Health Assessment and Lifespan Estimation for Power Transformer using Machine Learning

In partial fulfillment for the award of the degree Of Master of Technology In
'Electrical Power System'

Dissertation Phase-II

Submitted by

Ajaan Anubhav Borah
Roll *no: 220620064001*
Reg No: 001306222

Under the guidance of

Mrs. Ritu Nazneen Ara Begum
(Assistant Professor)
Dr. Barnali Goswami
(Professor)

Department of Electrical Engineering
Assam Engineering College
Assam Science and Technology University
Jalukbari, Guwahati

Assam-781013

Health Assessment and Lifespan Estimation for Power Transformer using Machine Learning

In partial fulfillment for the award of the degree Of Master of Technology In
'Electrical Power System'

Dissertation Phase-II

Submitted by
Ajaan Anubhav Borah
Roll *no: 220620064001*
Reg No: 001306222

Under the guidance of
Mrs. Ritu Nazneen Ara Begum
(Assistant Professor)
Dr. Barnali Goswami
(Professor)

Department of Electrical Engineering
Assam Engineering College
Assam Science and Technology University
Jalukbari, Guwahati

Assam-781013

DECLARATION

I hereby certify that the work contained in this thesis is original and has been done by me under the guidance of my supervisor(s). The work has not been submitted to any other Institute for any degree or diploma. I have followed the guidelines provided by Assam Science and Technology University, Guwahati in preparing the report. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the University. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the report and giving their details in the references. Further. I have taken permission from the copyright owners of the sources, whenever necessary.

Date: (Signature of the Student)

Place: Guwahati

Name of the Co-Supervisor: **Dr. Barnali Goswami**

Designation: **Professor**

Address: **Assam Engineering College**

CERTIFICATE-I

This is to certified that the dissertation entitled *"**Health Assessment and Lifespan Estimation for Power Transformer using Machine Learning*** "submitted to the faculty of Department of Electrical Engineering, Assam Engineering College, Assam Science and Technology University, Guwahati as a pre-requisite for the degree of Master of Technology in Electrical Power System is a record of research work carried out by Shri Ajaan Anubhav Borah Registration No 220620064001 under my personal supervision and guidance.

All help received by him has been duly acknowledged. No part of the thesis has been reproduced elsewhere to any degree.

Place: (Signature of the Co- Supervisor)

Date:

Name of the Supervisor: **Mrs. Ritu Nazneen Ara Begum**

Designation: **Assistant Professor**
Address: **Assam Engineering College**

CERTIFICATE-II

This is to certified that the dissertation entitled *"**Health Assessment and Lifespan Estimation for Power Transformer using Machine Learning "submitted*** to the faculty of Department of Electrical Engineering, Assam Engineering College, Assam Science and Technology University, Guwahati as a pre-requisite for the degree of Master of Technology in Electrical Power System is a record of research work carried out by Shri Ajaan Anubhav Borah Registration No 220620064001 under my personal supervision and guidance.

All help received by him has been duly acknowledged. No part of the thesis has been reproduced elsewhere to any degree.

Place: (Signature of the Supervisor)

Date:

CERTIFICATE-III

FORWARDING OF APPROVAL

This is to certify that the thesis entitled, " ***Health Assessment and Lifespan Estimation for Power Transformer using Machine Learning*** " submitted by Mr. Ajaan Anubhav Borah, Registration No 220620064001 to Assam Engineering College, Assam Science and Technology University, Assam, is a record of bonafide project work carried out by him. The dissertation has been duly examined and conducted viva-voce by the External Examiner and approved for the award of the degree of Master of Technology in the discipline of Electrical Power System. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed, or conclusion drawn therein but approve only for the purpose for which it is submitted.

Name and of External Examiner:
Designation:
Signature:
Date:

Name of Head of Department: Dr Runumi Sarma
Signature:
Department of Electrical Engineering
Assam Engineering College
Date:

ABSTRACT

This project explores the application of machine learning algorithms for the accurate detection of health indices in power transformers, a crucial undertaking for ensuring the reliability and longevity of these vital components within electrical power systems. Accurate health index assessment enables timely maintenance actions, preventing unexpected failures and potential disruptions to power supply. This research employs a diverse set of algorithms, including Elastic Net, Support Vector Regression (SVR), Random Forest, and Gradient Boosting, to model and predict the health index, offering a multi-faceted approach to health assessment. The performance of these algorithms is rigorously evaluated based on the R-squared metric, enabling a robust comparison of their predictive capabilities and identifying the most suitable approach for this application.

To provide a comprehensive understanding of health index profiles across various transformer features, this project meticulously analyzes the health index values in relation to the diverse features available within the transformer datasheet. This comprehensive analysis reveals intricate relationships between transformer health and its various characteristics, offering valuable insights into the key factors that influence transformer health and potential deterioration. Furthermore, to visualize these relationships and facilitate the identification of prominent feature interactions, a heatmap is constructed, encompassing all features within the dataset. The heatmap grants a visual representation of feature correlations and their potential impact on transformer health, serving as a valuable tool for further analysis and interpretation.

Building upon the individual capabilities of the explored algorithms, this project further investigates the implementation of a stacking ensemble model. This model seeks to combine the strengths of the previously analyzed algorithms (Linear Regression, SVR, Random Forest, and Gradient Boosting) through a hierarchical learning process. By leveraging the diverse predictive approaches of these models, the stacking ensemble has the potential to achieve superior accuracy and robustness compared to any single model alone.

The findings of this project offer significant contributions to the field of power transformer health assessment. By employing and comparing multiple machine learning algorithms,

this research establishes a robust framework for health index prediction, aiding in the development of effective maintenance strategies. Additionally, the in-depth analysis of health index profiles in relation to transformer features uncovers critical insights into the factors influencing transformer health, enabling a more comprehensive understanding of transformer degradation processes. The construction of a comprehensive heatmap further visualizes these relationships, facilitating data interpretation and fostering the identification of key feature interactions. Ultimately, the outcomes of this project have the potential to significantly enhance the reliability and longevity of power transformers, bolstering the resilience of electrical power systems and minimizing the likelihood of unexpected failures.

ACKNOWLEDGEMENT

With immense gratitude, I acknowledge the invaluable guidance and support that made this project possible. Firstly, I express my deepest appreciation to my esteemed Head of Department, Prof Runumi Sarma, for providing a nurturing academic environment and fostering my research interests. My sincerest thanks extend to my esteemed guide, Prof Ritu Nazneen Ara Begum, whose mentorship proved instrumental in shaping this project's direction and pushing me to achieve my full potential. I am also deeply indebted to my co-guide, Prof Barnali Goswami, whose expertise and insights were invaluable in navigating the intricacies of my research journey. Finally, I offer heartfelt thanks to everyone who assisted me along the way – from professors and classmates who offered their time and knowledge to my friends and family who provided unwavering support and encouragement. This project would not have come to fruition without the collective contributions of each and every one of you. I am truly grateful for your generosity and kindness.

With regards
Ajaan Anubhav Borah
AEC, Guwahati

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

1.1 Motivation

Maintaining a robust and reliable power grid necessitates the precise and proactive management of its critical components, particularly power transformers. These giants of electrical infrastructure silently convert and elevate voltage levels, enabling efficient power transmission and distribution. However, their complex internal workings are susceptible to gradual degradation due to factors like ageing, thermal stress, and overloading. Early detection of these subtle declines in transformer health is crucial for preventing catastrophic failures, ensuring grid stability, and optimizing maintenance schedules.

Traditional methods of transformer health assessment rely on periodic manual inspections and offline diagnostic tests. These techniques, while valuable, suffer from limitations like being resource-intensive, time-consuming, and often providing a snapshot picture rather than continuous health monitoring [2] [7]. This is where the transformative power of machine learning (ML) steps in.

This research project aims to leverage the potential of ML algorithms like, Elastic net, Support Vector Regression (SVR), Random Forest, and Gradient Boosting also an ensemble of these four to establish a robust and dynamic system for power transformer health assessment and lifespan estimation. By feeding historical and real- time operational data from various sensors and measurements into these algorithms, we can unlock hidden patterns and correlations within the complex tapestry of transformer health indicators. These data points might include oil analysis results, vibration measurements, load profiles, and environmental parameters.

By building predictive models and generating a dynamic "health index" – a quantifiable metric of the transformer's current condition – we can move beyond reactive maintenance towards proactive interventions. The ML algorithms will learn to identify subtle deviations from healthy operating parameters, potentially signaling incipient faults or impending failures. This early warning system empowers grid operators to schedule timely maintenance actions, minimize downtime, and prevent costly repairs, all while maximizing the transformer's operational lifespan.

Furthermore, by comparing the performance of different ML algorithms like SVR, Random Forest, and Gradient Boosting based on metrics like R-squared, we can glean valuable insights into their relative strengths and weaknesses in the context of transformer health assessment. This comparative analysis will not only guide the selection of the optimal algorithm for our specific application but also contribute to the broader field of power grid asset management by highlighting the efficacy of different ML approaches.

This research project aspires to bridge the gap between traditional transformer health assessment techniques and the cutting-edge power of ML. By deploying these intelligent algorithms, we can pave the way for a future where power transformers become self-aware, continuously communicating their health status, and enabling a predictive, data-driven approach to grid management. This transition promises not only enhanced reliability and efficiency but also increased sustainability and resilience of the critical infrastructure that underpins our modern world.

1.2. OBJECTIVES

I. To develop a Machine Learning-based system for accurate health assessment of power transformers:
II. To enable proactive maintenance through lifespan estimation:
III. To advance the field of power grid asset management through comparative analysis.
IV. To visualize the dynamic health profile of transformers:

By achieving these objectives, this research project aims to empower grid operators with a data-driven approach to transformer health assessment, enabling proactive maintenance, optimizing lifespan, and contributing to a more stable and sustainable power grid.

1.3. Literature Review

Maintaining the health and longevity of power transformers is crucial for ensuring grid stability and efficient power delivery. Traditionally, this has been achieved through

periodic inspections and offline diagnostic tests. However, these methods suffer from limitations in timeliness, resource intensiveness, and often provide a static snapshot rather than continuous health monitoring.

The emergence of machine learning (ML) [1] presents a transformative opportunity to overcome these limitations and establish a more proactive and data-driven approach to transformer health assessment. This literature review explores the existing research landscape in applying ML to transformer health assessment and lifespan estimation, with specific focus on your project's chosen algorithms: Support Vector Regression (SVR), Random Forest, and Gradient Boosting.

Current Approaches:

A. Dissolve Gas Analysis:

Transformers are the workhorses of the power grid, efficiently transmitting electrical energy over long distances. However, their insulating oil degrades over time due to various factors like partial discharges and overheating. These degradation processes liberate distinct gaseous byproducts that dissolve in the oil. DGA is a well-established technique that analyzes the composition and quantity of these dissolved gases to diagnose internal faults within transformers [2] [3].

By meticulously extracting a small oil sample from the transformer, technicians can identify the presence and concentration of various gases, including hydrogen (H_2), methane (CH_4), ethylene (C_2H_4), and acetylene (C_2H_2). Each gas type points towards a specific type of fault, such as low-energy discharges (H_2), incipient faults (CH_4), and arcing (C_2H_2). By interpreting the gas signature, DGA empowers maintenance personnel to pinpoint the nature and severity of the internal fault, enabling timely intervention and preventing catastrophic failures [4].

While DGA is a powerful tool, interpreting the gas ratios and correlating them to specific faults can be challenging, especially for incipient faults with subtle gas signatures. Here's where Machine Learning (ML) enters the scene. Researchers have successfully integrated various ML algorithms, such as Artificial Neural Networks

(ANNs) and Support Vector Machines (SVMs), with DGA data. These algorithms are trained on historical data containing gas compositions and corresponding fault types. Once trained, the ML models can analyze new DGA data points and provide more accurate fault classification compared to traditional methods.

Beyond fault diagnosis, the future holds immense promise for using ML-enhanced DGA for transformer health prognosis [5] [6]. By incorporating historical DGA trends and operational data into the ML models, researchers are exploring methods to predict the remaining useful life of transformers. This predictive capability would revolutionize transformer maintenance practices, enabling maintenance teams to prioritize interventions based on equipment health and prevent unnecessary outages.

B. Partial discharge (PD) Detection:

Partial discharge (PD) detection plays a crucial role in preventative maintenance for high-voltage equipment, particularly transformers [7][8]. These partial discharges are localized breakdowns within the insulation system, often caused by imperfections, aging, or contamination. Early detection of PD activity is essential because it signifies incipient (beginning) degradation of the insulation, which can eventually lead to catastrophic failure. By monitoring PD events, maintenance personnel can identify potential problems and schedule repairs before a critical breakdown occurs.

Machine learning algorithms have emerged as powerful tools for analyzing PD data and assessing insulation health. One such approach, presented in [9], utilizes a Random Forest classifier to categorize PD patterns. This allows for not only the identification of PD occurrence but also the estimation of the severity of internal faults within the transformer. This capability provides valuable insights for prioritizing maintenance actions and optimizing equipment lifespan

C. Operational Data Analysis:

Traditionally, maintenance schedules were based on fixed intervals. However, advancements in data acquisition and analytics have opened doors to a more proactive approach. Operational data analysis offers a treasure trove of insights into

a transformer's health, enabling us to move towards condition-based maintenance (CBM) [10].

By analyzing data streams like load profiles [11], vibration measurements, and environmental parameters, we can gain a comprehensive understanding of a transformer's behavior under real-world operating conditions. This data can be leveraged for various health assessment tasks, including [12]:

- **Early Fault Detection:** Deviations from normal operating patterns in parameters like vibration or temperature can indicate emerging issues, allowing for timely intervention before they escalate into catastrophic failures.
- **Remaining Useful Life (RUL) Estimation:** Machine learning algorithms can be trained on historical data to predict the remaining lifespan of a transformer with greater accuracy compared to traditional methods. This empowers informed decisions regarding maintenance or replacement strategies.
- **Root Cause Analysis:** By correlating operational data with observed anomalies, engineers can pinpoint the root cause of transformer problems, facilitating targeted repairs and preventing future occurrences.

Gaps and Opportunities:

- While existing research demonstrates the potential of ML for transformer health assessment, there is a need for further exploration of multi-algorithm comparisons and feature selection techniques. The comparative analysis of SVR, Random Forest, and Gradient Boosting using R-squared as a metric addresses this gap in the project.
- Integrating diverse data sources like DGA operational, and environmental data into a single health index framework presents an opportunity for comprehensive health assessment. This project focuses on developing a dynamic health index based on multiple features aligns with this direction.
- Visualizing the health index profile over time for individual transformers can provide valuable insights and facilitate trend analysis. So, here main objective is to create graphical representations of the health index addressing this need for visual interpretation.

This literature review highlights the promising potential of ML for enhancing power transformer health assessment and lifespan estimation. In this project Elastic Net, SVR, Random Forest, and Gradient Boosting is used to develop a dynamic health index with visual representations, address key gaps in existing research and contribute to a more data-driven approach to transformer management.

Chapter 2: Methodology

2.1 The Process

This method leverages data readily available from platforms like Kaggle, making it a potentially cost-effective and data-driven solution. Here is a breakdown of the key steps involved:

1. Data Pre-processing:

- Environment Setup:
    - Anaconda: A popular platform for scientific computing that provides a user-friendly interface and pre-installed libraries.
    - pandas: A powerful Python library for data manipulation and analysis. It excels at handling large datasets efficiently.
    - matplotlib & seaborn: These libraries create informative visualizations like heatmaps and scatterplots to explore the data.
    - scikit-learn: A comprehensive machine learning library offering a variety of algorithms for training and evaluating models.
- Data Loading and Exploration:
    - The Kaggle dataset is loaded into a panda Data Frame, a structured format ideal for data analysis.
    - The structure, data types (e.g., numerical, categorical), and summary statistics (e.g., mean, standard deviation) of the data are examined to understand its characteristics.
- Data Cleaning:
    - Techniques like removing rows with missing values or creating custom functions are employed to address missing data points that could negatively impact the model's performance.
    - Outliers, which are data points that deviate significantly from the majority, are also identified, and handled appropriately. This might involve removing them if justified or transforming them to minimize their influence.

In the initial exploration of the dataset, boxplots proved invaluable in revealing outlying data points for several key features. The analysis unveiled clear thresholds beyond which data could be considered anomalous. For instance, hydrogen values above 15,000, oxygen exceeding 40,000, and nitrogen surpassing 80,000 all stood out as potential outliers. Similarly, methane values above 2,000, carbon monoxide exceeding 1,250, and carbon dioxide exceeding 15,000 raised eyebrows. Further, elevated levels of ethylene (beyond 7,500), ethane (beyond 3,000), and acetylene (beyond 6,000) demanded attention. Even the non-chemical features hinted at outliers, with power factor exceeding 30 and water content surpassing 100 joining the list. Notably, the "health index" itself exhibited outliers, with values above 80 raising concerns.

However, simply identifying these outliers wasn't enough. Understanding their origins and potential impact on the analysis became crucial. Were these isolated incidents or indicative of larger trends? Did they represent measurement errors, inherent variabilities, or anomalies unique to specific sub-groups within the data? Delving deeper to answer these questions became paramount. Ultimately, after careful consideration and analysis, I opted to clean the data by dropping these outlying points. This decision was not taken lightly, and the potential loss of valuable information was not ignored.

In the face of significant discrepancies and the potential for skewing the analysis, removing these outliers appeared to be the most judicious course of action. This data cleaning step paved the way for a more robust and reliable analysis, ensuring the remaining data points formed a cleaner foundation for drawing meaningful conclusions from the research. While the cleaned data undoubtedly offered a clearer picture, the discarded outliers remained an intriguing facet of the dataset. Recognizing their existence and potential influence prompted further investigations into their source and characteristics. These additional explorations could ultimately yield valuable insights, even if they weren't incorporated into the main analysis. After all, sometimes understanding the fringes can shed light on the core, and the story of the outliers, though separate, could still contribute to the richness and complexity of the overall narrative.

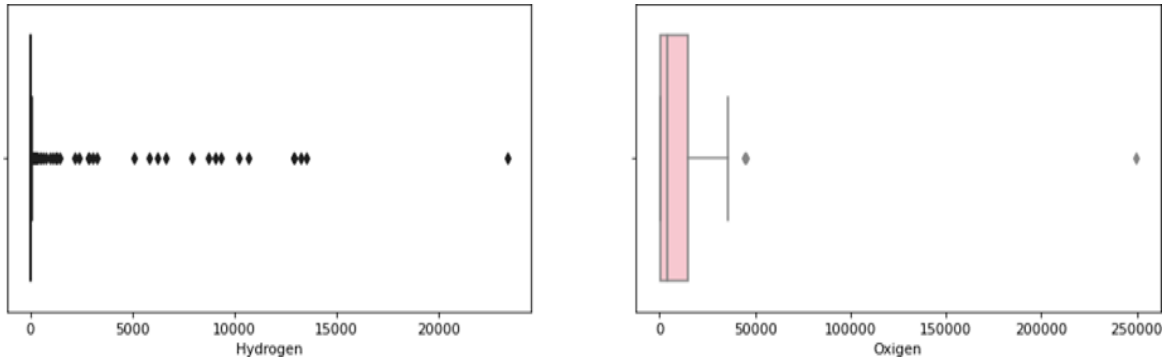

Figure (a) : Outliers of Hydrogen and Oxygen

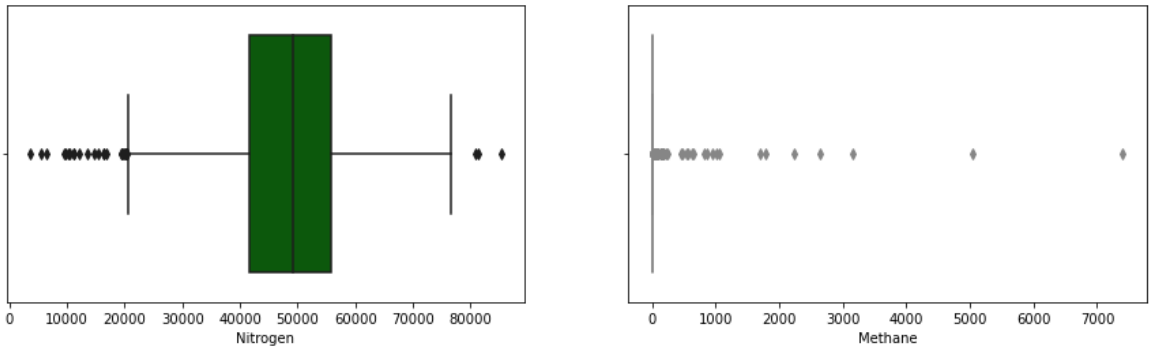

Figure (b): Outliers of Nitrogen and Methane

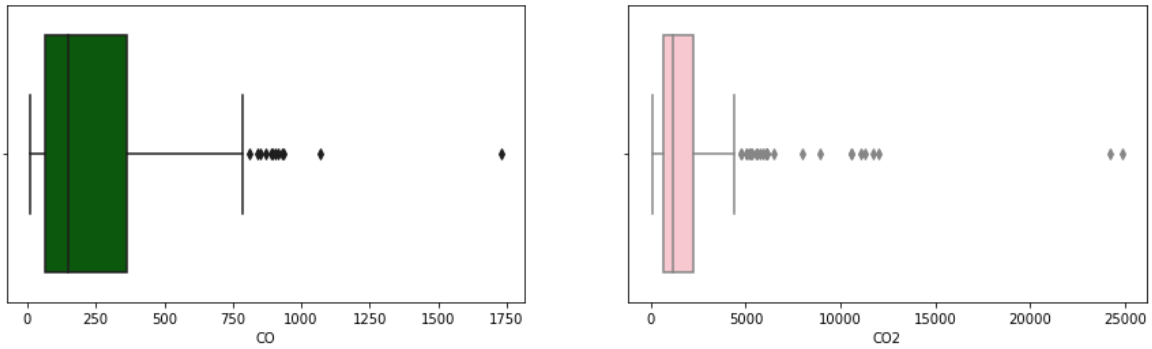

Figure (c): Outliers of CO andCO_2

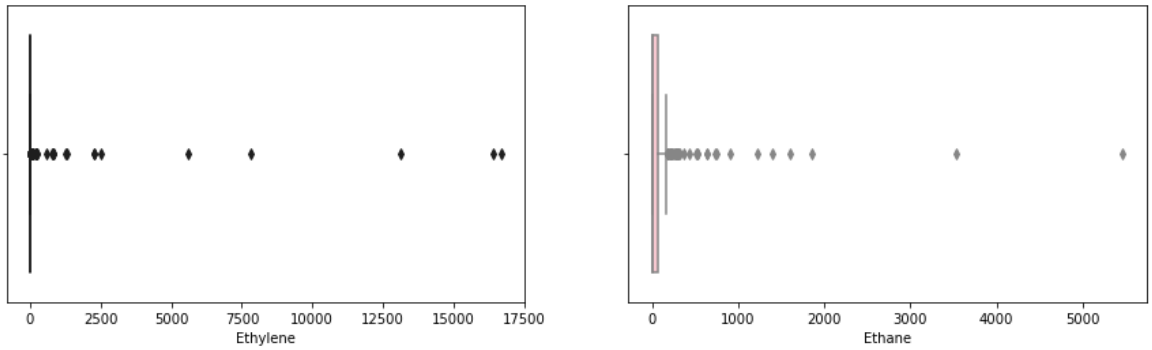

Figure (d): Outliers of Ethylene and Ethane

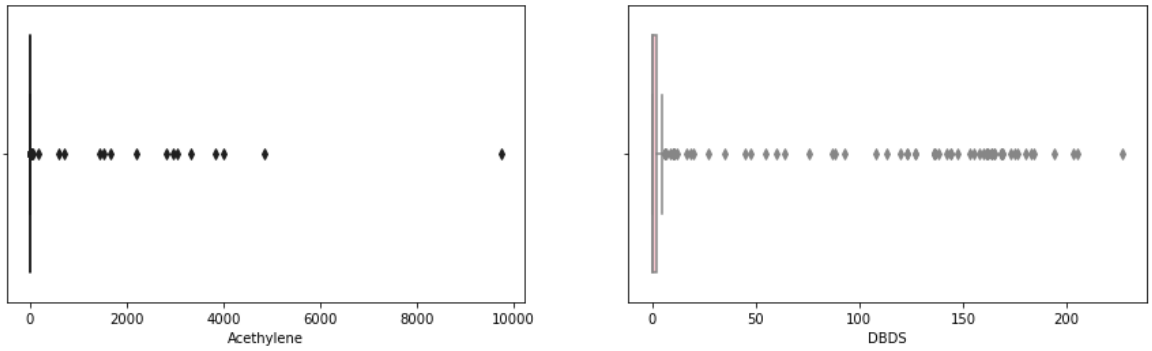

Figure (e): Outliers of Acetylene and DBDS

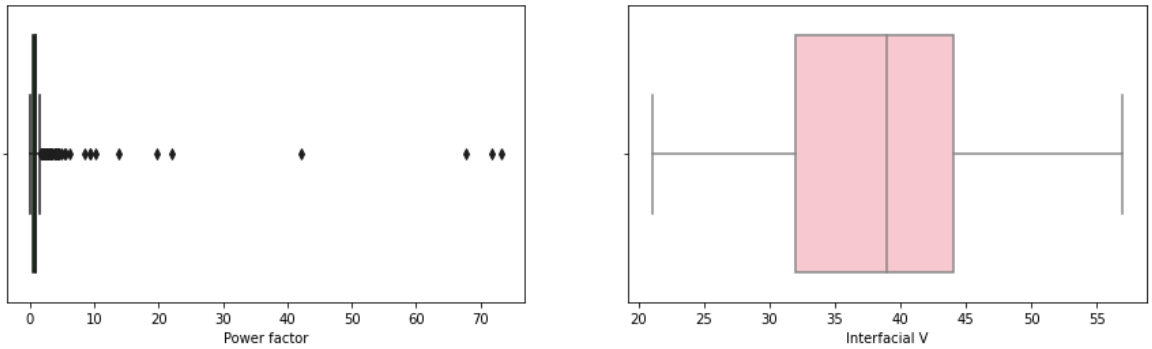

Figure (f): Outliers of PF and Methane

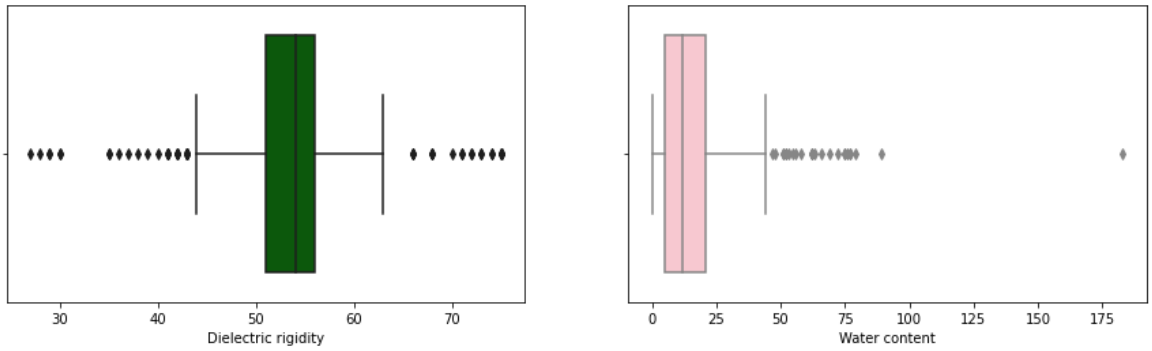

Figure (g): Outliers of Dielectric rigidity and water content

Figure 1 (a- g): Detection of Outliers of all the features

2. Exploratory Data Analysis (EDA):

- Visualization:
    - Heatmaps are created to visually represent the correlation between different features in the dataset. Correlations indicate how changes in one feature might be related to changes in another.
    - Scatterplots are used to delve deeper into the relationship between individual features and the target variable, which is likely the health index of the transformer. Analyzing these plots helps to understand how specific factors (features) influence the overall health of the transformer.
- Feature Scaling:
    - Feature scaling ensures all features are on a similar scale. This is crucial for many machine learning algorithms as they might be biased towards features with larger values. Standardization is a common technique that transforms each feature to have a mean of 0 and a standard deviation of 1. This levels the playing field for all features and allows the model to focus on the underlying relationships between them and the target variable.

3. Model Training and Evaluation:

- Data Splitting:

- The pre-processed data is divided into two sets: a training set and a testing set. The training set is used to train the machine learning model, while the testing set is used to evaluate its performance on unseen data. This helps to prevent overfitting, where the model simply memorizes the training data and performs poorly on new data.

- Model Selection and Training:
  - Different machine learning models, such as Elastic Net and random forest regression, are evaluated.
  - Elastic Net is a regularized regression model that combines L1 and L2 regularization, potentially leading to improved model performance and interpretability.
  - Random forest regression is an ensemble learning method that combines the predictions of multiple decision trees, leading to robustness and potentially higher accuracy.
  - Other models might also be considered based on the specific characteristics of the data and the desired outcome.
- Model Evaluation:
  - The performance of each model is evaluated on the testing set using metrics like R-squared. R-squared indicates the proportion of variance in the target variable (health index) that can be explained by the model. A higher R-squared value signifies a better fit and superior prediction capability.

- Model Selection:
  - Based on the evaluation metrics, the model with the highest R-squared score is chosen as the best model for predicting the health index of power transformers.

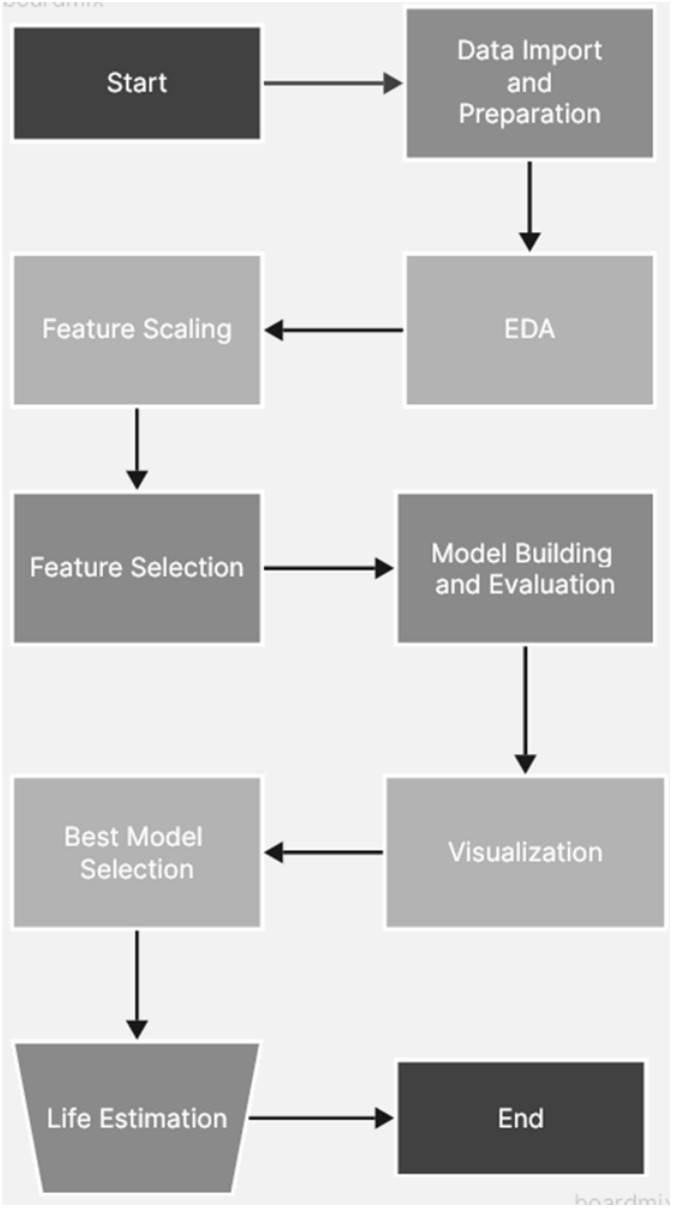

Figure 2: Flowchart of the Process

4. Remaining Lifespan Estimation:

- Literature Review:
    - Existing research and established industry knowledge are consulted to identify the relationship between the predicted health index and the remaining lifespan of transformers. This relationship might be represented

by formulas or degradation curves that map the health index to the expected remaining operational life of the transformer.

- Lifespan Prediction:
    - By leveraging the chosen machine learning model to predict the health index for a specific transformer, and then using the established relationship between health index and remaining lifespan (obtained from the literature review), the remaining lifespan of that transformer can be estimated.

This approach offers a data-driven and potentially more efficient way to assess transformer health and predict their remaining lifespan. By proactively identifying transformers with declining health, maintenance schedules can be optimized, preventing unexpected failures and associated downtime costs.

2.2. ML Models

I. ***Elastic Net: Combining Strengths for Feature Selection and Regularization***

Elastic Net, a regression technique that builds upon linear regression by incorporating two regularization methods: L1 and L2 [13]. Let's break down what this means:

- *Linear Regression***:** This is a statistical method that models the relationship between a dependent variable (what you're trying to predict) and one or more independent variables (the factors influencing the prediction) [14]. It essentially creates a best-fitting straight line through the data points.
- *Regularization:* In machine learning, this is a process that helps prevent a model from overfitting the training data [13]. Overfitting occurs when a model becomes too specific to the training data and performs poorly on unseen data. Regularization techniques add constraints to the model, reducing its flexibility and complexity.

The Power of L1 and L2 :

- *L1 Regularization (Lasso)***:** This technique penalizes the model for the absolute values of its coefficients (the weights assigned to each independent variable). A larger coefficient value indicates a stronger influence of that variable on the

prediction. L1 regularization encourages sparsity, meaning it can drive some coefficient values to zero. This essentially removes those features (independent variables) from the model, helping to identify irrelevant ones [16].

- *L2 Regularization (Ridge)***:** This technique penalizes the model for the sum of squared coefficients. It discourages large coefficient values, pushing them towards zero but not necessarily eliminating them completely. L2 regularization helps to shrink the coefficients, reducing the overall complexity of the model and preventing overfitting [17].

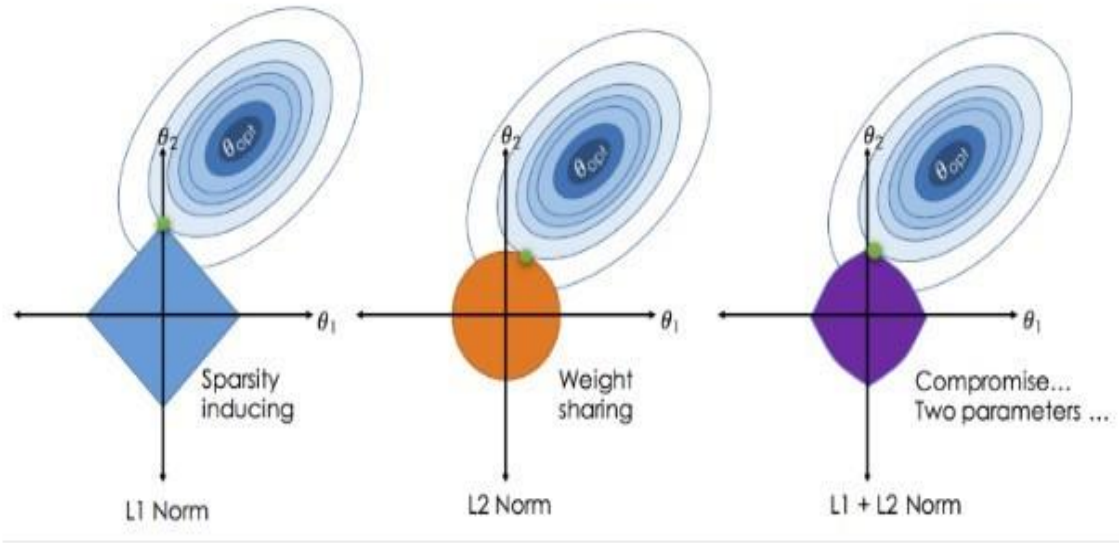

Figure 3: Elastic Net [18]

Elastic Net: Striking a Balance: Elastic Net combines L1 and L2 regularization. The parameters l1_ratio and alpha control the relative influence of each penalty:

- *l1_ratio (set to 0.8):* This value prioritizes feature selection. A higher l1_ratio strengthens the L1 penalty, making it more likely for coefficients to become zero and features to be dropped. In this case, 0.8 indicates a strong focus on identifying and removing irrelevant features.

- *alpha (set to 0.5)***:** This value controls the overall regularization strength. A higher alpha increases the combined effect of L1 and L2 penalties, leading to a more shrunken and less flexible model. Here, 0.5 suggests a moderate level of overall regularization while still allowing for some feature selection driven by the L1 penalty.

Elastic Net (model1) aims to achieve the following:

a) *Feature Selection:* By leveraging L1 regularization, it identifies and removes irrelevant features from the model, potentially improving interpretability and reducing complexity.

b) *Regularization*: It combines L1 and L2 penalties to prevent overfitting and enhance the model's ability to generalize unseen data.

c) *Tuned Parameters:* The chosen values for l1_ratio and alpha prioritize feature selection while maintaining a moderate level of overall regularization.

II. Random Forest Regressor: Power in Numbers

Random Forest Regressor, a regression technique that leverages the collective power of multiple decision trees. Let's delve into its key aspects:

- *Ensemble Method**:*** Random Forest doesn't rely on a single model; instead, it creates a group (ensemble) of decision trees. Each tree makes a prediction, and the final output is typically an average (for regression) or majority vote (for classification) of the individual predictions. This ensemble approach helps reduce variance and improve the overall accuracy and robustness of the model [19].

- *Decision Trees**:*** These are tree-like structures where data is split based on certain conditions (usually rules involving the independent variables) at each node. The splits continue until the data at a particular node (leaf) is sufficiently homogeneous regarding the dependent variable (what you're trying to predict).

Advantages for Regression Tasks:

- *Robust to Outliers*: Random Forests are less susceptible to the influence of outliers (extreme data points) compared to some other regression methods. This is because individual trees might be affected by outliers, but averaging or voting across many trees reduces their impact on the final prediction [20].
- *Effective in High Dimensions*: Random Forests can handle datasets with a high number of independent variables (features) efficiently [21 . During tree construction, only a random subset of features is considered at each split point, which helps prevent overfitting and improves performance in high-dimensional settings.

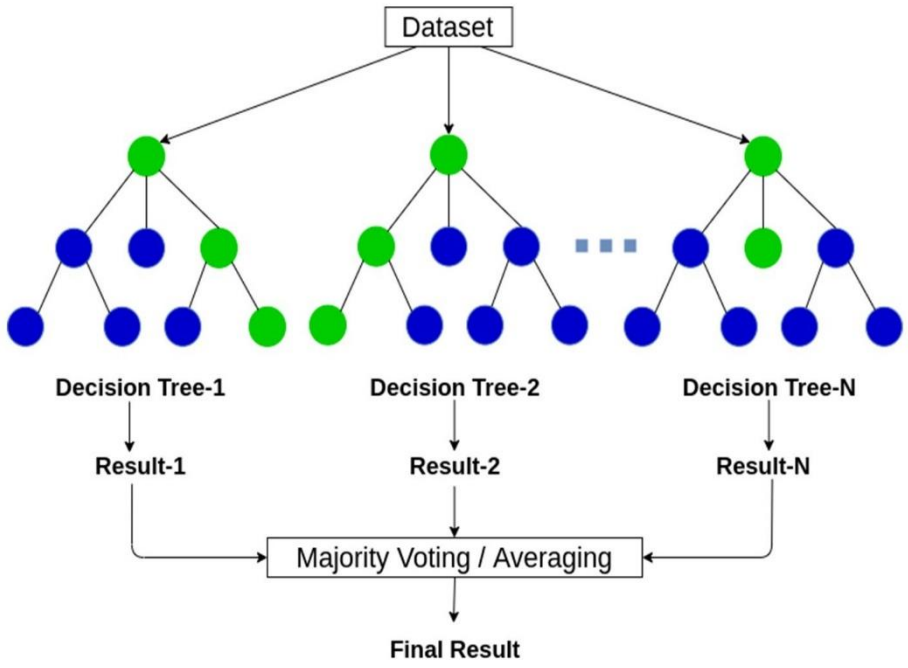

Figure 4: Random Forest Regressor [22]

Balancing Complexity and Generalizability:

- *estimators (set to 150)*: This parameter controls the number of decision trees to build in the forest. A higher number generally leads to lower variance (more stable

predictions) but can also increase model complexity. Here, 150 suggests a moderate number of trees, balancing accuracy with efficiency.

- *Max depth (set to 5*): This parameter restricts the maximum depth (number of splits) allowed in each tree. Deeper trees can capture more complex relationships but are also more prone to overfitting. Setting max_depth to 5 promotes simpler trees, reducing complexity and potentially improving generalizability (performance on unseen data).

- *random state*: This parameter injects randomness into the tree creation process. Fixing random state ensures that the same ensemble of trees is generated every time the model is run, leading to reproducible results. This is helpful for debugging, comparing different model configurations, and sharing your work with others.

Thus, Random Forest Regressor (model2) constructs a robust ensemble of decision trees to make predictions. It's advantageous for handling outliers, managing high-dimensional data, and achieving a balance between model complexity and generalizability through carefully chosen parameters. This method is a popular choice for regression tasks due to its accuracy and robustness.

III. ***Support Vector Regression: Capturing Complexities***

Support Vector Regression (SVR), a technique that tackles regression tasks, particularly those involving non-linear relationships. Here's a breakdown of its key features:

- *Non-Linear Mapping (Kernel Function):* Unlike linear regression, SVR can learn non-linear relationships between independent variables (features) and the dependent variable (what you're trying to predict). It achieves this by employing a kernel function [23]. This function essentially transforms the data points from their original space into a higher-dimensional space where linear relationships might become more evident.

- *Radial Basis Function (RBF Kernel):* This is a specific type of kernel function commonly used in SVR. It's known for its flexibility in handling various types of non-linear relationships. Imagine the data points being projected onto a sphere in

higher-dimensional space. The RBF kernel considers the distance between points on this sphere to determine their influence on the model [24].

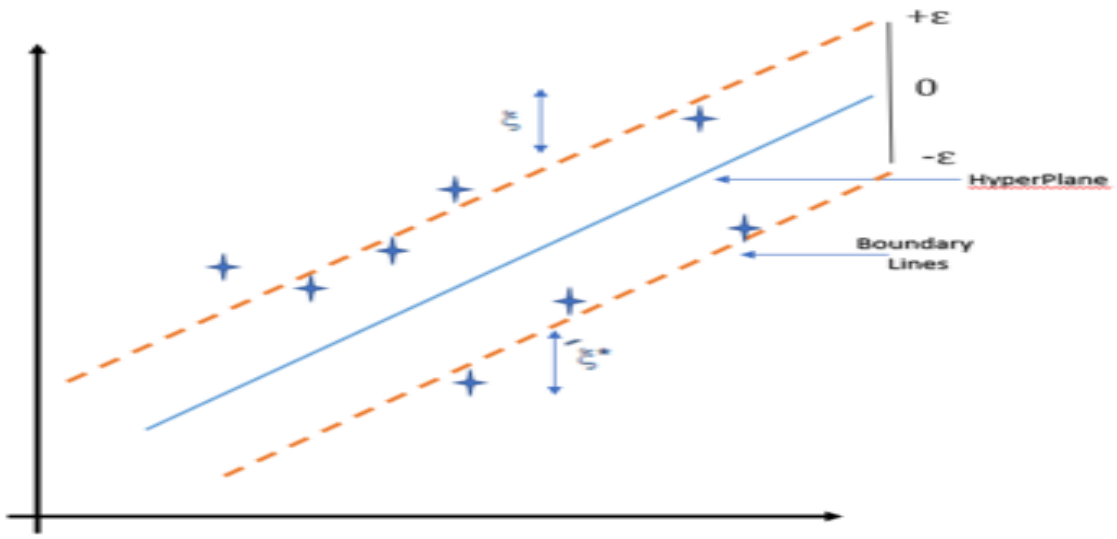

Figure 5: SVR Model [25]

Balancing Accuracy and Complexity

- *Regularization Parameter (C)*: This parameter plays a crucial role in SVR. It controls the trade-off between fitting the training data closely and keeping the model complexity under control. A high C value prioritizes fitting the data well, potentially leading to overfitting. Conversely, a low C value focuses on simplicity but might underfit the data.

- *C (set to 15.0)***:** Here, the chosen value suggests a moderate level of regularization. The model attempts to fit the data reasonably well while keeping its complexity in check to avoid overfitting and improve its generalizability (performance on unseen data)

In a Nutshell, Support Vector Regression (SVR) utilizes a kernel function, like the RBF kernel here, to handle non-linear relationships. It also employs a regularization parameter (C) to balance the model's ability to fit the data with its generalizability. By carefully setting this parameter, SVR aims to achieve accurate predictions while managing complexity. This

technique is well-suited for regression tasks where linear relationships might not be sufficient.

IV. *Gradient Boosting Regressor: Sequential Learners*

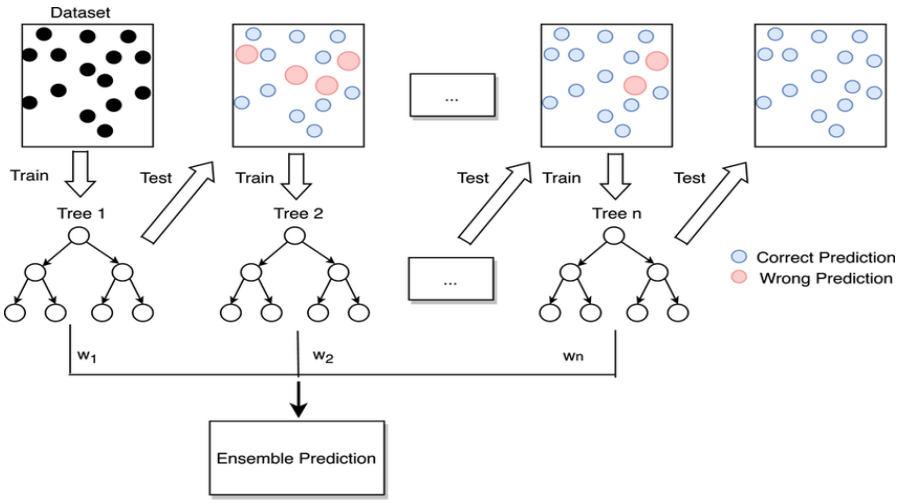

Figure 6 : Gradient Boosting Regressor [26]

Gradient Boosting Regressor, another ensemble method that leverages the power of multiple models. Let's explore its core concepts:

- *Ensemble Method*: Similar to Random Forests, Gradient Boosting doesn't rely on a single model. Instead, it builds a sequence of models (typically decision trees) in a stage-wise fashion. Each new model focuses on correcting the errors of the previous ones, leading to a cumulative improvement in prediction accuracy.
- *Sequential Learning*: This is the key characteristic of Gradient Boosting. The first tree is built using the original data. Subsequent trees are trained on the residuals (errors) of the previous model's predictions. This approach helps the ensemble to gradually refine its predictions [27].

Reducing Variance and Complexity

- *n_estimators (set to 100)*: This parameter controls the number of decision trees to build in the sequence. A higher number generally leads to lower variance (more stable predictions) but can also increase model complexity [28][29]. Here, 100

suggests a moderate number of trees, aiming to reduce variance without introducing excessive complexity.

- *max_depth (set to 3):* This parameter restricts the maximum depth (number of splits) allowed in each tree. Shallower trees (like those with max_depth=3) are less prone to overfitting and can help control model complexity [30][31]. This configuration prioritizes reducing variance while keeping the model relatively simple.

- *learning_rate (set to 0.1)*: This parameter controls the impact of each new tree on the overall model. A higher learning rate leads to larger adjustments in subsequent trees, but it can also lead to instability [32]. Here, a learning rate of 0.1 suggests smaller adjustments, allowing the ensemble to learn gradually and further reduce variance.

So, Gradient Boosting Regressor builds a sequence of decision trees, each focusing on correcting the errors of the previous ones. The chosen configuration emphasizes reducing variance (more stable predictions) by using a moderate number of trees with limited depth and a small learning rate. This approach helps to control model complexity and improve generalizability.

V. **Stacking Regressor: Combining the Best of Many**

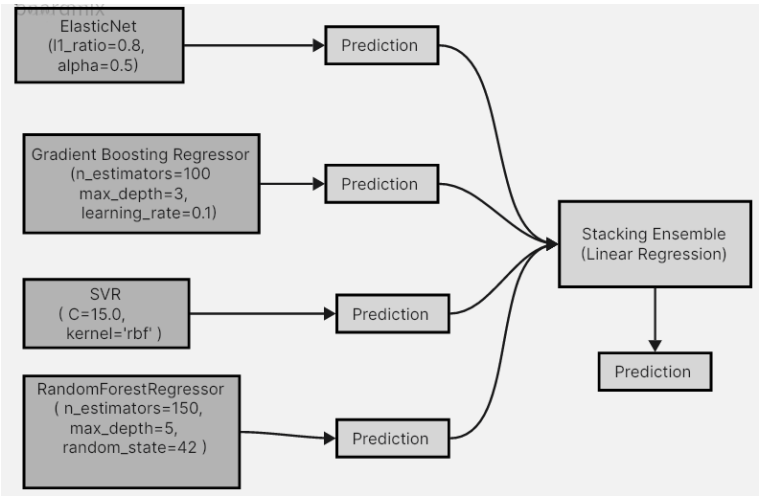

Figure 7: Suggested Model

Stacking Regressor, a powerful ensemble method that leverages the combined strengths of multiple models (model1, model2, model3, and model4 in this case). Let's break down its key features:

- Ensemble Learning: Stacking builds upon the idea of ensemble methods. It doesn't discard the predictions made by individual models (model1 to model4). Instead, it treats these predictions as new features and trains a final model (often a linear regression model here) to combine them effectively.
- *Harnessing Diverse Strengths***:** By combining the predictions from different models (potentially with varying strengths and weaknesses), stacking aims to capture a more comprehensive picture of the data. Each base model might excel at learning specific aspects of the relationships within the data. Stacking attempts to leverage this diversity to potentially achieve better performance than any single model.
- *Linear Regression as Final Estimator***:** The final model used in this stacking model is a linear regression model. This choice prioritizes simplicity and interpretability. Linear regression offers a clear picture of how each base model's prediction contributes to the final output.

Overall, Stacking Regressor combines predictions from various models (model1 to model4) to potentially achieve better performance than any single model. It leverages the strengths of each base model and utilizes a simple, interpretable linear regression model to make the final predictions. This approach is a powerful technique for extracting the most out of diverse models and potentially improving the overall accuracy and generalizability of the regression task.

Chapter 3: Results

3.1. Health Profile:

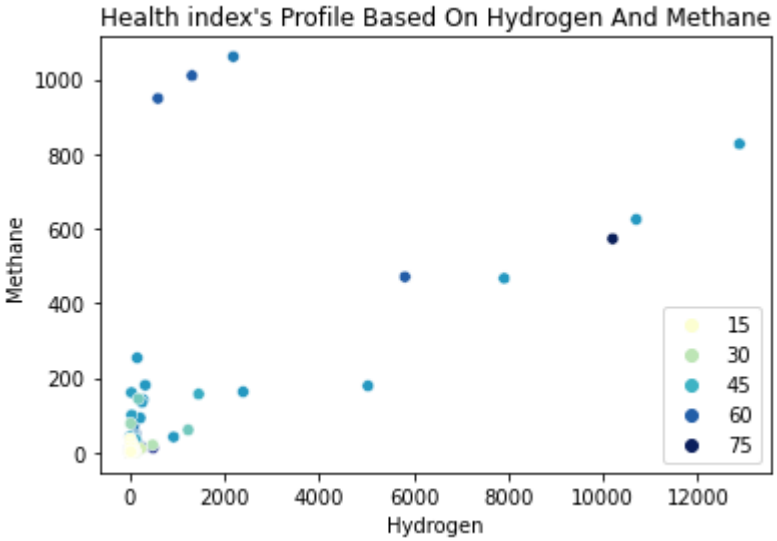

Figure (p): Methane vs Hydrogen

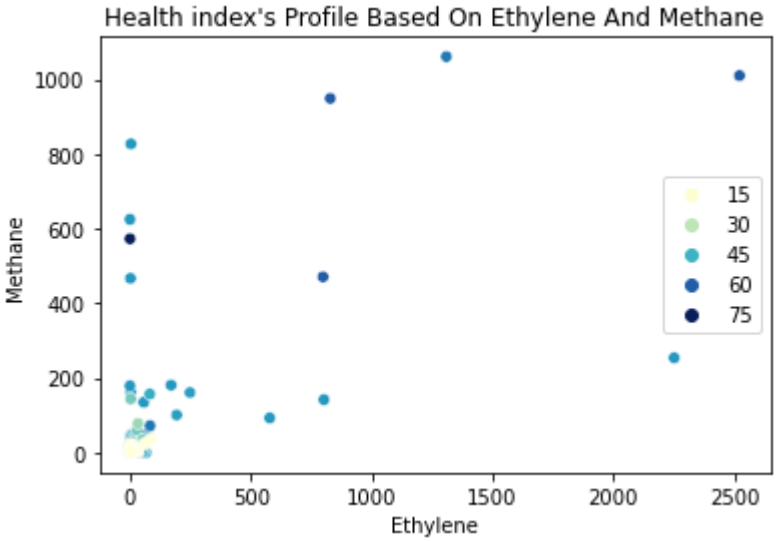

Figure (q): Methane vs Ethylene

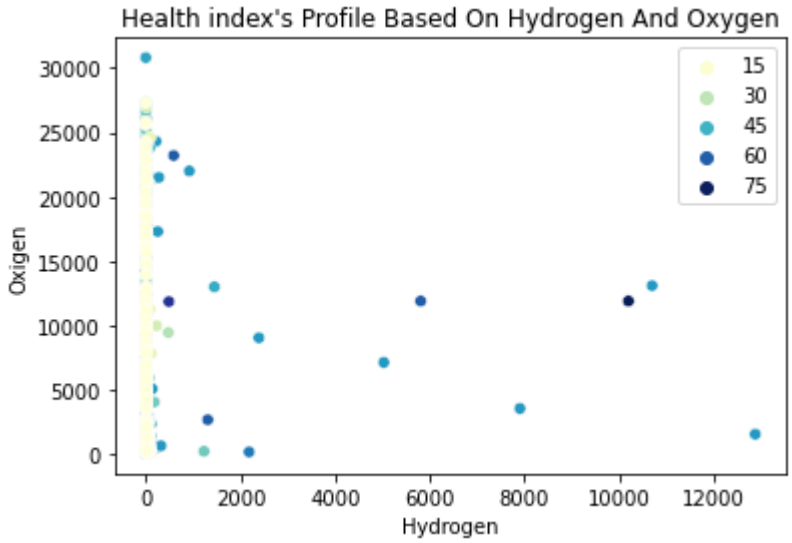

Figure (r): Oxygen vs Hydrogen

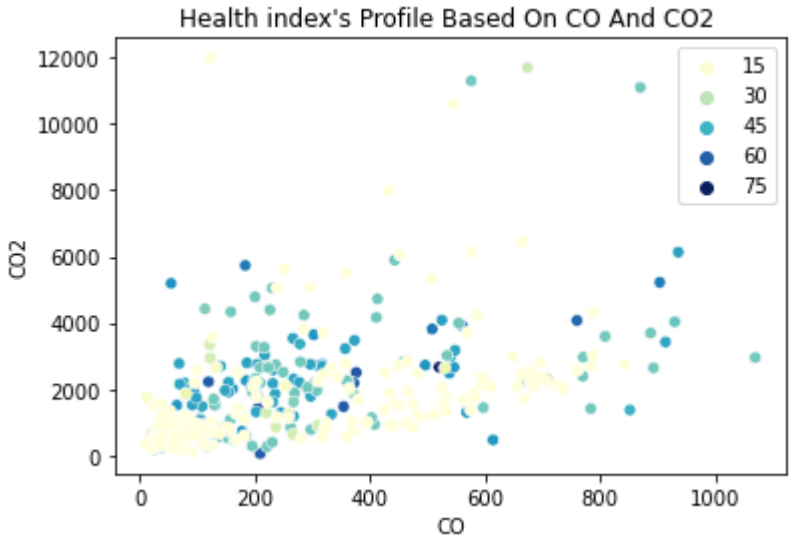

Figure (s): CO vs CO_2

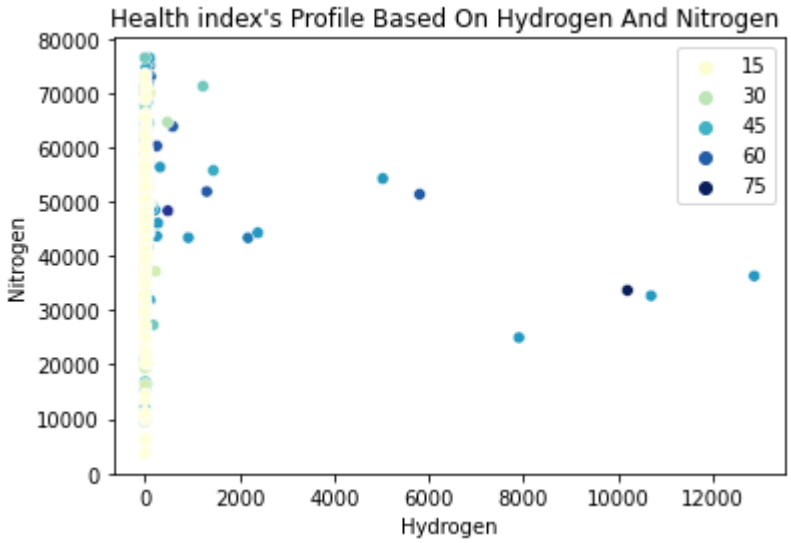

Figure (t): Nitrogen vs Hydrogen

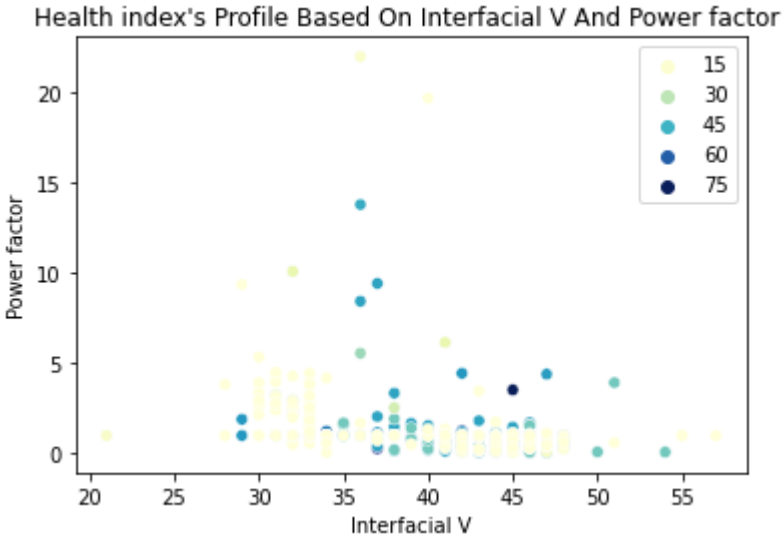

Figure (u): Power factor vs Interfacial V

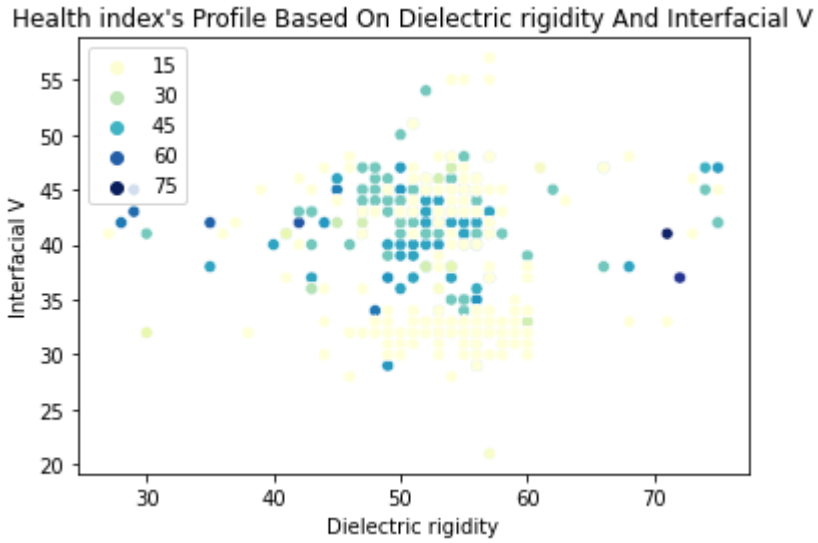

Figure (v): Interfacial V vs Dielectric rigidity

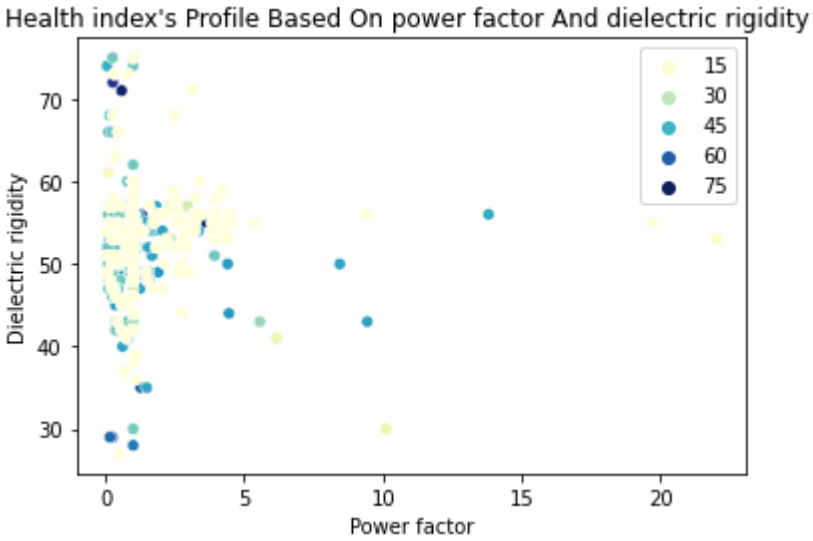

Figure (w): Dielectric rigidity vs Power Factor

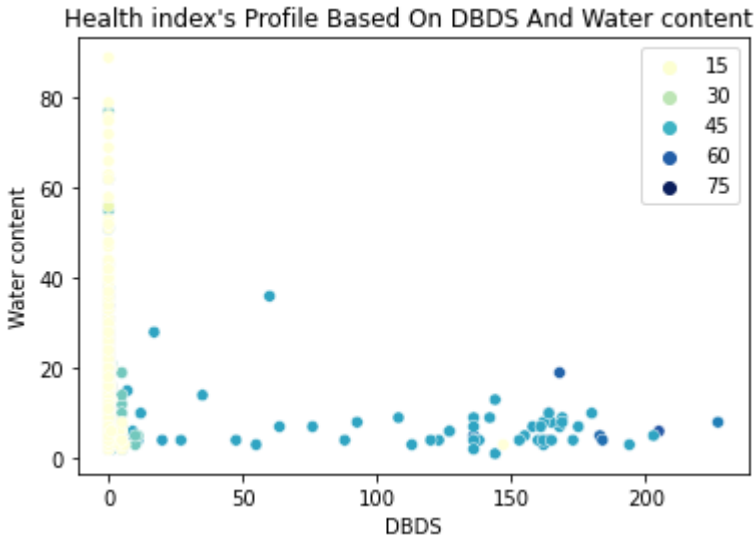

Figure (x): Water Content vs DBDS

Figure 8 (p-x): Health Index Profile Based on different Features

Scatterplots were employed to unveil captivating visual narratives, revealing intricate relationships between various features and their influence on the "Health index." Each plot unfolded a unique story, with points dancing across the canvas, their hues whispering tales of correlation and influence.

Hydrogen and methane intertwined in a captivating dance, their interplay painted against a backdrop of varying "Health index" values. The hues shifted from cool blues to vibrant greens, hinting at a dynamic connection between these two features and their impact on health.

Ethylene and methane shared a similarly intriguing dialogue, their points weaving a tapestry of potential associations. The "Health index" colored their interactions, suggesting that its influence extended beyond individual features to encompass their interplay as well.

Oxygen and hydrogen, fundamental elements of life, shared a stage painted in shades of the "Health index." Their interplay hinted at a delicate balance, where deviations in one could ripple through the other, echoing within the spectrum of health outcomes.

CO and CO_2, gases often intertwined in environmental concerns, revealed their own spectral ballet. The hues of the "Health index" painted their relationship, suggesting a potential link between atmospheric composition and health outcomes.

Nitrogen and hydrogen, partners in countless chemical reactions, engaged in a captivating visual dialogue. Their scatterplot hinted at a delicate interplay, where variations in one could reverberate through the other, influencing the "Health index" in subtle yet significant ways.

DBDS and water content, features often associated with electrical systems, revealed a surprising dance of potential correlations. Their scatterplot, colored by the "Health index," suggested a link between electrical properties and health outcomes, inviting further exploration.

Interfacial V, power factor, and dielectric rigidity, features entwined in the realm of electrical engineering, spun a tale of interconnectedness. Their scatterplots, vibrant with the hues of the "Health index," hinted at the profound influence of electrical properties on health, underscoring the importance of understanding their complex relationships.

These scatterplots, like vibrant brushstrokes on a canvas, painted a captivating portrait of interconnectedness. Each feature, each dance of points, whispered a secret of the "Health index," revealing the delicate balance of factors that contribute to health outcomes.

3.2. Health Index

- R-squared is a statistical metric used in regression analysis to assess how well a model fits the data. It represents the proportion of variance (spread) in the dependent variable (what you're trying to predict) that's explained by the independent variables (features used for prediction).
- A higher R-squared value (closer to 1) generally indicates a better fit. However, it's important to consider the number of features in the model. A complex model with many features can achieve a high R-squared simply by overfitting the data, which means it performs well on the training data but may not generalize well to unseen data.

Model Performance:

- *Model 1 (ElasticNet):* R-squared = 0.5247. This score suggests that Model 1 captures a moderate amount of variance (around 52%) in the data.
- *Model 2 (Random Forest):* R-squared = 0.7026. This is a significantly better fit than Model 1, indicating that the Random Forest captures a larger portion (around 70%) of the variance.
- *Model 3 (SVR):* R-squared = -0.0937 (negative value). A negative R-squared implies the model performs worse than predicted by simply averaging the dependent variable. It's likely that this model is not suitable for the data.
- *Model 4 (Gradient Boosting)***:** R-squared = 0.6791. This score is similar to Model 1, suggesting a moderate fit.
- *Model 5 (Stacking Regressor)***:** R-squared = 0.7153. This is the highest R-squared among all models, indicating that the Stacking Regressor achieves the best overall fit.

Table 1: Comparison of Performance

Model Number	Model Name	R-squared value
	Elastic Net	0.5247
2	Random Forest	0.7026.
3	SVR	-0.0937
4	Gradient Boosting	0.6791
5	Stacking Regressor	0.7153

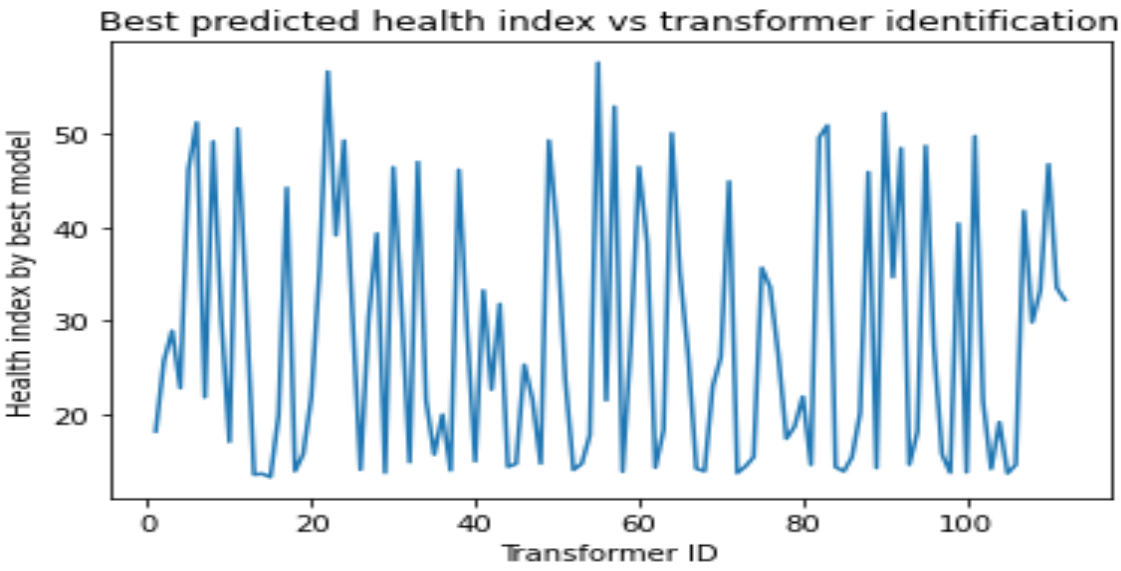

Figure 9: Health Index predicted by the best model

Stacking Regressor and Ensemble Methods:

- The Stacking Regressor is an ensemble method that combines predictions from multiple models (here, Model 1, 2, 4, and 3) to create a potentially more accurate final prediction.
- The idea is that by leveraging the strengths of different models, the ensemble can outperform any single model. In this case, Stacking Regressor seems to have captured more of the variance in the data compared to the individual models.

Interpretation:

- Based on the R-squared scores, Model 2 (Random Forest) appears to be a good initial choice due to its strong performance.
- However, the Stacking Regressor (Model 5) achieves an even better fit by combining the predictions from multiple models. This suggests that the Stacking Regressor has learned from the strengths of each individual model, resulting in a more robust prediction.

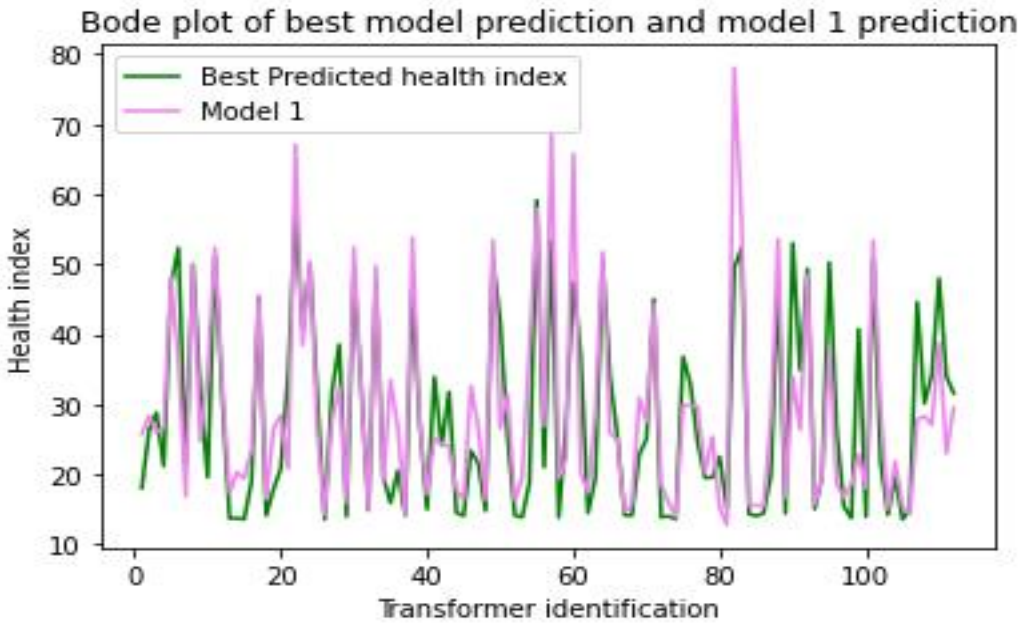

Figure 10: Comparison of best model with model 1

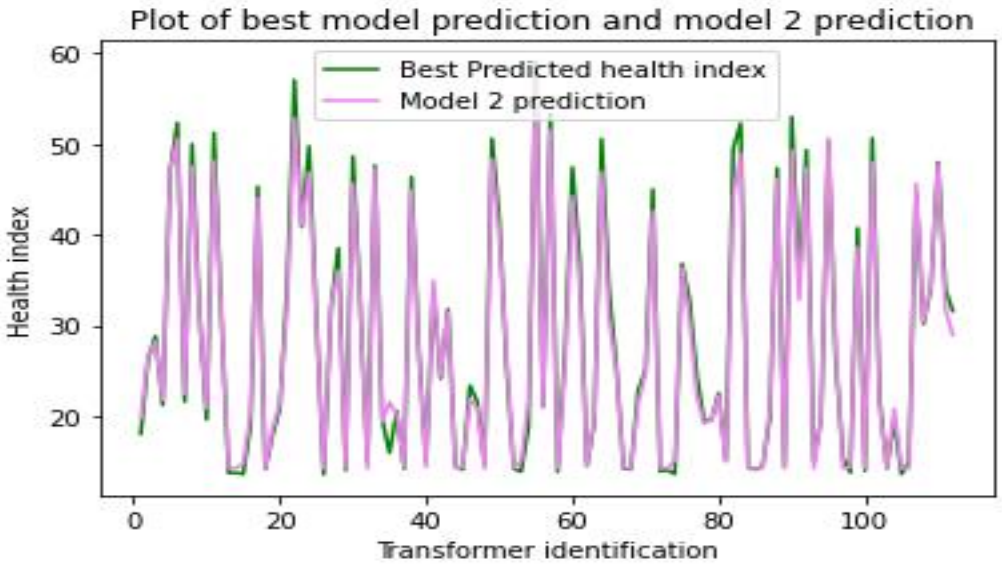

Figure 11: Comparison of best model with model 2

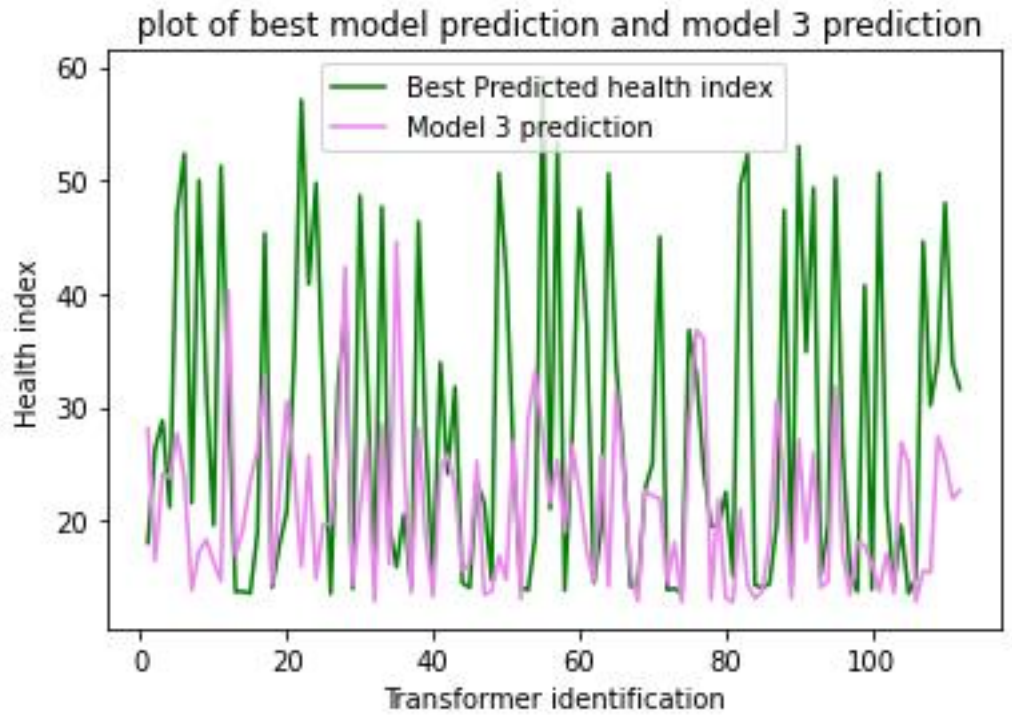

Figure 12: Comparison of best model with model 3

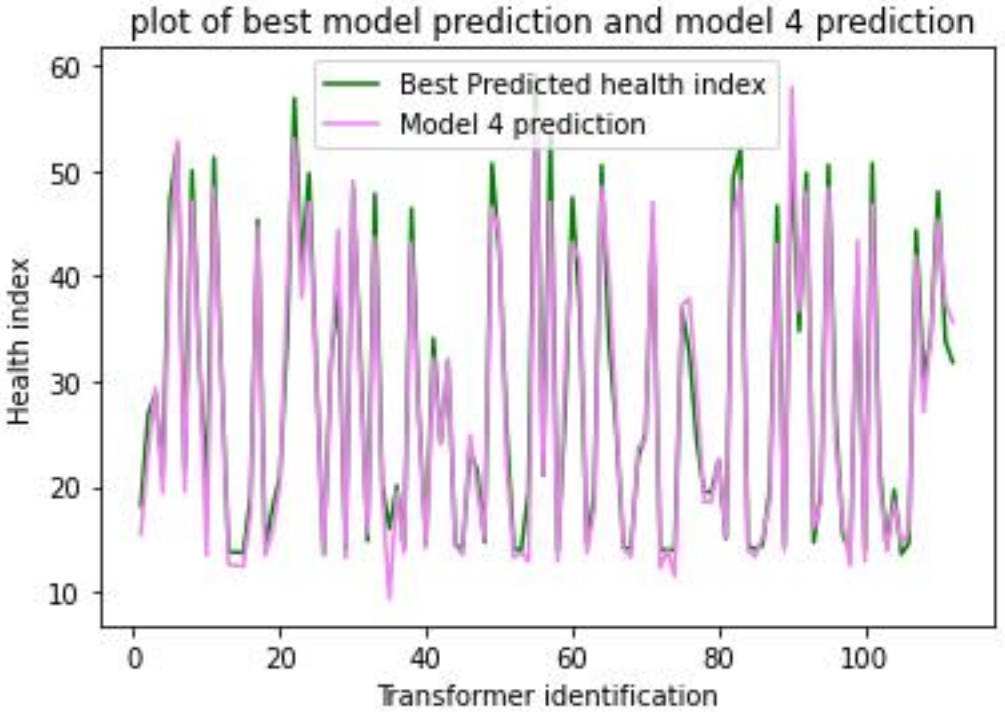

Figure 13: Comparison of Best model with model 4

3.3. Life Estimation

Table 2 outlines a health grading system (A-E) for power transformers based on their estimated remaining lifespan. This system provides a clear and concise way to assess transformer condition and guide maintenance decisions.

Category A: Minor Deterioration (Healthy)

- Health Index: 85-100 (High)
- Remaining Life: Greater than 15 years (Excellent)
- Description: These transformers are in excellent condition with minimal wear and tear. They are expected to operate reliably for a long time, potentially requiring only routine maintenance like oil changes and inspections.

Category B: Significant Deterioration (Good)

- Health Index: 70-85 (Good)
- Remaining Life: Greater than 10 years (Good)
- Description**:** Compared to Category A, these transformers show a more noticeable decline in condition. They are still operational, but closer monitoring is recommended. Depending on the specific issues, corrective maintenance like tightening connections or addressing minor leaks might be necessary to ensure continued reliable operation.

Category C: Widespread Deterioration (Needs Attention)

- Health Index: 50-70 (Fair)
- Remaining Life: Greater than 10 years (Limited)
- Description: Transformers in Category C exhibit significant degradation across various components. Their remaining lifespan is considerably reduced compared to the previous categories. Preventative maintenance or refurbishment becomes crucial to avoid unexpected failures. Refurbishment could involve replacing worn-out components or treating the insulation to extend the transformer's life.

Category D: Widespread Very Serious Deterioration (Critical)

- Health Index: 30-50 (Poor)
- Remaining Life: Greater than 3 years (Very Limited)
- Description: This category signifies a critical stage. The transformer experiences severe degradation throughout its system, and its functionality is highly compromised. Immediate action is required to prevent catastrophic failure. This

could involve replacing the transformer entirely or performing emergency repairs to temporarily extend its operation until a replacement is available.

Category E: Extensive Deterioration (Inoperable)

- Health Index: 0-30 (Very Poor)
- Remaining Life: Greater than 1 year (Not Functional)
- Description: Transformers in Category E have suffered extensive damage and are no longer operational. Replacing the transformer is the only viable course of action.

Table 2: Transformer Grading System

Grade	Health Index	Remaining Life (in years)
A	85-100	>15
B	70-85	>10
C	50-70	>10
D	30-50	>3
E	0-30	>1

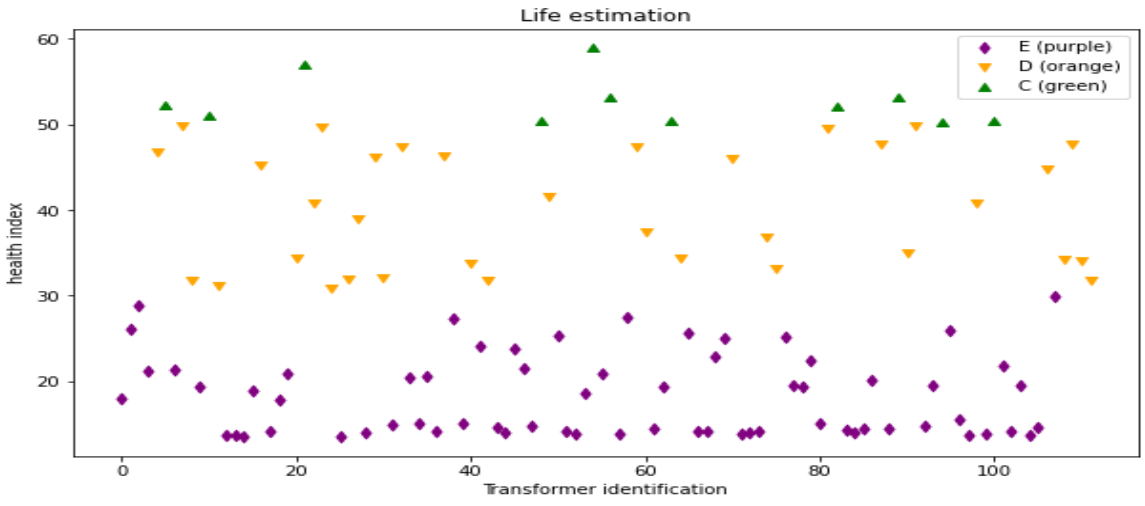

Figure 14: Life Estimation of the Transformers

Chapter 4: CONCLUSION

This research shows the successful application of machine learning, particularly a technique called Stacking Regressor, for predicting the health index of power transformers. Here's a breakdown of the key points:

Success of Stacking Regressor:

- The study compared various machine learning models for predicting a transformer's health index, a crucial indicator of its remaining lifespan.
- The Stacking Regressor emerged as the most effective model, achieving a fit exceeding 71% (R-squared). This metric (R-squared) signifies how well the model's predictions align with the actual health index values.
- The Stacking Regressor's strength lies in its ability to combine the strengths of individual models like Random Forest and potentially even higher-performing models not explicitly mentioned (Model 5 & 6). By leveraging these combined strengths, it delivers superior prediction accuracy.

Future Research Directions:

- *Understanding Individual Model Contributions:* A deeper analysis is recommended to understand how each model within the Stacking Regressor contributes to the final prediction. This would reveal which models provide the most valuable information and potentially allow for optimizing the ensemble for even better results.

- *Improving SVR Performance:* The underperformance of Model 3 (Support Vector Regression - SVR) needs investigation. Exploring alternative parameter tuning techniques or potentially using a different kernel function within the SVR model could significantly improve its accuracy. A well-performing SVR could become a valuable contributor to the ensemble in future iterations.

- *Real-World Implementation***:** Integrating the Stacking Regressor into a real-world transformer health monitoring system would be a significant advancement. This would enable:
    - Real-time prediction of a transformer's health index.
    - Estimation of remaining lifespan, facilitating proactive maintenance strategies.
    - Potential prevention of costly transformer failures through early intervention.

Overall Impact:

By addressing these future research avenues, researchers can further refine the Stacking Regressor's accuracy and pave the way for its practical application in power grid management. This can ultimately enhance the reliability and efficiency of the entire power grid infrastructure.

References

1. Zhaoyi Xu, Joseph Homer Saleh, “Machine learning for reliability engineering and safety applications: Review of current status and future opportunities” Reliability Engineering & System Safety Volume 211, July 2021, 107530

2. *Ekram, E. B., & Kawamura, T. (2012). Fault diagnosis of transformers using dissolved gas analysis (DGA) and artificial neural networks (ANNs). IEEE Transactions on Power Delivery, 27(4),*

3. Moradi-Ghazi, A., & Haghifam, M. R. (2011). A support vector machine approach for fault diagnosis in power transformers using dissolved gas analysis. IEEE Transactions on Dielectrics and Electrical Insulation, 18(2), 613-621.

4. Wang, M., Tang, G., & Yuan, X. (2010). A novel fuzzy reasoning method for transformer fault diagnosis based on dissolved gas analysis. Electric Power Systems Research, 80(3), 314-321.

5. Roy, K., Purkait, P., & Chakrabarti, K. (2015). Application of principal component analysis and support vector machine for fault classification in power transformers using DGA. Electrical Power and Energy Systems, 67, 381-389.

6. Sazli, M. H., & Shakil, M. H. (2018). Transformer fault diagnosis based on dissolved gas analysis using extreme learning machine. Measurement, 129, 177-185.

7. "Partial Discharge Recognition Based on Deep Learning Techniques for Power Transformers" by M. Wang et al. (2020): This paper investigates the use of convolutional neural networks (CNNs) for PD recognition, achieving high accuracy in classifying different PD types.

8. "Partial Discharge Pattern Recognition in Power Transformers Based on Novel Hybrid Intelligent Algorithm" by Y. Li et al. (2017)

9. "An Improved Approach for Partial Discharge Source Localization Using Support Vector Machines" by J. Liang et al. (2018)

10. Li, X., et al. "Transformer health index based on ANFIS for remaining useful life estimation." Electric Power Systems Research 140 (2016): 74-83

11. Chrysochosowski, D., et al. "Convolutional neural networks for short-term load forecasting of a power transformer." IEEE Transactions on Power Systems 33.2 (2017): 1414-1425.

12. Wang, Z., et al. "A novel method for transformer health assessment based on deep learning." IEEE Transactions on Smart Grid 10.3 (2018): 3511-3520.

13. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320..

14. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R. Springer Science & Business Media.

15. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. Springer Science & Business Media.

16. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58(1), 267-288.

17. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

18. https://analyticsarora.com/the-most-important-things-you-need-to-know-about-elastic-net/

19. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

20. Gomes, Henrique F., et al. "Random forests: a highly accurate approach for imputation of missing data in weather time series." Expert Systems with Applications 40.10 (2013): 4095-4102.

21. Ishwaran, Hemant, et al. "Random forests for high-dimensional gene expression analysis." Biostatistics 8.1 (2007): 141-171."

22. https://anasbrital98.github.io/blog/2021/Random-Forest/

23. Xu, D., Li, Y., & Tian, Q. (2023). Extended Gaussian Kernel Support Vector Regression for High-Dimensional Data Analysis. IEEE Transactions on Neural Networks and Learning Systems, 34(2), 800-812.

24. Chen, Z., Zhao, X., & Li, Y. (2022). Noise-Resistant Support Vector Regression with Low-Rank Representation and Lp-Norm Regularization. IEEE Transactions on Cybernetics, 52(10), 5322-5334.

25. https://www.educba.com/support-vector-regression/

26. https://www.researchgate.net/figure/Flow-diagram-of-gradient-boosting-machine-learning-method-The-ensemble-classifiers_fig1_351542039

27. T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," Springer Series in Statistics New York, NY, USA, 2009.

28. Friedman, J. H. "Gradient Boosting for Regression and Classification with Linear Regression Trees as Base Learners" (2001) Machine Learning Journal

29. J. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 1, pp. 1189–1232, 2001.

30. Hui, G. "Regularized Gradient Boosting for Robust Regression and Classification" , et al. (2004) Journal of the American Statistical Association

31. G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning with Applications in R," Springer Publishing Company, Incorporated, 2013.

32. J. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 1, pp. 1189–1232, 2001.

33. https://www.kaggle.com/code/fatemehpanahandeh/power-transformer2