the training is carried out in two categories to evenly distribute the data: Training with Unbalanced data, which is carried out with 470 samples in which 352 samples are considered for training data and the rest 118 samples are used for testing data. Training with balanced data, which is carried out with 1425 samples
in which 1068 samples were considered for training data and the rest 357 samples were used for testing data. Table 4 below indicates the encoding of the HI in terms of the aforesaid categories.

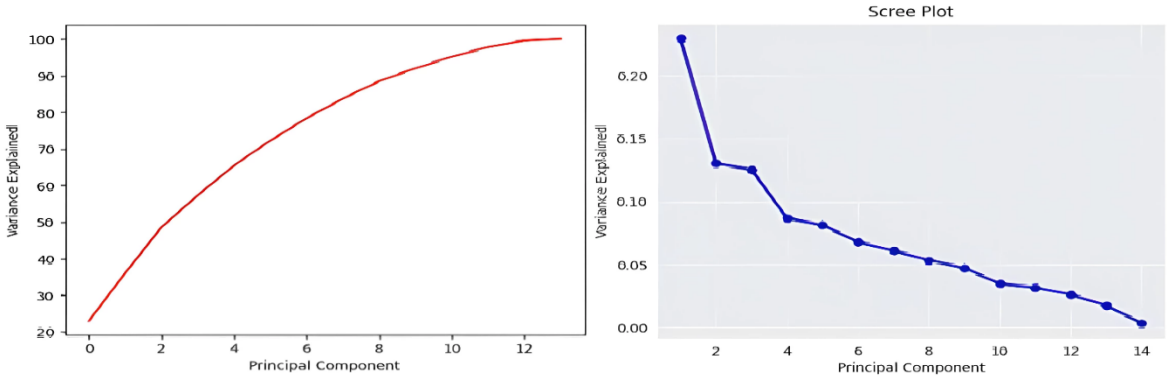

Figure 3: PCA: Principal Component Analysis (a) PCA plot (b) PCA-scree plot

Table 4: HI values encoded into categories

HI value	Encoded Condition
0	Very Poor
1	Poor
2	Fair
3	Good
4	Very Good

Consequently, the accuracy scores of HI prediction are enhanced considerably, and the details of the same are presented in this section. It is observed that RF and XGBoost Classifiers with data balancing using the 'IMB learn' library outperformed all others. The encoded categorical indicators are observed for the best fit RF classifier model to showcase the classification of the transformer's health condition as very poor, poor, fair, good, very good so that service engineers can take necessary action.

Table 5: Accuracy of the ML based Classifiers

Name of the ML Classifier	RF	KNN	SVM	XGBOOST
Accuracy Score without balancing using 'IMB learn' library	0.77	0.517	0.703	0.79
Accuracy Score with balancing using 'IMB learn' library	0.969	0.713	0.57	0.963

Thus, the proposed system can effectively contribute to the predictive maintenance.

5. Conclusion

Prediction of transformer HI was done using SVM. RF, XGBoost, kNN classifiers . The accuracy of RF, Knn, and XGBoost have increased with the oversampling technique of balancing, whereas the accuracy of SVM has decreased. XGBoost has shown superior performance with an accuracy of 79% without data balancing, and Random Forest showed superior performance with an accuracy of 96.9% with data balance. The use of these

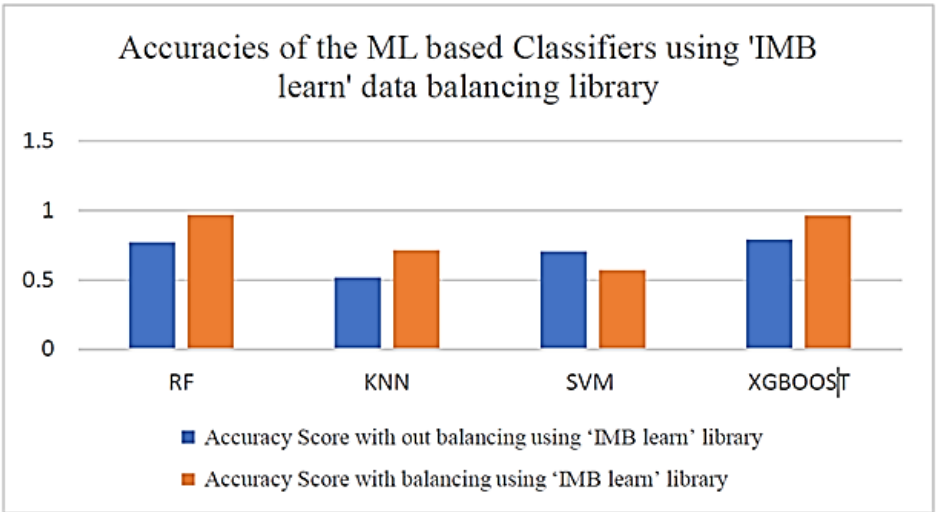

Figure 4: Test results: Accuracy of ML based Classifiers

classifiers and the data balancing technique, has served the purpose of predicting the transformer HI with higher accuracy for RF classifier.

References:

[1] S. V. Kulkarni and S.A. Khaparde, *Transformer Engineering. CRC Press*, 2004.

[2] G. Fotis, V. Vita, and T. I. Maris, "Risks in the European Transmission System and a Novel Restoration Strategy for a Power System after a Major Blackout," *Applied Sciences*, vol. 13, no. 1, pp. 83–83, Dec. 2022, Available: https://doi.org/10.3390/app13010083.

[3] V. Vita, G. Fotis, C. Pavlatos, and V. Mladenov, "A New Restoration Strategy in Microgrids after a Blackout with Priority in Critical Loads," *Sustainability*, vol. 15, no. 3, pp. 1974–1974, Jan. 2023, Available: https://doi.org/10.3390/su15031974.

[4] M. Zafeiropoulou *et al.*, "Forecasting Transmission and Distribution System Flexibility Needs for Severe Weather Condition Resilience and Outage Management," *Applied Sciences*, vol. 12, no. 14, p. 7334, Jan. 2022, Available: https://doi.org/10.3390/app12147334.

[5] Rajendra Prasad Upputuri, C. Vyjayanthi, and K Jaison, "Modeling and Detection of Inter-turn Faults in Distribution Transformer," *2019 8th International Conference on Power Systems (ICPS)*, Dec. 2019, Available: https://doi.org/10.1109/icps48983.2019.9067533.

[6] S. Sarkar, T. Sharma, A. Baral, B. Chatterjee, D. Dey, and S. Chakravorti, "An expert system approach for transformer insulation diagnosis combining conventional diagnostic tests and PDC, RVM data," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 21, no. 2, pp. 882–891, Apr. 2014, Available: https://doi.org/10.1109/tdei.2013.004052.

[7] S. Tenbohlen, S. Coenen, M. Djamali, A. Müller, M. H. Samimi, and M. Siegel, "Diagnostic Measurements for Power Transformers," *Energies*, vol. 9, no. 5, p. 347, May 2016, Available: https://doi.org/10.3390/en9050347.

[8] J. N'cho, I. Fofana, Y. Hadjadj, and A. Beroual, "Review of Physicochemical-Based Diagnostic Techniques for Assessing Insulation Condition in Aged Transformers," *Energies*, vol. 9, no. 5, p. 367, May 2016, Available: https://doi.org/10.3390/en9050367.

[9] S. Agarwal, "Data Mining: Data Mining Concepts and Techniques," *IEEE Xplore*, Dec. 01, 2013, Available: https://ieeexplore.ieee.org/abstract/document/6918822 (accessed Jan. 24, 2021).

[10] D. R. Morais and J. G. Rolim, "A Hybrid Tool for Detection of Incipient Faults in Transformers Based on the Dissolved Gas Analysis of Insulating Oil," *IEEE Transactions on Power Delivery*, vol. 21, no. 2, pp. 673–680, Apr. 2006, Available: https://doi.org/10.1109/tpwrd.2005.864044.

[11] P. Mirowski and Y. LeCun, "Statistical Machine Learning and Dissolved Gas Analysis: A Review," in *IEEE Transactions on Power Delivery*, vol. 27, no. 4, pp. 1791-1799, Oct. 2012, Available: 10.1109/TPWRD.2012.2197868.

[12] Z. M. Çınar, A. Abdussalam Nuhu, Q. Zeeshan, O. Korhan, M. Asmael, and B. Safaei, "Machine Learning in Predictive Maintenance towards Sustainable Smart Manufacturing in Industry 4.0," *Sustainability*, vol. 12, no. 19, p. 8211, Oct. 2020, Available: https://doi.org/10.3390/su12198211.

[13] Y. Jiang, B. Cukic, and Y. Ma, "Techniques for evaluating fault prediction models," *Empirical Software Engineering*, vol. 13, no. 5, pp. 561–595, Aug. 2008, Available: https://doi.org/10.1007/s10664-008-9079-3.

[14] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, Jul. 2013, Available: https://doi.org/10.1109/icccnt.2013.6726842.

[15] N. A. Bakar and A. Abu-Siada, "Fuzzy logic approach for transformer remnant life prediction and asset management decision," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 23, no. 5, pp. 3199–3208, Oct. 2016, Available: https://doi.org/10.1109/tdei.2016.7736886.

[16] N. Islam, R. Khan, S. K. Das, S. K. Sarker, M. M. Islam, M. Akter, and S. M. Muyeen, "Power transformer health condition evaluation: A deep generative model aided intelligent framework," *Electric Power Systems Research*, vol. 218, p. 109201, 2023. [Online]. Available: https://doi.org/10.1016/j.epsr.2023.109201.

[17] Z. Xing and Y. He, "Multimodal Mutual Neural Network for Health Assessment of Power Transformer," *IEEE Systems Journal*, pp. 1–10, 2023. DOI: 10.1109/JSYST.2023.3237225. [Online].Available:https://www.researchgate.net/publication/367455842_Multimodal_Mutual_Neural_Network_for_Health_Assessment_of_Power_Transformer.

[18] L. Jin, D. Kim, K. Y. Chan, and A. Abu-Siada, "Deep Machine Learning-Based Asset Management Approach for Oil-Immersed Power Transformers Using Dissolved Gas Analysis," *IEEE Access*, vol. 12, pp. 27794–27809, 2024, doi: 10.1109/ACCESS.2024.3366905.

[19] S. Fei and X. Zhang, "Fault diagnosis of power transformer based on support vector machine with genetic algorithm," *Expert Systems with Applications*, vol. 36, no. 8, pp. 11352–11357, Oct. 2009, Available: https://doi.org/10.1016/j.eswa.2009.03.022.

[20] H. Yuan, G. Wu, and B. Gao, "Fault diagnosis of power transformer using particle swarm optimization and extreme learning machine based on DGA," *High Volt. Appar.,* vol. 52, pp. 176–180, Nov. 2016, Available: 10.13296/j.1001-1609.hva.2016.11.029

[21] X. Shi, Y. Zhu, X. Ning, L. Wang, G. Sun, and G. Chen, "Transformer fault diagnosis based on deep auto-encoder network," *Electric Power Automation Equipment* , vol. 36, pp. 122–126, May 2016.

[22] Z. Xing, Y. He, J. Chen, X. Wang, and B. Du, "Health evaluation of power transformer using deep learning neural network," *Electric Power Systems Research*, vol. 215, Part B, p. 109016, 2023. DOI: 10.1016/j.epsr.2022.109016.

[23] A. Basuki and Suwarno, "Online Dissolved Gas Analysis of Power Transformers Based on Decision Tree Model," *in 2018 Conference on Power Engineering and Renewable Energy (ICPERE),* Solo, Indonesia, 2018, pp. 1-6, Available: 10.1109/ICPERE.2018.8739761.

[24] P. Sarajcev, D. Jakus, J. Vasilj and M. Nikolic, "Analysis of Transformer Health Index Using Bayesian Statistical Models," in *2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech)*, Split, Croatia, 2018, pp. 1-7.

[25] K. Leauprasert, T. Suwanasri, C. Suwanasri and N. Poonnoy, "Intelligent Machine Learning Techniques for Condition Assessment of Power Transformers," in *2020 International Conference on Power, Energy and Innovations (ICPEI),* Chiangmai, Thailand, 2020, pp. 65-68,

Available: 10.1109/ICPEI49860.2020.9431460.

[26] Atul Jaysing Patil, A. Singh, and R. K. Jarial, "An Integrated Fuzzy based Online Monitoring System for Health Index and Remnant Life Computation of 33 kV Steel Mill Transformer," Feb. 2020, Available: https://doi.org/10.1109/i4tech48345.2020.9102698.

[27] "IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers," in *IEEE Std C57.104-2008 (Revision of IEEE Std C57.104-1991)*, vol., no., pp. 1-36, 2 Feb. 2009, Available: 10.1109/IEEESTD.2009.4776518.

[28] "IEC 60567:2023 | IEC Webstore," *webstore.iec.ch*. https://webstore.iec.ch/publication/70013 (accessed Dec. 31, 2023).

[29] "Brown Boveri Review," 1974. Accessed: Dec. 31, 2023. [Online]. Available: https://library.e.abb.com/public/01ea301de5f64f6c8317bb1e28b6c2b2/bbc_mitteilungen_1974_e_12.pdf

[30] R. Rogers, "IEEE and IEC Codes to Interpret Incipient Faults in Transformers, Using Gas in Oil Analysis," *IEEE Transactions on Electrical Insulation*, vol. EI-13, no. 5, pp. 349–354, Oct. 1978, Available: https://doi.org/10.1109/tei.1978.298141.

[31] Arias, Ricardo; Mejia Lara, Jennifer , "Data for: Root cause analysis improved with machine learning for failure analysis in power transformers", Mendeley Data, V1, Available: 10.17632/rz75w3fkxy.1

[32] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal: Promoting Communications on Statistics and Stata*, vol. 20, no. 1, pp. 3–29, Mar. 2020, Available: https://doi.org/10.1177/1536867x20909688.

[33] G. Guo, H. Wang, D. A. Bell, Y. Bi, " KNN Model-Based Approach in Classification", Available: https://www.researchgate.net/publication/2948052_KNN_Model-Based_Approach_in_Classification.

[34] T. Evgeniou and M. Pontil, "Support Vector Machines: Theory and Applications," *Machine Learning and Its Applications*, pp. 249–257, 2001, Available: https://doi.org/10.1007/3-540-44673-7_12.

[35] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery*. pp. 785-794, USA, 2016, Available: https://doi.org/10.1145/2939672.2939785.