International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(19s) 2026 pp. 1-13
https://doi.org/10.70917/ijcisim-2026-4997
Article

Machine Learning Based Prediction of Transformer Health Index and Remaining Life Using Transformer Oil Parameters

Salava V. Satyanarayana[1], Dr. P. Suresh Kumar[2]

[1] Department of Electrical and Electronics Engineering, Dr. YSR ANU College of Engineering and Technology, Guntur, Andhra Pradesh, India.
Email: pshve2011@gmail.com
[2] Department of Mechanical Engineering, R.V.R. & J.C. College of Engineering, Guntur, Andhra Pradesh, India.
Email: prohansuresh@gmail.com

Abstract: Power transformers are important components of electrical power systems and it is essential that they operate reliably if continuous power supply is to be maintained. Because they are constantly subjected to electrical, thermal, mechanical, and environmental stresses, the insulation suffers degradation, a fact that can be seen in the physicochemical properties and in the characteristics of the dissolved gases in the transformer oil. Traditional methods of assessing condition depend on periodic testing of the oil and on the judgement of experts, a procedure which is time-consuming and open to personal bias. The present paper describes a machine learning-based approach for predicting the Transformer Health Index (HI) and the Remaining Life (RL) by means of the diagnostic parameters of transformer oil. The model used in this study is based on a dataset consisting of 470 samples of transformer oil and including 14 diagnostic features, namely the concentrations of dissolved gases (H_2, O_2, N_2, CH_4, CO, CO_2, C_2H_4, C_2H_6 and C_2H_2), Dibenzyl Disulfide (DBDS), power factor, interfacial voltage, dielectric rigidity and water content. Following data preprocessing and feature engineering, Random Forest Regression is applied in order to identify the complex nonlinear relationships between the oil parameters and the condition of the transformer. The model's performance is assessed using the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE) and the coefficient of determination (R^2). The results of the experiments show that the proposed framework is able to make accurate predictions of both the Health Index and the Remaining Life, thus allowing a reliable evaluation of the transformer's condition. The method developed supports predictive maintenance by allowing for timely planning of maintenance, reducing the occurrence of unexpected transformer failures, extending the service life of the assets and improving the overall efficiency of asset management for transformers in modern power systems ..
Keywords: Power Transformer,Health Index,Remaining Life Prediction,Random Forest Regression,Machine Learning

1. Introduction

Power transformers are essential elements of today's electrical power systems and act as the main means of transforming voltage throughout the processes of power generation, transmission and distribution. The reliability and continuous operation of power networks are greatly influenced by the condition of these transformers. If transformers fail unexpectedly, this can lead to extended outages, heavy repair expenses, environmental hazards, and large economic losses. Because of this, there is now a great deal of concern among electric utilities all over the world regarding the accurate evaluation of transformer condition Transformer insulation is continually deteriorating during its service life as a result of thermal aging, electrical discharges, oxidation, moisture entering the insulation, contamination, and mechanical stresses. The changes caused by these degradation processes modify the physical and chemical properties of the transformer insulating oil, which is why oil analysis is one of the most effective non-invasive methods for evaluating the condition of a transformer. Various parameters including the concentrations of dissolved gases, dielectric strength, water content, interfacial tension, acidity, power factor, and sulphur compounds offer useful information about the internal condition of the transformer. In particular, Dissolved Gas Analysis (DGA) has been widely used for detecting early faults and insulation deterioration by examining the gases produced under abnormally operating conditions [1].

Power transformers are essential elements of today's electrical power systems and act as the main means of transforming voltage throughout the processes of power generation, transmission and distribution. The reliability and continuous operation of power networks are greatly influenced by the condition of these transformers. If transformers fail unexpectedly, this can lead to extended outages, heavy repair expenses, environmental hazards, and large economic losses. Because of this, there is now a great deal of concern among electric utilities all over the world regarding the accurate evaluation of transformer condition.Transformer insulation is continually deteriorating during its service life as a result of thermal aging, electrical discharges, oxidation, moisture entering the insulation, contamination, and mechanical stresses. The changes caused by these degradation processes modify the physical and chemical properties of the transformer insulating oil, which is why oil analysis is one of the most effective non-invasive methods for evaluating the condition of a transformer. Various parameters including the concentrations of dissolved gases, dielectric strength, water content, interfacial tension, acidity, power factor, and sulphur compounds offer useful information about the internal condition of the transformer. In particular, Dissolved Gas Analysis (DGA) has been widely used for detecting early faults and insulation deterioration by examining the gases produced under abnormally operating conditions [2].

Machine learning has become a promising approach to assessing the condition of transformers since it is capable of modelling complex nonlinear relationships directly from historical operational data. By using supervised learning algorithms, it is possible to detect hidden patterns in thevarious parameters of transformer oil and to predict health-related indices with a high degree of accuracy. Of the different machine learning algorithms available, Random Forest Regression has attracted a great deal of attention because of its robustness in avoiding overfitting, its ability to deal with correlated features, its resistance to noise, and its capacity to estimate nonlinear relationships without the need for explicit mathematical models [3].

Main contributions of this work.

A machine learning based framework for Transformer Health Index (THI) prediction based on the transformer oil diagnostic parameters.

Implementation of a Random Forest Regression (RFR) approach for estimating the life remaining of the transformer.

All fourteen dissolved gas and oil quality parameters were studied jointly in order to extract better transformer degradation signature.

Correlation and prediction study was made between Health Index and Remaining life of the transformer.

The predictive framework offers an intelligent decision support tool for transformer maintenance and asset management

2. Methodology

The proposed machine learning framework and workflow to predict the Transformer Health Index (HI) and Remaining Life (RL) utilizing transformer oil's diagnostic parameter has also been shown in this chapter. This frame work has combined the use of the Transformer oil's condition monitoring parameters along with the application of supervised machine learning to form a condition monitoring system capable of determining condition which will be highly relevant in the concept of smart asset management, predictive maintenance and life asset modeling. These processes can be broadly classified under stages such as Data Acquisition, Pre-processing, Feature Selection, Model Building, HI and RL Prediction, Evaluation and visualization of result [4].

TABLE I. Transformer Oil Diagnostic Parameters

Feature	**Description**
Hydrogen (H_2)	Indicates partial discharge and low-energy faults
Oxygen (O_2)	Oxidation indicator
Nitrogen (N_2)	Atmospheric gas concentration
Methane (CH_4)	Thermal fault indicator
Carbon Monoxide (CO)	Cellulose insulation degradation
Carbon Dioxide (CO_2)	Paper insulation aging
Ethylene (C_2H_4)	High-temperature overheating
Ethane (C_2H_6)	Moderate thermal faults
Acetylene (C_2H_2)	Electrical arcing
DBDS	Corrosive sulfur compound
Power Factor	Dielectric loss indicator
Interfacial Voltage	Oil oxidation condition
Dielectric Rigidity	Insulation strength
Water Content	Moisture contamination

The dataset utilized for this research comprises a total of 470 transformer oil samples, where one sample corresponds to the condition of one power transformer. A total of 14 diagnostic transformer oil data are presented as input features, and the life Expectation (Remaining life) and Health Index are represented as two output features of samples. Dataset is built up based on a large amount of record on power transformer oil analysis, including measurements that address to DGA and some physical and chemical properties of oil, giving an overview of power transformer insulation health condition.

Target Variable(s):

Health Index (HI): Numerical value that reflects the health of transformer.

Life Expectation (LE): Total life expectation in years. Here LE is also referred to as Remaining Life (RL).

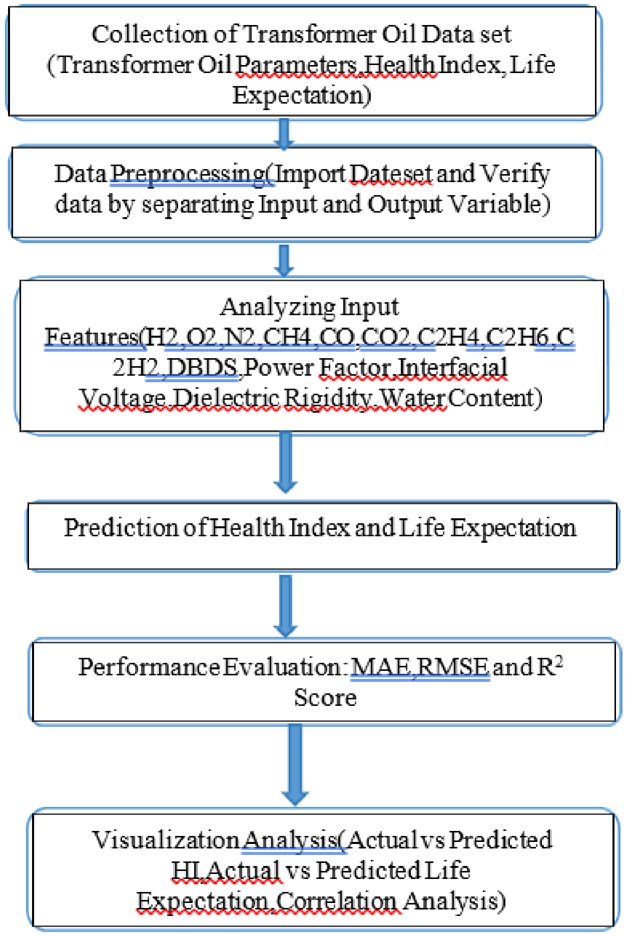

Fig.1. Proposed Workflow of Prediction of Health Index and Life Expectation

The proposed process begins with collecting a dataset of transformer oil containing 470 transformer sample and 14 diagnostic oil parameters along with Health Index (HI) and Life Expectation (Remaining Life) values. The dataset is imported to the Python programming environment and preprocessed with the verification of the data and differentiating the predictor variables from the target variables. The fourteen parameters of transformer oil are used for forming the input feature matrix and Health Index and Remaining Life are used as independent target variables. The two Random Forest Regression models are developed using 200 decision trees with fixed random state. The first model predicts the transformer Health Index and the second model estimates the Remaining Life. The prediction performance is evaluated with the help of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and

coefficient of determination (R^2). Finally, the output of the model is analyzed by means of analysis of Actual vs. Predicted plots, Health Index vs. Remaining Life scatter plots, importance of the feature analysis, and correlation analysis which provides a clear understanding about the condition of the transformer and remaining service life [5].

Prediction of Health Index and Life Expectation

The proposed machine learning framework was implemented using Python in the Jupyter Notebook environment. Data preprocessing and model development were performed using the Pandas, NumPy, Scikit-learn, Matplotlib, and XGBoost libraries. The transformer oil dataset was preprocessed by removing missing values, eliminating outliers, and normalizing the input features using standard scaling. The dataset was divided into training and testing subsets using an 80:20 ratio to evaluate the predictive capability of the developed models.

```
In [1]: # ==========================================================
        # Transformer Health Index & RUL Prediction
        # Supervised Machine Learning
        # ==========================================================

        import pandas as pd
        import numpy as np

        from sklearn.model_selection import train_test_split
        from sklearn.model_selection import cross_val_score
        from sklearn.preprocessing import StandardScaler

        from sklearn.linear_model import LinearRegression
        from sklearn.tree import DecisionTreeRegressor
        from sklearn.ensemble import RandomForestRegressor
        from sklearn.ensemble import GradientBoostingRegressor
        from sklearn.neighbors import KNeighborsRegressor
        from sklearn.svm import SVR

        from sklearn.metrics import mean_absolute_error
        from sklearn.metrics import mean_squared_error
        from sklearn.metrics import r2_score

        import matplotlib.pyplot as plt

        # ---------------------------------------------------------
        # Load Dataset
        # ---------------------------------------------------------

        df = pd.read_csv("Health index1.csv")

        print(df.head())

        print(df.info())

        # ---------------------------------------------------------
        # Features and Targets
        # ---------------------------------------------------------

        X = df.drop(['Health index','Life expectation'],axis=1)

        y_health = df['Health index']

        y_rul = df['Life expectation']

        # ---------------------------------------------------------
        # Feature Scaling
        # ---------------------------------------------------------

        scaler = StandardScaler()

        X_scaled = scaler.fit_transform(X)

        # ---------------------------------------------------------
        # Train Test Split
        # ---------------------------------------------------------

        X_train, X_test, y_train, y_test = train_test_split(
            X_scaled,
            y_health,
            test_size=0.2,
            random_state=42)
```

Fig.2. Importing the Data Set and splitting it into Training and Test Sets

As per Fig.2. The transformer health dataset (Health index1.csv) was loaded into the Jupyter Notebook after using the Pandas library. The feature matrix (X) was formed after omitting the Health Index and Life Expectancy columns, whereas the Health Index and Life Expectancy were used as target variables. The input features were standardized by employing StandardScaler from the Scikit-learn library to remove variations due to the differences in feature scales and enhance the learning ability of the regression models. Then the standardized dataset was divided into training and testing sets using the train_test_split() function with a split of 80:20 and random state of 42. The training dataset was then used for building supervised machine learning models to predict the Transformer Health Index and Life Expectancy, while the testing dataset was used for the assessment of the forecasting performance of the models [6].

```
   Hydrogen  Oxigen  Nitrogen  Methane   CO   CO2  Ethylene  Ethane  \
0      2845    5860     27842     7406   32  1344     16684    5467
1     12886      61     25041      877   83   864         4     305
2      2820   16400     56300      144  257  1080       206      11
3      1099      70     37520      545  184  1402         6     230
4      3210    3570     47900      160  360  2130         4      43

   Acethylene  DBDS  Power factor  Interfacial V  Dielectric rigidity  \
0           7  19.0          1.00             45                   55
1           0  45.0          1.00             45                   55
2        2190   1.0          1.00             39                   52
3           0  87.0          4.58             33                   49
4           4   1.0          0.77             44                   55

   Water content  Health index  Life expectation
0              0          95.2              19.0
1              0          85.5              19.0
2             11          85.3              19.0
3              5          85.3               6.0
4              3          85.2               6.0
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 470 entries, 0 to 469
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Hydrogen             470 non-null    int64
 1   Oxigen               470 non-null    int64
 2   Nitrogen             470 non-null    int64
 3   Methane              470 non-null    int64
 4   CO                   470 non-null    int64
 5   CO2                  470 non-null    int64
 6   Ethylene             470 non-null    int64
 7   Ethane               470 non-null    int64
 8   Acethylene           470 non-null    int64
 9   DBDS                 470 non-null    float64
 10  Power factor         470 non-null    float64
 11  Interfacial V        470 non-null    int64
 12  Dielectric rigidity  470 non-null    int64
 13  Water content        470 non-null    int64
 14  Health index         470 non-null    float64
 15  Life expectation     470 non-null    float64
dtypes: float64(4), int64(12)
memory usage: 58.9 KB
None
```

Fig.3. Importing the Complete Data set of Transformer oil with all Parameters

Fig.3.demonstrated that the transformer oil dataset has been imported successfully into Python. The dataset is made up of 470 samples of transformer oil and consists of 16 attributes including dissolved gas concentrations (Hydrogen, Oxygen, Nitrogen, Methane, CO, CO_2, Ethylene, Ethane, and Acetylene), oil quality parameters (DBDS, Power Factor, Interfacial Tension, Dielectric Rigidity, and Water Content), and the target variables Health Index and Life Expectation. The figure shows the first few records of the dataset and the output of the info() function showing that all of the 470 entries are accurate with no missing values. The dataset contains both integer and floating-point features and has a total memory consumption of approximately 58.9 KB implying that it is very well organized and suitable for further processing [7].

Fig. 4 explains the procedure for training and evaluating machine learning models to assess transformer life expectancy based on transformer oil properties. From the database, the input features were determined by deleting the Life Expectancy and Health Index columns, while Life Expectancy served as the dependent variable. Input feature unification was achieved with the help of the StandardScaler. In addition, the data was split into two parts: 80% for training and 20% for testing purposes. Six different regression models were utilized to perform the training process: Linear Regression, Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbors (KNN) as well as Support Vector Regression (SVR). MAE, RMSE, as well as coefficient of determination (R^2) were employed to evaluate the estimated model performance. As it can be seen from the results attained by the experiments, Random Forest Regressor proved to be the most effective in terms of prediction performance because it had the lowest MAE (4.53), lowest RMSE (6.85), and the highest value of R^2 score (0.836). This way, it can be argued that Random Forest

is an effective tool for predicting the life expectancy of transformers using their oil features [8]

```
In [3]: X = df.drop(['Health index','Life expectation'],axis=1)

        y = df['Life expectation']

        X_scaled=scaler.fit_transform(X)

        X_train,X_test,y_train,y_test=train_test_split(
            X_scaled,
            y,
            test_size=0.2,
            random_state=42)

        for name,model in models.items():

            model.fit(X_train,y_train)

            pred=model.predict(X_test)

            mae=mean_absolute_error(y_test,pred)

            rmse=np.sqrt(mean_squared_error(y_test,pred))

            r2=r2_score(y_test,pred)

            print(name)

            print("MAE :",mae)

            print("RMSE :",rmse)

            print("R2 :",r2)

            print()
```

```
Linear Regression
MAE : 9.147274856429494
RMSE : 11.517209110803115
R2 : 0.5332995019694478

Decision Tree
MAE : 4.8446808510638295
RMSE : 9.521923192118058
R2 : 0.6809982188259096

Random Forest
MAE : 4.526867021276597
RMSE : 6.847790364416833
R2 : 0.835014919430064

Gradient Boosting
MAE : 4.983202677019718
RMSE : 7.242921287662739
R2 : 0.8154256740191393

KNN
MAE : 6.642340425531917
RMSE : 9.733356771564123
R2 : 0.6666741117184627

SVR
MAE : 8.040232629906463
RMSE : 10.932404152312785
R2 : 0.5794911726460553
```

Fig.4.MAE,RMSE and R2 for the Various Machine Learning Algorithms

Fig.5 demonstrates the feature importance analysis carried out using the Random Forest Regressor to find the most crucial transformer oil parameters for predicting the transformer lifespan. After the model was developed consisting of 200 decision trees, the feature importance values were generated using the feature_importances_ attribute and ordered in decreasing order. The corresponding bar chart illustrates the contribution of each input feature for prediction. The analysis shows that Interfacial Tension is the first important parameter with an importance score equal to 0.483, followed by Water Content (0.145), Power Factor (0.078), DBDS (0.057), and Hydrogen (0.047). On the contrary, the parameters such as Ethane, Acetylene, CO_2, and Ethylene have low importance values which indicates their lesser contribution to the detection of transformer lifespan. Thus, this analysis shows how the Random Forest model indicates the most important diagnostic oil parameters affecting the insulation condition of the transformer and the time left to work [9]

```
In [3]: X = df.drop(['Health index','Life expectation'],axis=1)

y = df['Life expectation']

X_scaled=scaler.fit_transform(X)

X_train,X_test,y_train,y_test=train_test_split(
    X_scaled,
    y,
    test_size=0.2,
    random_state=42)

for name,model in models.items():

    model.fit(X_train,y_train)

    pred=model.predict(X_test)

    mae=mean_absolute_error(y_test,pred)

    rmse=np.sqrt(mean_squared_error(y_test,pred))

    r2=r2_score(y_test,pred)

    print(name)

    print("MAE :",mae)

    print("RMSE :",rmse)

    print("R2 :",r2)

    print()
```

```
Linear Regression
MAE : 9.147274856429494
RMSE : 11.517209110803115
R2 : 0.5332995019694478

Decision Tree
MAE : 4.8446808510638295
RMSE : 9.521923192118058
R2 : 0.6809982188259096

Random Forest
MAE : 4.526867021276597
RMSE : 6.847790364416833
R2 : 0.835014919430064

Gradient Boosting
MAE : 4.983202677019718
RMSE : 7.242921287662739
R2 : 0.8154256740191393

KNN
MAE : 6.642340425531917
RMSE : 9.733356771564123
R2 : 0.6666741117184627

SVR
MAE : 8.040232629906463
RMSE : 10.932404152312785
R2 : 0.5794911726460553
```

Fig.5. Analysis of all Parameters related to the Machine Learning Algorithms

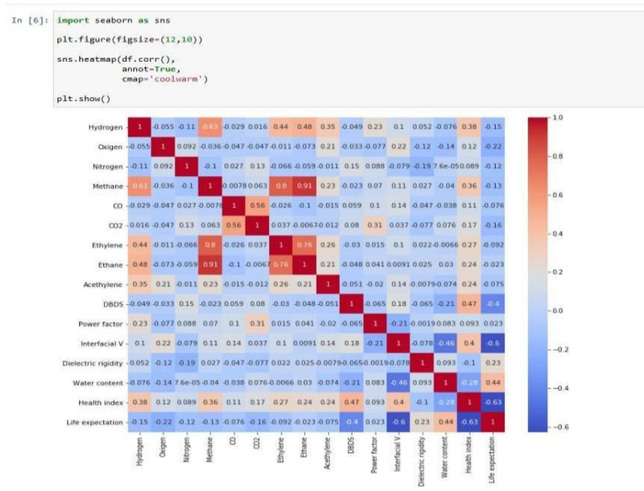

Fig.6. Correlation Map of the Transformer Oil Parameters

Fig.6 presents the correlation heatmap of the transformer oil data set has displayed the pairwise Pearson correlation coefficients among the input features and the target variable. The heatmap was generated using the Seaborn library, where the positive correlations have been shown with warmer colors and negative correlation with cooler colors that allow the identification of the linear relationships, feature dependencies, and multicollinearity within the data. The analysis shows strong positive correlations among the dissolved gas parameters such as Methane–Ethylene, CO–CO_2, and Ethylene–Ethane indicating their association with transformer insulation deterioration and faults. The Health Index has shown a strong positive correlation with the Life Expectation, meaning that transformers with a higher health index have a longer expected lifespan. Most other parameters show weak to moderate correlations indicating that each feature has its own information to provide with respect to the transformer's condition [10]

```
In [10]: import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.ensemble import AdaBoostRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.svm import SVR

from sklearn.metrics import r2_score

# -------------------------------------------------------
# Load Dataset
# -------------------------------------------------------
df = pd.read_csv("Health index1.csv")

# -------------------------------------------------------
# Features and Target
# -------------------------------------------------------
X = df.drop(["Health index","Life expectation"], axis=1)
y = df["Health index"]

# -------------------------------------------------------
# Feature Scaling
# -------------------------------------------------------
scaler = StandardScaler()
X = scaler.fit_transform(X)

# -------------------------------------------------------
# Train-Test Split
# -------------------------------------------------------
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.20,
    random_state=42
)

# -------------------------------------------------------
# Regression Models
# -------------------------------------------------------
models = {

    "Linear Regression": LinearRegression(),

    "Decision Tree": DecisionTreeRegressor(random_state=42),

    "Random Forest": RandomForestRegressor(
        n_estimators=200,
        random_state=42),

    "Gradient Boosting": GradientBoostingRegressor(
        random_state=42),

    "AdaBoost": AdaBoostRegressor(
        random_state=42),

    "KNN": KNeighborsRegressor(),

    "SVR": SVR()
}

# -------------------------------------------------------
# Model Evaluation
# -------------------------------------------------------
results = []

for name, model in models.items():

    model.fit(X_train, y_train)

    y_pred = model.predict(X_test)

    r2 = r2_score(y_test, y_pred)

    accuracy = r2 * 100

    results.append([

        name,

        round(accuracy,2),

        round(r2,4)
```

Fig.7. Accuracy for all the Machine Leaning Algorithms

```
# -------------------------------------------------------
# Display Results
# -------------------------------------------------------
results = pd.DataFrame(results,
                       columns=["Algorithm",
                                "Accuracy (%)",
                                "R2 Score"])

results = results.sort_values(by="Accuracy (%)",
                              ascending=False)

print(results)

results.to_csv("Algorithm_Comparison.csv", index=False)
```

	Algorithm	Accuracy (%)	R2 Score
2	Random Forest	74.61	0.7461
3	Gradient Boosting	73.27	0.7327
1	Decision Tree	56.43	0.5643
5	KNN	53.21	0.5321
0	Linear Regression	49.56	0.4956
4	AdaBoost	48.75	0.4875
6	SVR	41.96	0.4196

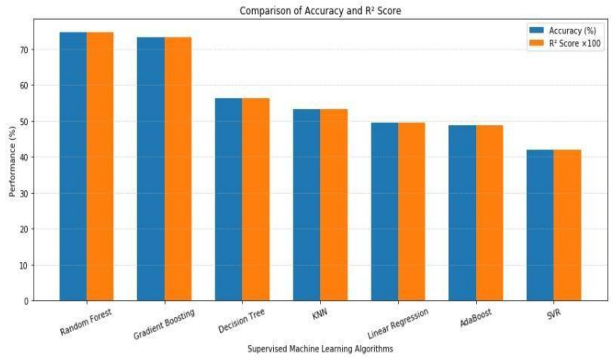

Fig.8. Accuracy Comparison of Different Supervised Machine Learning Algorithms

Fig.7&8 represents the use of various Artificial Intelligence methods to obtain and evaluate the quality of Transformer Health Index using the parameters of transformer oil. To begin, all relevant libraries of Python were used to carry out the Python code implementation, and the transformer oil data set was retrieved into the environment. Furthermore, since the Health Index column and Life Expectancy column were excluded from the data set, the input features were identified. After that, the feature data was standardized using the StandardScaler method, with the next step consisting of the train-test split of the data of 80/20 ratio for the unbiased evaluation. It is worth mentioning that the following 6 regression algorithms were utilized: Linear Regression, Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbors (KNN), and Support Vector Regression (SVR), while the determination coefficient (R^2) was used for both training and estimating, after which the prediction accuracy was considered to be equal to result obtained multiplied by 100. According to the results, the Random Forest Regressor achieved the highest prediction accuracy of 74.61% ($R^2 = 0.7461$), whereas Gradient Boosting got 73.27% ($R^2 = 0.7327$). Decision Tree and KNN were somewhat less effective with the prediction accuracies of 56.43% and 53.21%, respectively, whereas Linear regression, AdaBoost, and SVR were significantly less.

Fig.9. represents the prediction of Transformer Health Index and Life Expectation with the Random Forest Regressor using 5- fold cross-validation on the transformer oil dataset. Crossvalpredict() was used to obtain the predictions for each transformer in the dataset. The figure shows theactualversuspredicted values for bothtransformer healthindexandlife expectation for the first 15 transformers in the dataset and indicates a good approximation of these transformer quality parameters from transformer oil parameters by this regression technique

```python
In [25]: import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import KFold, cross_val_predict

df = pd.read_csv('Health index1.csv')
X = df.drop(columns=['Health index', 'Life expectation'])
y = df[['Health index', 'Life expectation']]

# Using 5-Fold Cross Validation for out-of-fold predictions across entire dataset
kf = KFold(n_splits=5, shuffle=True, random_state=42)
rf = RandomForestRegressor(n_estimators=100, random_state=42)

y_pred_cv = cross_val_predict(rf, X, y, cv=kf)

results_df = pd.DataFrame({
    'Row Index': df.index,
    'Actual Health Index': y['Health index'],
    'Predicted Health Index': np.round(y_pred_cv[:, 0], 2),
    'Actual Life Expectation': y['Life expectation'],
    'Predicted Life Expectation': np.round(y_pred_cv[:, 1], 2)
})

print("First 15 rows in exact dataset order:")
print(results_df.head(15).to_string(index=False))
```

```
First 15 rows in exact dataset order:
 Row Index  Actual Health Index  Predicted Health Index  Actual Life Expectation  Predicted Life Expectation
         0                 95.2                   67.29                     19.0                       16.68
         1                 85.5                   59.12                     19.0                       18.57
         2                 85.3                   62.82                     19.0                       16.22
         3                 85.3                   42.76                      6.0                       28.38
         4                 85.2                   59.21                      6.0                       15.20
         5                 75.6                   60.05                      6.0                       17.96
         6                 75.6                   55.60                      6.0                       18.76
         7                 73.2                   44.24                     19.0                       15.69
         8                 72.8                   44.07                      6.0                       17.29
         9                 68.0                   53.64                      6.0                        7.66
        10                 63.4                   49.62                      6.0                       18.64
        11                 63.4                   30.79                      6.0                       23.10
        12                 61.3                   50.57                      6.0                       13.25
        13                 60.5                   61.94                     19.0                       16.96
        14                 60.5                   60.37                     19.0                       19.09
```

Fig.9. Prediction of Transformer Health Index and Life Expectation

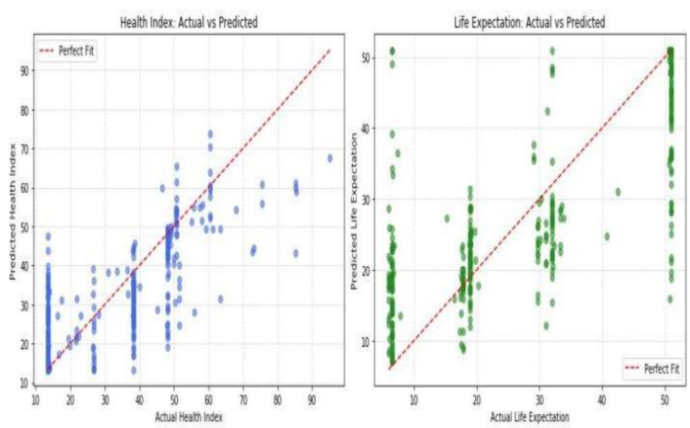

Fig.10. Scatter Plot of Actual and Predicted Health Index, Actual and Predicted Life Expectation

Fig.10 demonstrates the . Comparison of the predicted and actual values for the Transformer Health Index and Life Expectation via the Random Forest Regression model. The ideal prediction line is denoted with a dashed red line where the Predicted value matches the Actual value. In the Health Index plot, the majority of the points are situated closely on the ideal prediction line, meaning the model closely approximates the actual health index of the transformers with some exceptions. In addition, in the Life Expectation plot, the actual and predicted values align well across the different service ranges, showing a reliable predictive model. Slight deviations are visible as transformer degradation process is extremely complicated, but with all data points around the ideal prediction line, the Random Forest model accurately reflects how transformer oil features are related to the Health index and Life Expectation

TRANSFORMER REMAINING LIFE EXPECTATION BY HEALTH CONDITION

Condition_Category	Unit_Count	Avg_Health_Index	Actual_Avg_Life_Years	Predicted_Avg_Life_Years	Avg_Error_Years
1. Good Condition (Low HI)	285	14.43	42.11	41.07	5.01
2. Moderate Age (Medium HI)	163	44.67	19.78	21.06	6.47
3. Critical Fault (High HI)	22	69.74	13.13	18.21	6.54

INDIVIDUAL TRANSFORMER LIFE EXPECTATION CASE STUDIES

Health index	Condition_Category	Life expectation	Predicted_Life_Expectation_Years	Prediction_Error_Years
13.4	1. Good Condition (Low HI)	51.0	50.11	-0.89
51.5	2. Moderate Age (Medium HI)	29.1	36.02	6.92
48.2	2. Moderate Age (Medium HI)	17.9	17.82	-0.08
75.6	3. Critical Fault (High HI)	6.0	17.83	11.83
95.2	3. Critical Fault (High HI)	19.0	16.75	-2.25

Fig.11.Comparison of Actual and Predicted Transformer Life Expectancy for various Health Index Categories

From the above data, it clearly represents proposed machine learning framework that predicts Transformer HI and RLE from various transformer oil data. First, we import and preprocess a dataset containing 470 transformer oil samples. Then, the data is scaled and split into training (80%) and testing (20%) sets for model training. The trained models include Linear Regression, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, K-Nearest Neighbors (KNN), and Support Vector Regression (SVR).

They are tested with MAE, RMSE, and R score to identify the model with the highest prediction accuracy and lowest error. The highest performing model is Random Forest Regressor and it has the highest accuracy (R) and lowest error (MAE and RMSE). Analyzing the features revealed a group of most important transformer oil parameters (Interfacial Tension, Water Content, Power Factor, DBDS, Hydrogen) for HI prediction. The correlation heatmap between transformer oil parameters and confirms a high positive correlation between the Transformer Health Index (HI) and the Life Expectation of the transformer. Finally, the developed Random Forest Regressor model used for both HI and RLE prediction is verified using 5-fold cross-validation and shows the predicted values close to real values.

The scatter plots for actual against predicted values show the majority data points lying close to the ideal line with high accuracy. Finally, we define the transformer state as a "Good Condition", "Moderate Age", "Critical Fault" on basis ofHI valueand the forecasted remaining lifespan values concur with actual transformer lifespans. The developed Random Forest-based frame work proves to be an effective machine learning tool that could enable timely diagnostics and remaining life assessment of the transformer, offering new strategies for the predictive maintenance of power transformer assets and contributing positively to asset management and planning activities in the power industry.

4. Conclusion

In this, Machine learning approach is introduced to model Transformer Health Index (HI) and Remaining life (RL) of the transformer from its diagnostic oil parameters. An existing real world dataset containing 470 samples and 14 relevant diagnostic parameters is used for building regression models. The models are able to predict the condition of the transformer. An empirical evaluation was carried out to prove the applicability of ML approach. Based on the obtained experimental results, Random forest regression model out of othersupervised learning modelsperformed well by representing inherent nonlinearity existingbetween transformer oil attributes and target variables. It results in better predicted HI/RL values than others with higher value of R and minimum values forMAE and RMSE. By feature importance test, Interfacial Tension,Water Contents,Power Factor,DBDS andHydrogenparameters of transformer oil were found to be highly important predictors contributing to HI and RL prediction. Additionally,close match between observed and predicted HI/RL values demonstrated the viability of the presented framework. Thus, in addition to making predictions of HP and RL in an easy to implement and noninvasive manner, it assists in better operational risk management and provides reliable decision for the maintenance planning of transformer assets which leads to minimum accidental failure, extended asset service life, efficient performance of existing equipments and better for present electrical power systems performance

REFERENCES

1. A. Wen, B. Wen, K. Song, W. Liao, X. Wang and C. Zhang, "Research on Transformer Life Prediction Optimized by Artificial Intelligence Algorithm," 2025 8th International Conference on Energy, Electrical and Power Engineering (CEEPE), Wuxi, China, 2025, pp. 31-36, doi: 10.1109/CEEPE64987.2025.11033877.
2. P. Jarman, D. D. N. Gaunt and G. R. Morin, "Condition Assessment of Power Transformers Based on Health Index Value," Applied Sciences, vol. 9, no. 22, Art. no. 4877, Nov. 2019, doi: 10.3390/app9224877.
3. I. B. Taha, "Power Transformers Health Index Enhancement Based on Convolutional Neural Network After Applying Imbalanced-Data Oversampling," Electronics, vol. 12, no. 11, Art. no. 2405, 2023, doi: 10.3390/electronics12112405.
4. R. A. Prasojo, G. A. Sudrajad and Suwarno, "Health Index Prediction Using Artificial Neural Network (ANN) on Historical Data of Power Transformer," in 2023 IEEE 3rd International Conference in Power Engineering Applications (ICPEA), 2023, pp. 239–242.
5. M. Badawi, S. A. Ibrahim, D. E. A. Mansour, A. El-Faraskoury, S. A. Ward, K. Mahmoud, M. Lehtonen and M. M. F. Darwish, "Reliable Estimation for Health Index of Transformer Oil Based on Novel Combined Predictive Maintenance Techniques," IEEE Access, vol. 10,
6. pp. 25954–25972, 2022, doi: 10.1109/ACCESS.2022.3156102.
7. M. K. K. Alabdullh et al., "A Novel Method to Estimate the Lifetime of Mineral Oil-Type Power Transformers Based on the Analysis of Chemical and Physical Indicators Using Artificial Intelligence," Heliyon, vol. 10, no. 22, Art. no. e40447, Nov. 2024, doi: 10.1016/j.heliyon.2024.e40447.
8. S. T. Zahra, S. K. Imdad, S. Khan, S. Khalid and N. A. Baig, "Power Transformer Health Index and Life Span Assessment: A Comprehensive Review of Conventional and Machine Learning Based Approaches," Engineering Applications of Artificial Intelligence, vol. 139, Art. no. 109474, Jan. 2025, doi: 10.1016/j.engappai.2024.109474.
9. X. Zuo, Y. Li and Y. Zhang, "Calculation of a Health Index of Oil-Paper Transformers Insulation with Binary Logistic Regression," Mathematical Problems in Engineering, vol. 2016, Art. no. 6069784, 2016, doi: 10.1155/2016/6069784.
10. I. B. Taha, "Power Transformers Health Index Enhancement Based on Convolutional Neural Network After Applying Imbalanced-Data Oversampling," Electronics, vol. 12, no. 11, Art. no. 2405, 2023, doi: 10.3390/electronics12112405.
11. A. Nezami, M. Vakilian and M. Fardmanesh, "A Novel Artificial Neural Network Approach for Residual Life Estimation of Paper Insulation in Oil-Immersed Power Transformers," IET Electric Power Applications, vol. 18, no. 4, 2024, doi: 10.1049/elp2.12407