Iraqi Journal of Science, 2025, Vol. 66, No. 6, pp: 2581-2592
DOI: 10.24996/ijs.2025.66.6.31

ISSN: 0067-2904

Application of Data Science Techniques and Machine Learning based classifiers for Transformer Health Assessment

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: 29/9/2024 Accepted: 18/11/2024 Published: 30/6/2025

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. Principle Component Analysis (PCA) is used to estimate the principal components of the data, helping in selecting the features that 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 were 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 Neighbours.

1. Introduction

Power Transformers (PTs) are prominent equipment in the Power System Network (PSN). They essentially regulate the levels along the stretch of the transmission and distribution line from the generator to the final load center [1]. To ensure the healthy operation of PTs, power sectors consider certain maintenance strategies. There are three categories of maintenance. One among them is 'run-to-failure' , where the action is taken after the aftermath of a transformer failure. Another category is ‘preventive maintenance’, in which action is executed based on a predetermined schedule. The last category is ‘predictive maintenance’, the cheapest among all three. Predictive maintenance aims to evaluate the health of every device, enabling advanced detection of impending failures. Health assessment of PT is an important driving factor with regard to predictive maintenance. Opting for predictive maintenance would prove advantageous, enhancing the reliability of system operations and minimizing unexpected power supply interruptions. Both predictive maintenance and preventive maintenance offer shared benefits in this regard. However, predictive maintenance goes a step further by curbing

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

costs through the avoidance of unnecessary maintenance operations. Existing research and practical applications of predictive maintenance predominantly center on large PTs.

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 major role in transformer failures, it is required to instigate a weather monitoring system in transmission and distributed system to keep track of the conditions [4]. The other reasons could be cooling system failures, overloading, overvoltage, and overcurrent [5]. Therefore, power suppliers prioritize health assessment of transformers for effective and healthy operation.

2. Literature survey

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 Physico-chemical diagnostic methods employed for assessing the insulation condition in aging transformers. Developing fault prediction models often involves the utilization of 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 was significantly contributed to by DGA analysis. Transformer oil Dissolved Gas Analysis (DGA) presented in [10] is a useful aspect of transformer health assessment/ index. The techniques employed to evaluate transformer health rely on Dissolved Gas Analysis (DGA). DGA examines the concentration of particular gases, pertaining to 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 in 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 Dissolved Gas Analysis (DGA) data, water content, furan, and interfacial tension (IFT). The goal was to analyze the health conditions of the transformer and evaluate the remaining lifespan of the transformer based on operating temperature. The predominant focus in current research is on evaluating the Health Index (HI) of transformers by scrutinizing the deterioration of oil using Dissolved Gas Analysis (DGA). Dissolved gas analysis (DGA) has already been integrated with machine learning techniques, including proven methods, such as Artificial Neural Networks (ANN) [16-17] and fuzzy logic [18].

Additionally, support vector machines and deep belief networks have also been utilized, including the extreme learning machine (ELM) [19-21]. These approaches leverage DGA data from the past to identify patterns and gauge the health of transformers. An improved health

assessment utilizing a fuzzy logic system based on Dissolved Gas Analysis (DGA) alone is proposed in [22]. However, this focuses exclusively on DGA-based health evaluation. As presented in [23], online condition monitoring for High Impedance (HI) using Dissolved Gas Analysis (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 were compared with those derived from other models.

Another study by Sarajcev et al. in 2018 [24], presented a Bayesian multinomial logistic regression model for estimating transformer Health Index (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 the implementation of online learning/monitoring, offering potential benefits for health assessment.

The study discussed by Leauprasert in 2020 [25], presented the utilization of regression models to assess the Health Index (%HI) in terms of percentage for estimating the condition of the transformer. However, it highlights limitations, such as the lack of a proper elucidation of model parameters. Additionally, 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 33kV steel-mill transformers. It utilizes fuzzy models for computing the Health Index (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 those in generation and transmission. Although there is considerable ML-based HI predictions-based health assessment performed with higher accuracies, the 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 followed a specific methodology, and the different stages involved are detailed in this section. The entire implementation is based on the DGA, which is carried out in certain methods. Renowned methods of DGA are also presented in this section.

a. Dissolved Gas Analysis (DGA)

Dissolved Gas Analysis (DGA) is a diagnostic technique employed to evaluate the condition of PTs. This technique involves analyzing gases of the insulating oil pertaining to transformers. The identification and measurement of distinct gas concentrations in oil offers valuable insights into internal problems like overheating, electrical discharges, and insulation deterioration. DGA is 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 H2, CH4, C2H2, C2H4, C2H6, CO, and CO2, 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 C2H2/C2H4, C2H2/CH4, C2H4/C2H6, 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 as well as thermal faults of transformers. Limitations, such as inconsistencies

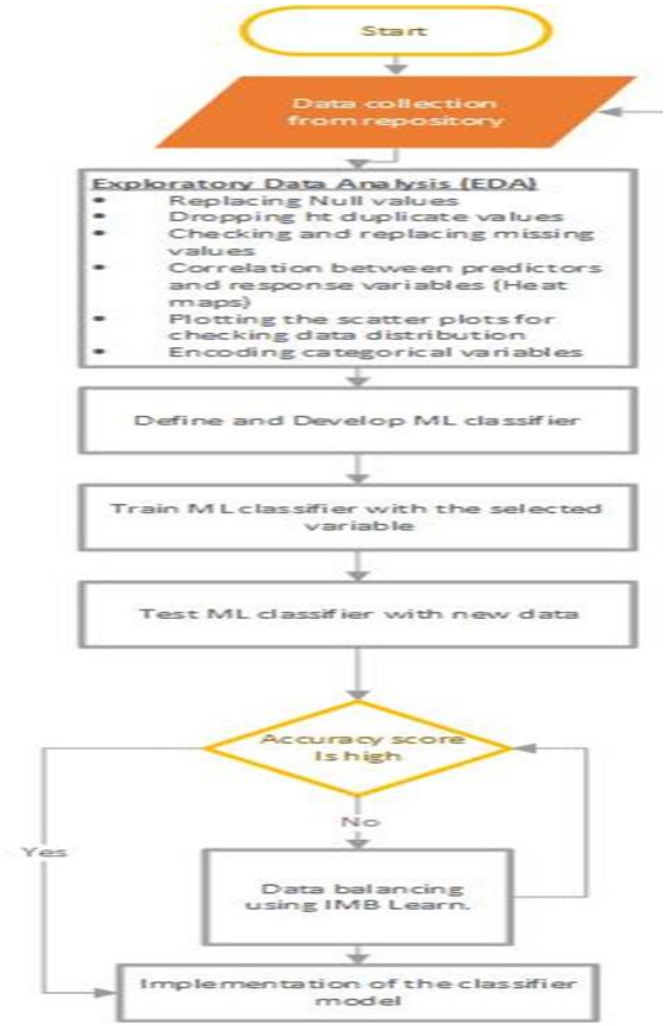

Figure1: Flow chart of ML classifier Implementation using DGA data

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.

Stage 1: Data acquisition

	Hydrogen	Oxigen	Nitrogen	Methane	CO	CO2	Ethylene	Ethane	Acethylene	DBDS	Power factor	Interfacial V	Dielectric rigidity	Water content	Health index
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
2	2820	16400	56300	144	257	1080	206	11	2190	1.0	1.00	39	52	11	85.3
3	1099	70	37520	545	184	1402	6	230	0	87.0	4.58	33	49	5	85.3
4	3210	3570	47900	160	360	2130	4	43	4	1.0	0.77	44	55	3	85.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
465	15	227	52900	3	60	853	3	84	0	0.0	1.00	32	56	28	13.4
466	15	334	47100	3	64	622	3	108	0	0.0	1.00	32	55	12	13.4
467	15	1280	35000	2	675	2530	0	0	0	5.0	0.30	45	58	8	13.4
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

470 rows × 15 columns

Figure 2: Transformer oil DGA data

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. The data has 470 samples of 15 variables, and they are Hydrogen, Oxygen, Nitrogen, Methane, CO, CO2, Ethelene, Ethane, Acethylene, DBDS, Power Factor, Interfacial V, Die Electric Rigidity, Water Content, and Health Index, these are listed in the 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 the field of electrical engineering, data science is used to decipher and enhance various facets of the field. Exploratory Data Analysis (EDA) is an effective tool in Data Science to preprocess the data and make it compatible for 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

The application of 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 with respect to HI predictions. Theoretical background about these classifiers is presented in this section.

a. Random Forest ML algorithm

Ho (1995) initially proposed this concept of the random-subspace method, but Breiman (2001) expanded this as Random Forest (RF). This model represents an algorithm whose base is 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. Even though interpretability becomes complicated, the prediction performance is enhanced. RF consistently delivers an accurate estimation of the error rate in comparison 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 neighbor 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 identifying an "optimal" hyperplane for solving the issue. The basic SVM structuring is linear, and in this, 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 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 this addresses the problem of decision trees prone to overfitting. Lastly, XGBoost, being a tree structure model, eliminates the need to normalize data collected by Phasor Measurement Units (PMU) in power systems. Additionally, it effectively handles missing values, rendering it suitable for PMU-based transient stability prediction by uncovering relationships between features and transient stability.

Stage 4: Define and develop ML based Classifiers

As described in the paragraph above, four different classifiers are considered in this paper, where RF, KNN and SVM classifiers are called from 'SCIKIT Learn' python library as functions, and the data is fit. Secondly, XGBoost classifier is imported from 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 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 work 'IMB learn' library is used to carry out the process of balancing the data.

4. Results: EDA, Training, Testing performed

To start with, the outcomes are of the EDA process, such as Correlation Matrix, Heat map, and Statistics of the data to examine the nature of the data and how well it suits the application considered. Correlation Matrix, Heat maps, and scatter plots are presented in Figures 3-5.

	Hydrogen	Oxigen	Nitrogen	Methane	CO	CO2	Ethylene	Ethane	Acethylene	DBDS	Power factor	Interfacial V	Dielectric rigidity	Water content	Health index
Hydrogen	1.000000	-0.054989	-0.114050	0.634567	-0.029333	0.016366	0.442336	0.482226	0.354975	-0.048598	0.227482	0.101098	0.051834	-0.076245	0.377388
Oxigen	-0.054989	1.000000	0.092438	-0.035593	-0.047161	-0.047253	-0.010588	-0.073319	0.207559	-0.032990	-0.076762	0.219819	-0.123626	-0.141714	0.121009
Nitrogen	-0.114050	0.092438	1.000000	-0.101023	0.027335	0.134220	-0.066404	-0.058571	-0.011292	0.152747	0.087955	-0.078527	-0.191573	0.000076	0.089455
Methane	0.634567	-0.035593	-0.101023	1.000000	-0.007771	0.063087	0.801129	0.914559	0.230079	-0.023055	0.070275	0.112699	0.026885	-0.039642	0.361770
CO	-0.029333	-0.047161	0.027335	-0.007771	1.000000	0.555432	-0.025930	-0.099748	-0.014762	0.059271	0.103007	0.144189	-0.046546	-0.038400	0.112751
CO2	0.016366	-0.047253	0.134220	0.063087	0.555432	1.000000	0.036822	-0.006660	-0.011840	0.080156	0.309407	0.037421	-0.076638	0.076053	0.168777
Ethylene	0.442336	-0.010588	-0.066404	0.801129	-0.025930	0.036822	1.000000	0.755344	0.255074	-0.029814	0.014871	0.103618	0.022044	-0.006573	0.271504
Ethane	0.482226	-0.073319	-0.058571	0.914559	-0.099748	-0.006660	0.755344	1.000000	0.206291	-0.047526	0.041412	0.009142	0.025386	0.029699	0.236507
Acethylene	0.354975	0.207559	-0.011292	0.230079	-0.014762	-0.011840	0.255074	0.206291	1.000000	-0.050700	-0.019516	0.136413	-0.007921	-0.074175	0.240143
DBDS	-0.048598	-0.032990	0.152747	-0.023055	0.059271	0.080156	-0.029814	-0.047526	-0.050700	1.000000	-0.064931	0.183434	-0.064712	-0.207384	0.468809
Power factor	0.227482	-0.076762	0.087955	0.070275	0.103007	0.309407	0.014871	0.041412	-0.019516	-0.064931	1.000000	-0.209686	-0.001867	0.082673	0.092729
Interfacial V	0.101098	0.219819	-0.078527	0.112699	0.144189	0.037421	0.103618	0.009142	0.136413	0.183434	-0.209686	1.000000	-0.078438	-0.464783	0.400216
Dielectric rigidity	0.051834	-0.123626	-0.191573	0.026885	-0.046546	-0.076638	0.022044	0.025386	-0.007921	-0.064712	-0.001867	-0.078438	1.000000	0.092870	-0.104426
Water content	-0.076245	-0.141714	0.000076	-0.039642	-0.038400	0.076053	-0.006573	0.029699	-0.074175	-0.207384	0.082673	-0.464783	0.092870	1.000000	-0.281165
Health index	0.377388	0.121009	0.089455	0.361770	0.112751	0.168777	0.271504	0.236507	0.240143	0.468809	0.092729	0.400216	-0.104426	-0.281165	1.000000

Figure 3: Correlation Matrix of DGA data

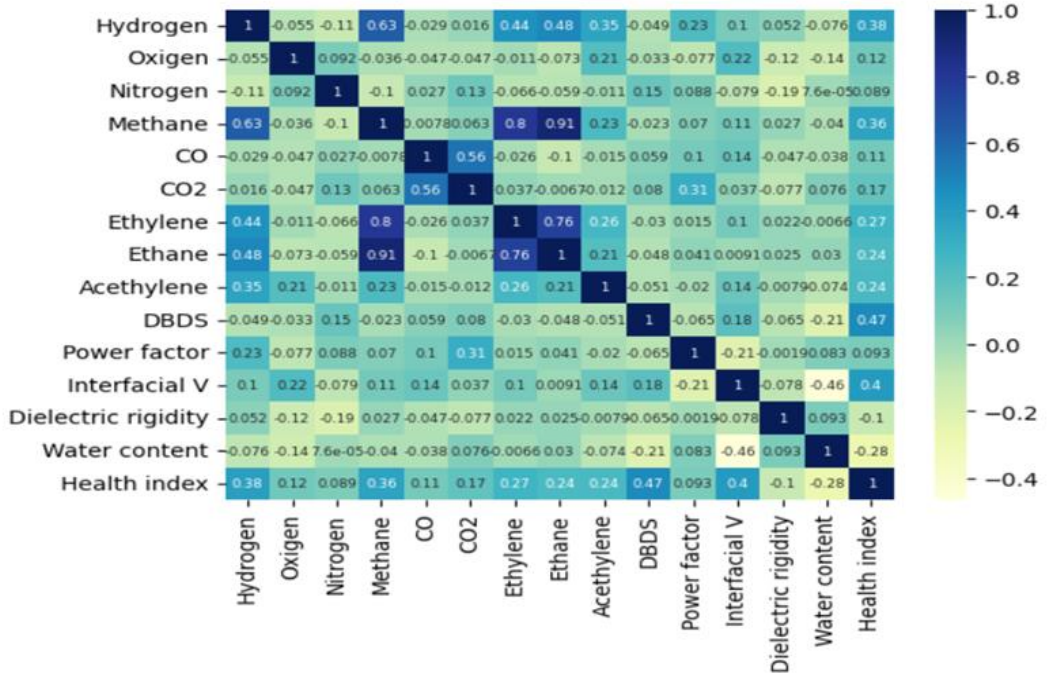

Figure 4: Heat map of DGA data

The data collected for training machine learning algorithms undergoes initial scrutiny through a heat map, enabling the analysis of relationships between different predictors and the target variable. The heat map presented in Figure 4 provides insights into the correlation among all the predictors.

The heat map is generated by coding in Python. Dibenzyl Disulfide (DBDS), Hydrogen, Methane, and Interfacial V have a positive correlation with the Health Index, whereas Dielectric rigidity and water content have a negative correlation with the Health Index. 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. Figure 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. Principal Component Analysis (PCA) is a dimensionality reduction technique, used for getting information from a high-dimensional space. 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 lower variation. In this context, dimensions refer to features that characterize the data.

	Hydrogen	Oxigen	Nitrogen	Methane	CO	CO2	Ethylene	Ethane	Acethylene	DBDS	Power factor	Interfacial V	Dielectric rigidity	Water content	Health index
count	470.000000	470.000000	470.000000	470.000000	470.000000	470.000000	470.000000	470.000000	470.000000	470.000000	470.000000	470.000000	470.000000	470.000000	470.000000
mean	404.261702	8357.372340	47759.561702	79.695745	244.000000	1816.414894	162.923404	81.940426	91.491489	17.036596	1.849043	38.434043	53.495745	16.282979	27.504043
std	2002.142678	14164.233283	13760.451816	489.320336	237.267485	2256.790519	1323.811504	342.573636	644.365828	46.735057	6.144009	6.178830	6.458906	17.115646	17.741458
min	0.000000	57.000000	3600.000000	0.000000	10.000000	48.000000	0.000000	0.000000	0.000000	0.000000	0.050000	21.000000	27.000000	0.000000	13.400000
25%	4.000000	496.000000	41700.000000	2.000000	66.000000	641.750000	0.000000	0.000000	0.000000	0.000000	0.570000	32.000000	51.000000	5.000000	13.400000
50%	9.000000	3810.000000	49100.000000	3.000000	150.500000	1125.000000	3.000000	4.000000	0.000000	0.000000	1.000000	39.000000	54.000000	12.000000	13.400000
75%	34.000000	14875.000000	55875.000000	7.000000	361.750000	2257.500000	6.000000	69.750000	0.000000	2.000000	1.000000	44.000000	56.000000	21.000000	38.550000
max	23349.000000	249900.000000	85300.000000	7406.000000	1730.000000	24900.000000	16684.000000	5467.000000	9740.000000	227.000000	73.200000	57.000000	75.000000	183.000000	95.200000

Figure 5: Statistics of the DGA data

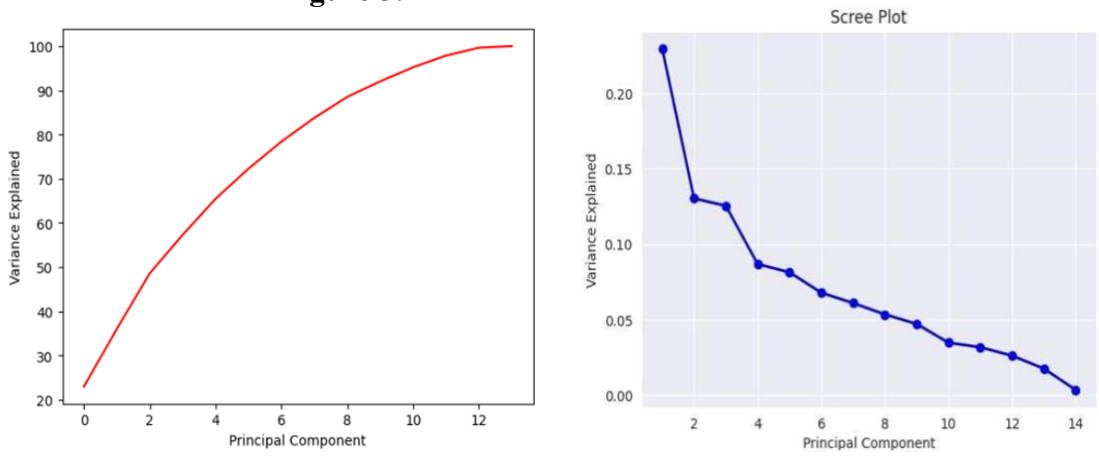

Figure 6: PCA: Principle Component Analysis (a) PCA plot (b) PCA- scree plot

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, along with all the features, is used for classification. PCA curve and Scree plot are presented in Figure 6, respectively.

As mentioned in the section before, RF, kNN, SVM, and XGBoost ML-based Classifiers are trained with 470 samples (352 Training data + 118 Testing data) of 15 variables of DGA data of transformer oil, which estimates the HI of the PT. Predictor variables considered are

Hydrogen, Oxygen, Nitrogen, Methane, CO, CO2, Ethelene, Ethane, Acethylene, DBDS, Power Factor, Interfacial V, Die Electric Rigidity, Water Content, and the Target or Response variable considered is Health Index (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 is carried out with 1425 samples in which 1068 samples are considered for training data and the rest 357 samples are used for testing data. 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. A comprehensive comparison between the classifiers is presented in Figure 7.

The encoded categorical indicators are observed for best fit RF classifier model to showcase the classification of the transformer's health condition as very poor, poor, fair, good, and very good, so that service engineers can take necessary action.

Table 1: 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. Figure 8 below indicates the encoding of the HI in terms of the foresaid categories.

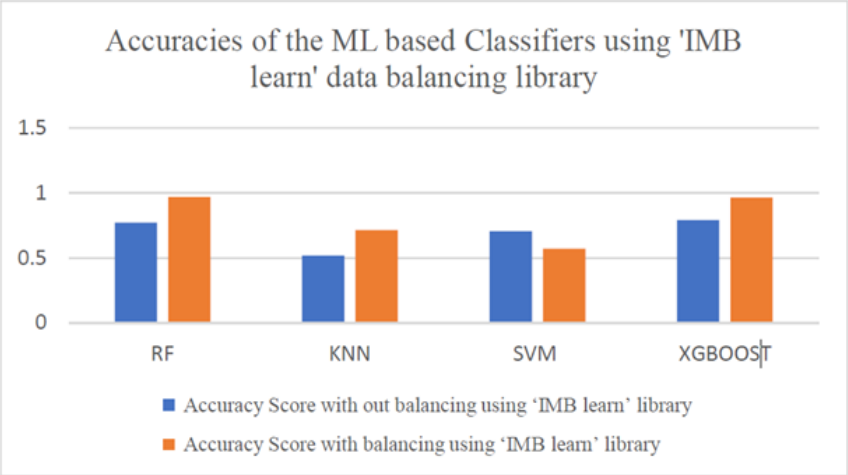

Figure 7: Test results: Accuracy of ML based Classifiers

HI value	Encoded condition
3	Very Good
3	Very Good
3	Very Good
3	Very Good
3	Very Good

Figure 8: HI values encoded into categories

5. Conclusion

Prediction of transformer health index was done using SVM. RF, XGBoost, kNN classifiers, . The accuracy of RF, Knn, XGBoost increased with the usage of the oversampling technique of balancing, whereas the accuracy of SVM has gone down. 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 classifiers, along with the data balancing technique, has served the purpose of predicting the transformer Health Index 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, doi: 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, doi: 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, doi: 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, doi: 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, doi: 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, doi: 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, doi: https://doi.org/10.3390/en9050367.

[9] S. Agarwal, "Data Mining: Data Mining Concepts and Techniques," *IEEE Xplore*, Dec. 01, 2013. 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, doi: 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, doi: 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, doi: 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, doi: 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, doi: 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, doi: https://doi.org/10.1109/tdei.2016.7736886.

[16] Y. Zhang, X. Ding, Y. Liu, and P. J. Griffin, "An artificial neural network approach to transformer fault diagnosis," *IEEE Transactions on Power Delivery*, vol. 11, no. 4, pp. 1836–1841, 1996, doi: https://doi.org/10.1109/61.544265.

[17] J.-H. Yi, J. Wang, and G.-G. Wang, "Improved probabilistic neural networks with self-adaptive strategies for transformer fault diagnosis problem," *Advances in Mechanical Engineering*, vol. 8, no. 1, p. 168781401562483, Jan. 2016, doi: https://doi.org/10.1177/1687814015624832.

[18] R. Naresh, V. Sharma, and M. Vashisth, "An Integrated Neural Fuzzy Approach for Fault Diagnosis of Transformers," *IEEE Transactions on Power Delivery*, vol. 23, no. 4, pp. 2017–2024, Oct. 2008, doi: https://doi.org/10.1109/tpwrd.2008.2002652.

[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, doi: 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, doi: 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," *Electr. Power Autom. Equip*, vol. 36, pp. 122–126, May 2016.

[22] E. Aburaghiega, M. Emad Farrag, D. Hepburn and A. Haggag, "Enhancement of Power Transformer State of Health Diagnostics Based on Fuzzy Logic System of DGA," 2018 Twentieth International Middle East Power Systems Conference (MEPCON), Cairo, Egypt, 2018, pp. 400-405, doi: 10.1109/MEPCON.2018.8635154.

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

[24] P. Sarajcev, D. Jakus, J. Vasilj and M. Nikolic, "Analysis of Transformer Health Index Using Bayesian Statistical Models," 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," 2020 International Conference on Power, Energy and Innovations (ICPEI), Chiangmai, Thailand, 2020, pp. 65-68, doi: 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, doi: 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, doi: 10.1109/IEEESTD.2009.4776518.Guide for the sampling of gases and of oil-filled electrical equipment and for the analysis of free and dissolved gases. IEC.

[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, doi: https://doi.org/10.1109/tei.1978.298141.

[31] Arias, Ricardo; Mejia Lara, Jennifer (2020), "Data for: Root cause analysis improved with machine learning for failure analysis in power transformers", Mendeley Data, V1, doi: 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, doi: https://doi.org/10.1177/1536867x20909688.

[33] "(PDF) KNN Model-Based Approach in Classification," *ResearchGate*. 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, doi: https://doi.org/10.1007/3-540-44673-7_12.

[35] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 785-794, 2016