**COMPARATIVE ANALYSIS OF XGBOOST AND RANDOM FOREST ALGORITHMS FOR TRANSFORMER FAILURE PREDICTION**

BY

**ABDULRAHMAN OPEYEMI ABDULKAREEM**

**19/67EC/00903**

**MARCH 2024**



**COMPARATIVE ANALYSIS OF XGBOOST AND RANDOM FOREST ALGORITHMS FOR TRANSFORMER FAILURE PREDICTION**

**BY**

**ABDULRAHMAN OPEYEMI ABDULKAREEM**

**19/67EC/00903**

**A Research Project Submitted to the Department Of Electrical And  
Computer Engineering, Faculty Of Engineering and Technology, Kwara  
State University, Malete, in Partial Fulfilment of the Requirements for  
the Award Of Bachelors of Engineering Degree (B.Eng) Degree In  
Electrical And Computer Engineering**

**MARCH 2024**

# DECLARATION

I hereby declare that this research project titled “**Comparative Analysis of Xgboost and Random Forest Algorithms for Transformer Failure Prediction**” is my work and has not been submitted byany otherperson for any degree or qualification at any higher institution. I also declare that theinformation provided therein are mine and those that are not mine are properlyacknowledged.

ABDULRAHMAN OPEYEMI ABDULKAREEM

Name of Student Signature and Date

# CERTIFICATION

This is to certify that this project titled “**Comparative Analysis of Xgboost and Random Forest Algorithms for Transformer Failure Prediction**” was carried out by **AbdulRahman Opeyemi AbdulKareem.** The project has been read and approved as meeting the requirements for the award of Bachelor of Engineering (B.Eng.) Degree in Electrical and Electronics Engineering in the department of Electrical and Computer Engineering, Faculty of Engineering and Technology, the Kwara State University, Malete.

**Engr. Dr. Bilkisu Jimada-Ojuolape** Date

Supervisor

**Engr. Dr. Abdulwaheed Musa** Date

Head of Department

External Examiner Date

# DEDICATION

This project is dedicated to Almighty Allah for his infinite mercy, wisdom, knowledge  
and protection in my life. Also, with all my heart, I dedicate this report to my ever present  
and supportive parents, my wonderful family, friends and well-wishers for their full  
support and encouragement throughout this journey.

# ACKNOWLEDGEMENT

My utmost gratitude, honors and adoration go to the Almighty Allah for his divine protection, wisdom, knowledge, guidance and protection over me throughout my course of study and this programme successfully.

I extend my deepest appreciation to my beloved parents and brothers for their invaluable moral, spiritual, and financial assistance throughout my studies.

Special thanks to my esteemed supervisor, Dr Bilkisu Jamada-Ojuolape, for her steadfast support, invaluable suggestions, and constructive criticisms that enriched the quality of this work. I am also grateful to the Head of Department, Dr AbdulWaheed Musa and other lecturers who have immensely contributed to my educational career by impacting on me knowledge and courage needed in the field of Electrical and Computer Engineering.

To my friends and course mates who have contributed in various ways to this phase of my life, I express my heartfelt gratitude.

Lastly, to all my guys who consistently kept me updated, regardless of how small the information may have seemed, you are much appreciated.

Thank you all immensely for making the journey fun and remarkable.

# TABLE OF CONTENT

[DECLARATION i](#_Toc173828196)

[CERTIFICATION ii](#_Toc173828197)

[DEDICATION iii](#_Toc173828198)

[ACKNOWLEDGEMENT iv](#_Toc173828199)

[TABLE OF CONTENT v](#_Toc173828200)

[LIST OF FIGURES viii](#_Toc173828201)

[LIST OF EQUATIONS ix](#_Toc173828202)

[LIST OF TABLES x](#_Toc173828203)

[CHAPTER ONE 1](#_Toc173828204)

[INTRODUCTION 1](#_Toc173828205)

[1.1 BACKGROUND OF STUDY 1](#_Toc173828206)

[1.2 PROBLEM STATEMENT 2](#_Toc173828207)

[1.3 AIM AND OBJECTIVES 2](#_Toc173828208)

[1.4 JUSTIFICATION 3](#_Toc173828209)

[1.5 SCOPE OF STUDY 3](#_Toc173828210)

[1.6 DEFINITION OF TERMS 4](#_Toc173828211)

[1.7 PROJECT LAYOUT 4](#_Toc173828212)

[CHAPTER TWO 7](#_Toc173828213)

[LITERATURE REVIEW 7](#_Toc173828214)

[2.1 INTRODUCTION 7](#_Toc173828215)

[2.2 THEORETICAL BACKGROUND 7](#_Toc173828216)

[2.2.1 MAINTENANCE 7](#_Toc173828217)

[2.2.2 TRANSFORMER FAULTS AND PREDICTIVE MAINTENANCE 9](#_Toc173828218)

[2.2.3 PRIDICTIVE MAINTENANCE AND TRANSFORMER DATA 11](#_Toc173828219)

[2.2.4 PREDICTIVE MAINTENANCE AND MACHINE LEARNING 12](#_Toc173828220)

[2.2.5 RANDOM FOREST ALGORITHM 14](#_Toc173828221)

[2.2.6 EXTRA GRADIENT BOOSTING (XGBOOST) ALGORITHM 16](#_Toc173828222)

[2.3 REVIEW OF RELATED WORKS 20](#_Toc173828223)

[CHAPTER 3 24](#_Toc173828224)

[METHODOLOGY 24](#_Toc173828225)

[3.1 INTRODUCTION 24](#_Toc173828226)

[3.2 STUDY FRAMEWORK 25](#_Toc173828227)

[3.2.1 DATASET 26](#_Toc173828228)

[3.2.2 MACHINE LEARNING TOOL 27](#_Toc173828229)

[3.2.3 PRE-PROCESSING DATA 28](#_Toc173828230)

[3.2.4 MODEL IMPLEMENTATION 28](#_Toc173828231)

[3.2.5 PERFORMANCE EVALUATION 29](#_Toc173828232)

[CHAPTER FOUR 33](#_Toc173828233)

[RESULT AND DISCUSSION 33](#_Toc173828234)

[4.0 INTRODUCTION 33](#_Toc173828235)

[4.1 DATA ANALYSIS AND MODEL IMPLEMENTATION 33](#_Toc173828236)

[4.1.1 DESCRIPTIVE STATISTICS OF DATA 33](#_Toc173828237)

[4.1.3 CORRELATION MARTRIX 34](#_Toc173828238)

[4.1.4 EXPLORATORY DATA ANALYSIS 35](#_Toc173828239)

[4.1.4 DATA PREPARATION 36](#_Toc173828240)

[4.1.5 MODEL TRAINING AND IMPLEMENTATION 36](#_Toc173828241)

[4.2 PERFOMANCE EVALUATION OF THE MODEL 39](#_Toc173828242)

[CHAPTER FIVE 41](#_Toc173828243)

[CONCLUSION AND RECOMENDATIONS 41](#_Toc173828244)

[5.1 CONCLUSION 41](#_Toc173828245)

[5.2 RECOMMENDATION 41](#_Toc173828246)

[5.3 LIMITATIONS 43](#_Toc173828247)

[REFERENCES 44](#_Toc173828248)

# LIST OF FIGURES

[Figure 1: Types of maintenance according to EN 13306 standard 8](#_Toc174710049)

[Figure 2: Electrical Faults of Transformer(Hussain et al., 2021) 9](#_Toc174710050)

[Figure 3:Supervised learning algorithm (Abbasi, 2021) 13](#_Toc174710051)

[Figure 4: Unsupervised learning algorithm(Abbasi, 2021) 13](#_Toc174710052)

[Figure 5: Deep learning algorithm(Abbasi, 2021) 14](#_Toc174710053)

[Figure 6: Random Forest Algorithm(Wang et al., 2023) 15](#_Toc174710054)

[Figure 7: Extreme Gradient Boosting Algorithm(Khan et al., 2022) 18](#_Toc174710055)

[Figure 8: Study Framework 25](#_Toc174710056)

[Figure 9: Confusion Matrix 31](#_Toc174710057)

[Figure 10: ROC curve (Abbasi, 2021) 33](#_Toc174710058)

[Figure 11: Correlation Matrix 36](#_Toc174710059)

[Figure 12: Data Visualization 38](#_Toc174710060)

[Figure 13: RF CM at 0.9 Figure 14: RF CM at 0.7 40](#_Toc174710061)

[Figure 13: RF CM at 0.9 Figure 14: RF ROC 40](#_Toc174710062)

[Figure 15: XGBoost CM at 0.5 Figure 16: XGBoost CM at 0.7 42](#_Toc174710063)

[Figure 15: XGBoost CM at 0.9 Figure 16: XGBoost ROC 42](#_Toc174710064)

[Figure 17: ROC Curve Comparison 44](#_Toc174710065)

# LIST OF EQUATIONS

[(1) 16](#_Toc160683469)

[**(2)** 16](#_Toc160683470)

[**(3)** 18](#_Toc160683471)

[**(4)** 19](#_Toc160683472)

[**(5)** 19](#_Toc160683473)

[**(6)** 19](#_Toc160683474)

[**(7)** 31](#_Toc160683475)

[**(8)** 31](#_Toc160683476)

[**(9)** 31](#_Toc160683477)

[**(10)** 32](#_Toc160683478)

[**(11)** 32](#_Toc160683479)

# LIST OF TABLES

[Table 1: Dataset Parameters 26](#_Toc173828174)

[Table 2: Descriptive Statistics of Data 34](#_Toc173828175)

[Table 3:Random Forest Perfomance 37](#_Toc173828176)

[Table 4: XGBoost Performance 38](#_Toc173828177)

[Table 5: Perfomance Evaluation 39](#_Toc173828178)

# ABSTRACT

In the field of power systems, electrical transformers are critical for the smooth and efficient transmission of electricity. However, transformer failures can cause significant disruptions, underscoring the need for effective and efficient maintenance strategies. This study embarks on an innovative exploration into predictive maintenance (PdM) for power grids, leveraging machine learning to preemptively identify transformer failures. The research compares the effectiveness of two advanced algorithms, Extreme Gradient Boosting (XGBoost) and Random Forest, using operational and historical data as a novel predictive tool. The paper highlights the inefficiencies of traditional maintenance strategies and positions PdM as a proactive alternative that promises enhanced grid resilience and optimized maintenance schedules.

The study begins with a comprehensive literature review, focusing on the theoretical background of transformer maintenance and the application of machine learning algorithms in failure prediction. Utilizing data from the kaggle database, the research employs rigorous preprocessing techniques and model development to evaluate the performance of both algorithms. The findings indicate that while both XGBoost and Random Forest demonstrate significant predictive capabilities, XGBoost outperforms Random Forest in terms of accuracy and efficiency. This research not only highlights the potential of machine learning in enhancing transformer maintenance strategies but also provides valuable insights for future studies aimed at improving predictive models in the electrical engineering domain. The results underscore the importance of adopting advanced analytical techniques to safeguard critical infrastructure and optimize maintenance practices in power systems.

# CHAPTER ONE

# INTRODUCTION

## BACKGROUND OF STUDY

For decades, the heart of any power grid, transformers, hum with the vital energy that fuels our world. Yet, despite their critical role, they remain vulnerable to breakdowns that cause costly downtime and disruptions. Traditional approaches, relying on fixed schedules or reactive repairs, often prove inefficient, leading to unnecessary maintenance and compromised power supply (Tianjin da xue et al., 2018a). Predictive maintenance (PdM) emerges as a game-changer, transforming equipment management from reactive to proactive, leveraging the power of several prominent mode of maintenance to analyze data, it anticipates equipment failures before they strike, enabling timely interventions and optimized maintenance schedules. (Carvalho et al., 2019).

This research delves into the exciting realm of PdM for electrical equipment using machine learning-based maintenance. By employing powerful machine learning algorithms such as Random Forest and Extreme Gradient Boosting (XGBoost), this research aims to identify patterns that might foreshadow transformer failures. Analyzing metrics like operational data, environmental factors, and historical failure records, we can predict potential breakdowns before they occur. Comparing the performance of these algorithms in terms of accuracy, efficiency, and interpretability will identify the most suitable tool for this task. Successful implementation could significantly reduce downtime, lower maintenance costs, and enhance grid resilience.

Ultimately, this project represents a crucial step towards smarter, more efficient, and more reliable transformers. By ensuring the uninterrupted flow of the vital energy that fuels our world, we can contribute to a brighter future for all.

## PROBLEM STATEMENT

In the realm of power grid management, the longevity and reliability of transformers, the core components sustaining our global energy infrastructure, face susceptibility to breakdowns, leading to costly disruptions. Conventional maintenance approaches, characterized by fixed schedules or reactive repairs, prove inefficient, resulting in unnecessary downtime and compromised power supply (Tianjin da xue et al., 2018a). Existing research has explored various AI algorithms for transformer predictive maintenance, including support vector machines (SVM), neural networks (NN), and decision trees (DT). However, these studies primarily focus on direct sensor data, limiting their applicability in scenarios with limited or unreliable sensor coverage.

## AIM AND OBJECTIVES

The aim of this project is to demonstrate the feasibility and advantages of utilizing a broader range of operational and historical data for transformer failure prediction through a comparative analysis of Random Forest and Extreme Gradient Boosting (XGBoost) algorithms, ultimately establishing a reliable predictive maintenance (PdM) strategy for transformers. The objectives of the project are:

1. Collect and analyze comprehensive transformer data.
2. Thoroughly evaluate the Random Forest algorithm's performance in predicting transformer failures based on the collected data.
3. Thoroughly evaluate the Extreme Gradient Boosting (XGboost) algorithm's performance in predicting transformer failures based on the collected data.
4. Determine the most effective algorithm.

## JUSTIFICATION

Traditional maintenance methods for transformers, relying on schedules or reactive repairs, struggle with efficiency and resource allocation(Tianjin da xue et al., 2018a). Unexpected breakdowns disrupt operations and cost dearly. Other predictive maintenance with AI algorithm offers a solution, but often relies on costly, sparse sensor data. This project explores the largely untapped potential of indirect prediction using abundant, readily available data. By comparing powerful algorithms like Random Forest and XGBoost in terms of accuracy, efficiency, and interpretability will identify the most suitable tool for this task. Successful implementation could significantly reduce downtime, lower maintenance costs, and enhance resilience.

## SCOPE OF STUDY

This project focuses on a comparative analysis of XGBoost and Random Forest algorithms for predicting transformer failures. By leveraging operational and historical data, the study aims to evaluate the effectiveness of these algorithms in forecasting potential breakdowns. The comparison will assess each algorithm's accuracy, efficiency, and interpretability in the context of transformer failure prediction. The objective is to identify the most effective tool for enhancing predictive maintenance strategies, ultimately contributing to reduced downtime, lower maintenance costs, and improved grid reliability.

## DEFINITION OF TERMS

**Predictive Maintenance**

Also known as Statistical-based maintenance. It is based on the continuous monitoring of the equipment or the machine. It employs prediction tools to measure when such maintenance actions are necessary, hence the maintenance can be scheduled. Furthermore, it allows failure detection at an early stage based on the historical data by utilizing those prediction tools such as machine learning methods, integrity factors, statistical inference approaches, and engineering techniques (Çinar et al., 2020).

**Machine Learning**

Machine learning refers to the ability of systems to learn from specific training data related to a particular problem, automating the creation of analytical models and addressing associated tasks (Çinar et al., 2020; Janiesch et al., 2021).

**Indirect Prediction of Transformer Failures**

Rather than directly monitoring the transformers, the project aims to use the patterns and anomalies observed to infer potential issues with transformers. By identifying correlations and trends in the historical data that precede transformer failures.

## PROJECT LAYOUT

The organizational structure outlined below is adhered to in the project report:

**Chapter 1:** Introduction

In this chapter, an overview of the research project is presented. It encompasses an introduction to the project, a delineation of the problems under scrutiny, the study's goals and objectives, the research's importance, the study's scope, and the layout of the project.

**Chapter 2:** Literature Review

This chapter reviews existing research on predictive maintenance (PdM) for transformers, with a focus on AI algorithms used for failure prediction. It evaluates studies on various approaches, including those that use direct sensor data, and identifies gaps that the current research aims to address.

**Chapter 3:** Methodology

This chapter details the process for collecting and preprocessing data related to transformer failures. It describes the techniques used to ensure data quality and prepare it for analysis, and specifies the evaluation metrics for comparing the performance of the Random Forest and XGBoost algorithms.

**Chapter 4:** Result and Analysis

In this chapter, a comparative analysis of Random Forest and XGBoost is performed, highlighting their strengths and weaknesses in predicting transformer failures based on data analyzed.

**Chapter 5:** Conclusion and Recommendation

This chapter summarizes the key findings of the comparative analysis, emphasizing the effectiveness of Random Forest and XGBoost for transformer failure prediction. It provides recommendations for improving predictive maintenance strategies based on the results of the analysis.

# CHAPTER TWO

# LITERATURE REVIEW

## INTRODUCTION

This chapter aims to review both past and present literature pertinent to this area of research. The research process commenced with a comprehensive review of journals and internet sites in the field of machine learning algorithms, specifically focusing on XGBoost and Random Forest, and their application in predicting transformer failures.

## THEORETICAL BACKGROUND

### MAINTENANCE

Like any complex machinery, transformers rely on diligent maintenance to maintain peak performance and minimize disruptive failures. According to EN 13306, maintenance encompasses a holistic approach of technical, administrative, and managerial actions throughout the transformer's lifespan(British Standards Institution, 2018). The maintenance of transformers in a power grid involves a set of essential activities aimed at ensuring the uninterrupted and efficient functioning of the equipment (Rojek et al., 2023). These maintenance tasks include repairing, replacing components, routine checking, adjusting parameters, testing, measuring, and fault-finding. The execution of maintenance involves implementing a defined strategy through action plans. These plans ensure adherence to guidelines, maintaining direction and facilitating subsequent activities necessary for the system's maintenance. According to standard EN 13306, these maintenance strategies can be classified in a number of ways (corrective maintenance (CM), preventive maintenance (PM), predictive maintenance (PdM)) (Coandǎ et al., 2020). The figure below shows the schematic representation of the maintenance types.

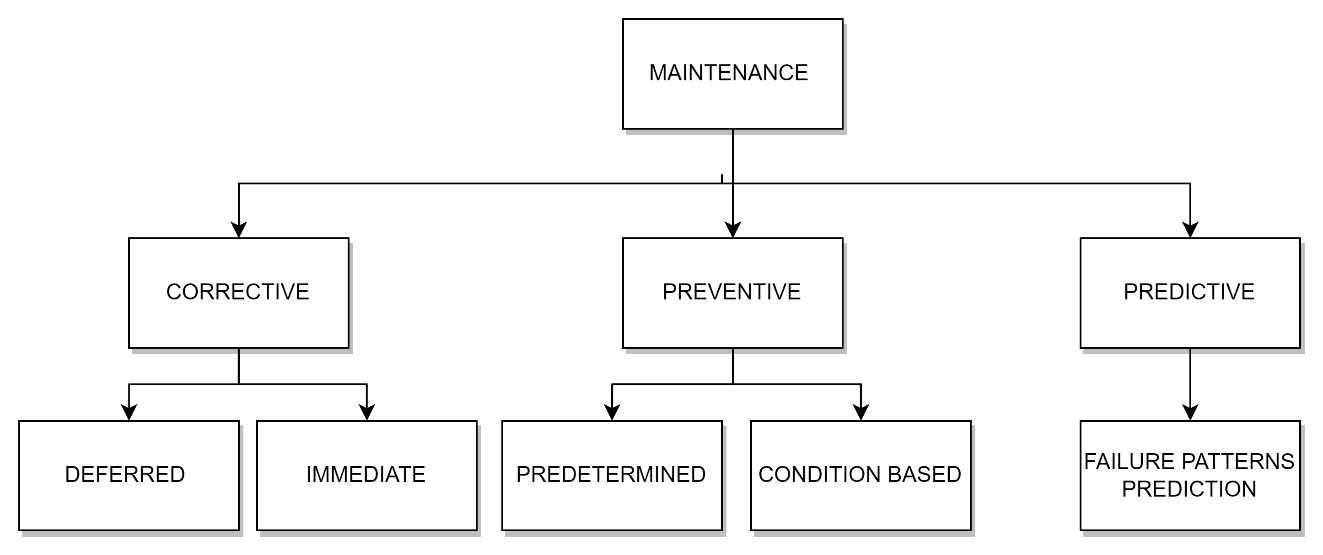


Figure 1: Types of maintenance according to EN 13306 standard

The strategies of Corrective Maintenance (CM) and Preventive Maintenance (PM) have been employed since the early 1990s. Corrective maintenance aims to restore a system after a failure, often resulting in unpredictable consequences and higher costs. Preventive maintenance involves planned interventions to keep equipment in good condition, intending to address issues before they lead to failure. However, the challenge lies in flawlessly scheduling maintenance well in advance. Increasing demands for system reliability have diminished the effectiveness of preventive maintenance, with the downside of not always considering the actual system state, leading to potential unnecessary procedures and additional costs(Coandǎ et al., 2020; Rojek et al., 2023).

In recent years, a third strategy, Predictive Maintenance (PdM), has gained prominence. Predictive maintenance represents a significant advancement over corrective and preventive maintenance strategie(Coandǎ et al., 2020)s. It leverages the power of data, analytics, and machine learning to predict equipment failures before they occur, enabling timely intervention. Predictive maintenance involves continuously monitoring the condition of the transformer and analyzing this data to identify signs of potential failures. This is achieved using advanced machine learning algorithms that can learn from historical data to identify patterns that precede a failure(Carvalho et al., 2019).

### TRANSFORMER FAULTS AND PREDICTIVE MAINTENANCE

Transformers, often unsung heroes within the power grid, play a crucial role in stepping down high-voltage electricity to ensure its safe delivery to homes and businesses. However, these intricate machines are susceptible to malfunctions, and when transformer faults arise, they can pose significant challenges to grid reliability and stability(Hussain et al., 2021). These faults can be broadly categorized into internal and external faults. As depicted in **Figure below**.

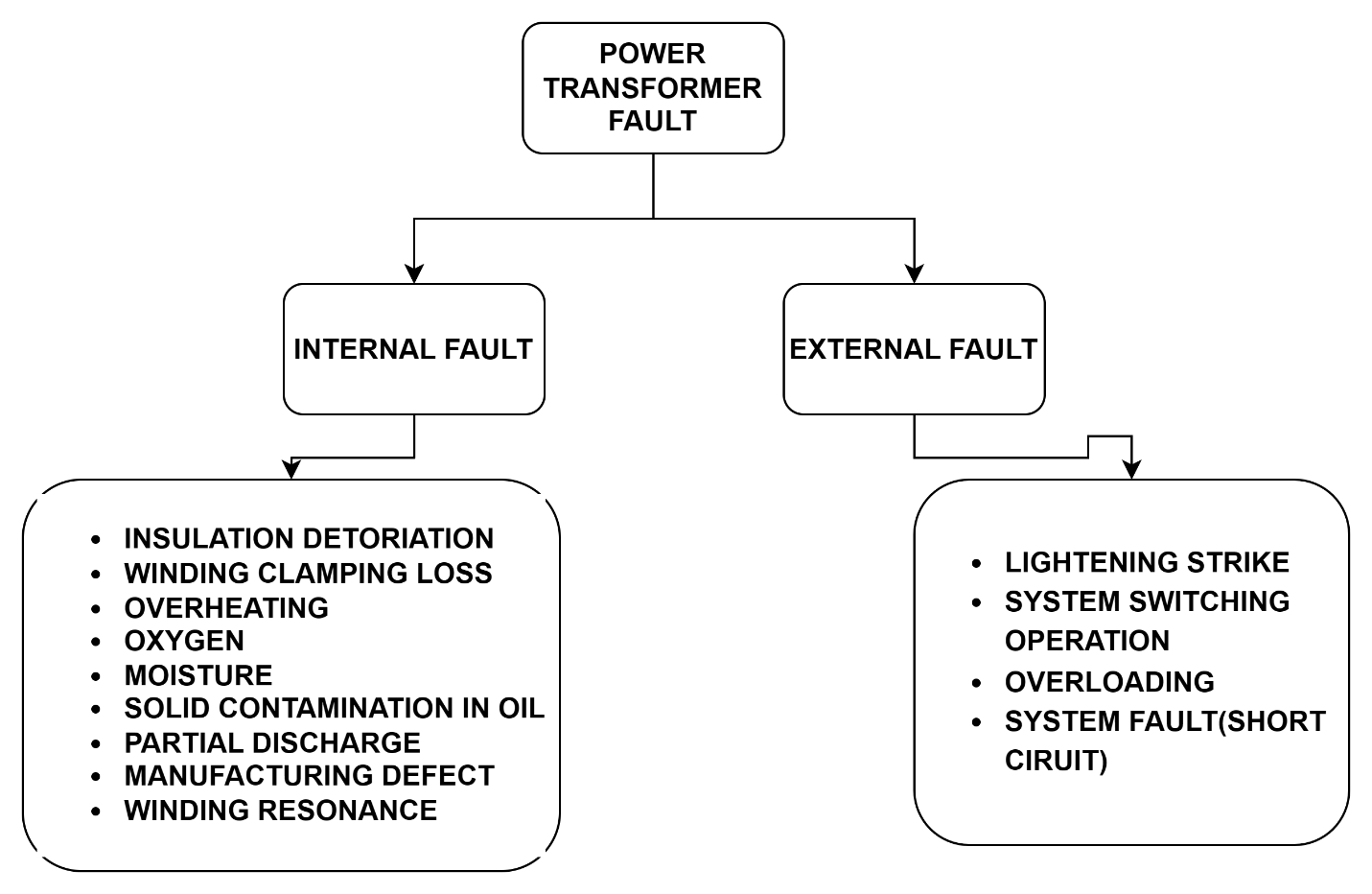


Figure 2: Electrical Faults of Transformer

**Internal Faults** which constitute approximately 70% to 80% of transformer faults, originate from minor discharges within the transformer insulation, initially existing as transient states(Hussain et al., 2021). These faults can manifest in various areas of the transformer, including the winding (affected by axial displacement, buckling deformation, disc space variation, and short-circuited turns), tank, insulating oil (impacted by oxidation, water penetration, dissolution due to temperature rise, and acidity), core (experiencing insulation failure and shorted laminations), terminal (affected by open leads, loose connections, and short circuits), cooling system, and tap changer (experiencing mechanical or electrical issues, short circuits, and overheating)(Hussain et al., 2021).

**External faults** are faults that occur outside the transformer, such as in the power system or the load. They can be caused by events such as lightning strikes, short circuits, overloads, or mechanical damage(Hussain et al., 2021). External faults can affect the transformer by inducing overvoltages, overcurrents, or abnormal temperature rises.

However, rather than adhering to a predetermined schedule (preventive maintenance) or responding reactively to failures (corrective maintenance), a proactive maintenance strategy is employed based on the predicted health of transformers. Predictive maintenance, utilizing machine learning, analyzes historical data encompassing load patterns, voltage levels, and frequency fluctuations. By leveraging this data, the system can predict potential transformer failures. This proactive approach enables scheduled maintenance interventions, aiming to prevent failures and minimize downtime, contributing to a more efficient and reliable transformer.(Tianjin da xue et al., 2018b).

### PRIDICTIVE MAINTENANCE AND TRANSFORMER DATA

Operational and historical data for transformers are crucial for ensuring their reliable performance and preventing failures. This data includes parameters such as temperature, load conditions, and historical failure records. By monitoring this data, utilities and maintenance teams can identify potential issues before they escalate, enabling them to take preventive measures. This approach helps avoid expensive repairs and power outages, and ensures public safety.

**Temperature Data**: This includes measurements of temperature levels at various points within the transformer. Monitoring temperature is critical for ensuring proper equipment function and preventing overheating, which can lead to failures (Liu et al., 2022).

**Load Data**: This pertains to measurements of electricity demand or load levels on the transformer. Understanding load patterns helps operators anticipate peak demand periods, plan for capacity requirements, and optimize resource allocation (Liu et al., 2022).

**Historical Failure Data**: This includes records of past transformer failures and maintenance activities. Analyzing historical failure data helps identify common failure modes and trends, guiding effective maintenance strategies (Abbasi, 2021; Marcelino et al., 2021; Wang et al., 2023).

**Operational Data**: This encompasses measurements of various operational parameters such as voltage, current, and insulation resistance. Monitoring operational data ensures that the transformer operates within acceptable limits and helps identify issues affecting performance (Abbasi, 2021; Marcelino et al., 2021; Wang et al., 2023).

Operational and historical data for transformers can provide valuable insights for predictive maintenance programs. By monitoring these metrics over time, operators can identify trends, patterns, and early indicators of equipment degradation or impending failures. This proactive approach allows for timely maintenance interventions, reducing the risk of unplanned outages and optimizing asset performance.

### PREDICTIVE MAINTENANCE AND MACHINE LEARNING

Machine learning, a subset of artificial intelligence, involves developing algorithms and statistical models that enable computer systems to learn and make predictions or decisions without being explicitly programmed(Abbasi, 2021; Marcelino et al., 2021; Wang et al., 2023). Machine learning, along with the Internet of Things (IoT), plays a pivotal role in predictive maintenance. IoT devices, essentially sensors or equipment, continuously feed real-time data to centralized systems. This influx of data, combined with machine learning models, allows for incredibly precise predictive maintenance schedules(Marcelino et al., 2021). The three main ML techniques employed are.

1. **Supervised learning algorithms** like XGBoost and Random Forest excel at pattern recognition. Trained on historical data of transformer failures and corresponding sensor readings, they learn to identify the intricate relationships between sensor data and impending faults (Janiesch et al., 2021). Just like an experienced doctor analyzing test results, these algorithms can trigger alarms based on subtle anomalies in real-time data, predicting failures before they occur.



Figure 3:Supervised learning algorithm (Abbasi, 2021)

1. **Unsupervised learning algorithms** like k-means clustering take a different approach. They explore vast amounts of sensor data, searching for hidden patterns and groupings that might not be readily apparent to traditional methods (Çinar et al., 2020). Similarly, in fault detection, it can uncover new failure modes or early signs of trouble, providing deeper insights into the health of the equipment.



Figure 4: Unsupervised learning algorithm(Abbasi, 2021)

1. **Deep learning** takes things a step further with its ability to analyze complex data streams like vibration signals or infrared images(Breviglieri et al., 2021a; Janiesch et al., 2021). Convolutional neural networks (CNNs) can be trained on thousands of transformer images, both healthy and faulty. They meticulously dissect each image, extracting minute features and patterns that human eyes might miss.



Figure 5: Deep learning algorithm(Abbasi, 2021)

In predictive maintenance, machine learning algorithms analyze historical data to identify patterns that precede a failure. This allows for proactive scheduling of maintenance activities, potentially preventing failures and reducing downtime(Coandǎ et al., 2020). Predictive maintenance has been reported to reduce breakdowns by 70%, increase productivity by 25%, and lower maintenance costs by 25%(Rojek et al., 2023). This approach, combined with machine learning, is reshaping how businesses operate, making them more proactive, efficient, and resilient.

### RANDOM FOREST ALGORITHM

Random Forest is a robust ensemble learning technique that leverages the collective power of multiple decision trees to tackle complex problems in supervised learning(Wang et al., 2023). It’s versatile and can be applied to both Classification and Regression tasks in Machine Learning. This ensemble learning method combines a multitude of sensor readings and historical trends, enhancing accuracy and resilience in pinpointing potential failures(Wang et al., 2023). As the name implies, a “Random Forest” is a classifier comprising numerous decision trees on various subsets of the given dataset. It averages the results to enhance the predictive accuracy of the dataset(Wang et al., 2023). The more trees in the forest, the higher the accuracy, which helps prevent overfitting. The below diagram explains the working of the Random Forest algorithm:

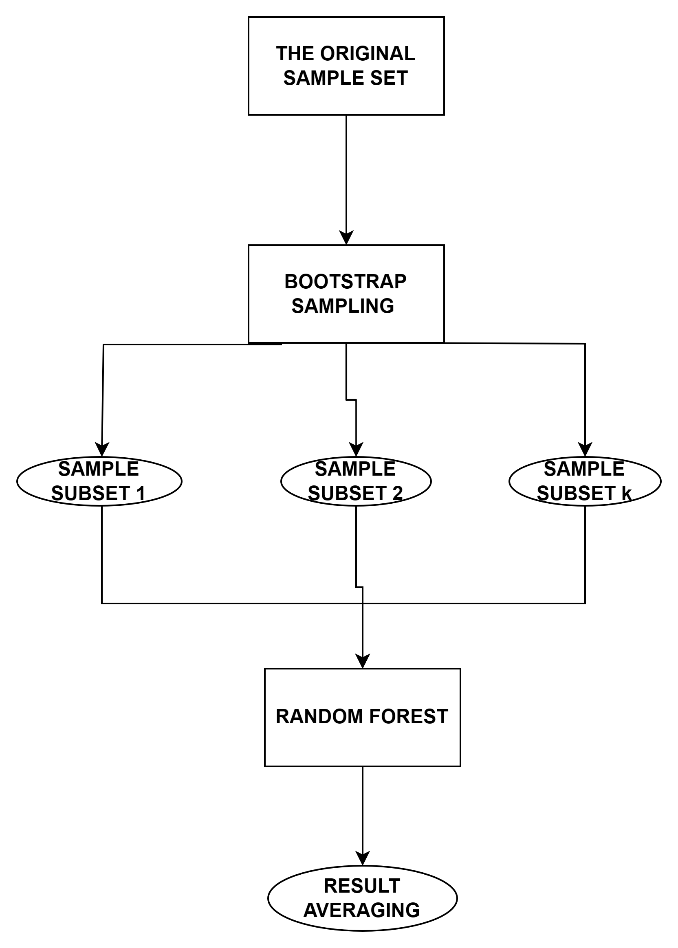


Figure 6: Random Forest Algorithm(Wang et al., 2023)

In the context of **classification**, the algorithm begins by randomly sampling subsets of the training data with replacement. For each subset, decision trees are constructed using group of sensor readings and historical trends. During the construction of these trees, a subset of features is randomly selected at each node. The optimal feature and split point are chosen based on their ability to minimize the Gini impurity after splitting. Gini impurity is a measure of the uncertainty or impurity of a set of samples, with lower values indicating purer nodes(Wang et al., 2023).

The Gini impurity, denoted as Gini(D), for a dataset D with C classes is calculated using the formula(Wang et al., 2023):

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

Here, pi​ is the probability of class i in node D.

While classification focuses on distinct categories, Random Forest also ventures into the realm of **regression.** Here, each tree becomes a treasure hunter, analyzing relationships between features and actual values to estimate the remaining lifespan of equipment based on their current readings(Wang et al., 2023). The final predicted value for a sample is the average of predictions by all the individual trees, calculated as(Wang et al., 2023):

|  |  |  |
| --- | --- | --- |
|  |  | **(2)** |

Here,represents the predicted value by the i-th decision tree for sample x and k is the number of decision tree.

### EXTRA GRADIENT BOOSTING (XGBOOST) ALGORITHM

Gradient Boosting is a machine learning technique that builds a strong predictive model by combining the predictions of several weaker models. It’s particularly useful for regression and classification problems. The core idea is to construct each weak learner based on the gradient direction of the loss function, which leads to iterative refinement of predictions and a robust model.

Building upon this concept, XGBoost, or “Extreme Gradient Boosting”, a scalable machine learning system for tree boosting. It’s widely used to achieve state-of-the-art results on data challenges such as Kaggle competitions. Developed by Chen and Guestrin, XGBoost uses Classification and Regression Trees (CART) as the base classifier and integrates it with gradient boosting(Chen et al., 2019). The algorithm adds a regularization term to the loss function, reducing model complexity and achieving a balance between model accuracy and complexity. Each time a new CART is added, the prediction residuals of the previous CART are fitted, and the accumulated prediction results of all CARTs yield the final model results(Wang et al., 2023). This makes XGBoost a highly efficient, flexible, and portable tool for machine learning tasks. The below diagram explains the working of the XGBoost algorithm:

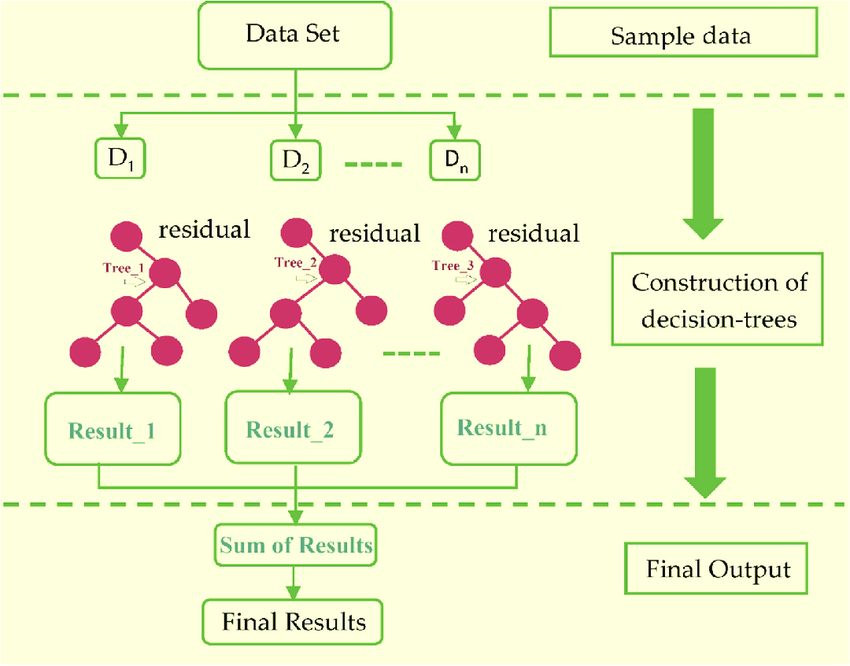


Figure 7: Extreme Gradient Boosting Algorithm(Khan et al., 2022)

The mathematical equations below guide how the model learns from data for prediction(Chen et al., 2019). The loss function, regularization term, and objective function all work together to minimize the difference between the predicted and actual transformer states, while preventing overfitting.

|  |  |  |
| --- | --- | --- |
|  |  | **(3)** |

This is the overall function that the XGBoost algorithm aims to minimize. It consists of the loss term, which measures the difference between the predicted and actual target values, regularization term, Ω(ft​), which prevents overfitting by adding a penalty for complexity and iteration t.

|  |  |  |
| --- | --- | --- |
|  |  | **(4)** |

This term is used to control the complexity of the model. It penalizes the model as the number of leaves (T) in the tree increases and as the leaf weights (w) become larger. The parameters γ and λ control the extent of regularization.

|  |  |  |
| --- | --- | --- |
|  |  | **(5)** |

This equation is a simplification of the objective function using a second-order Taylor expansion. It approximates the loss function around the current estimate, which makes the optimization problem easier to solve(Chen et al., 2019).

|  |  |  |
| --- | --- | --- |
|  |  | **(6)** |

Where Ij​ is the instance set of leaf j, and wj​ is the score assigned to leaf j. This is the final form of the objective function that the algorithm minimizes at each step. It is a sum over all leaves of the tree. For each leaf, it calculates a score based on the sum of the gradients and Hessian values of the instances assigned to that leaf.

## REVIEW OF RELATED WORKS

This section contains comprehensive review of past works that are related to this study as well as the strength, weakness and the methods adopted in each review.

In recent study, (Wang et al., 2023) present a novel approach named TPE-XGBoost for diagnosing transformer faults using incomplete data. This methodology utilizes Bayesian optimization to fine-tune the hyperparameters of the XGBoost model, showcasing superior performance in comparison to alternative machine learning algorithms. A notable strength of this method lies in its capability to effectively handle incomplete datasets, as evidenced by its robust performance. However, the study highlights a limitation regarding reduced diagnostic accuracy when the rate of missing data exceeds 20%, emphasizing the necessity for further enhancement, particularly in scenarios with a high missing data rate exceeding 30%.

Introduced by (Chen et al., 2019), a methodology for predicting transient stability status in power systems using the XGBoost model. Key features of the generator’s state are extracted and redundant ones are removed. The paper emphasizes the XGBoost model as a competitive technology for transient stability prediction due to its advantages as a tree structure model that does not require data normalization and can effectively handle missing values. Despite its advantages, the paper calls for more empirical validation and real-world application to fully evaluate its effectiveness.

The study by (Zhang et al., 2019), which explores the use of the XGBoost algorithm for diagnosing bearing faults in complex industrial environments. The research compares XGBoost with alternative tree models and highlights its superior performance in terms of both training time and accuracy. Notably, the paper emphasizes the importance of managing model complexity through regular coefficients and employing Bayesian optimization for parameter tuning. However, the study acknowledges limitations, including the need for high-quality data and challenges related to model generalizability and interpretability. Overall, the findings provide valuable insights into the potential of advanced machine learning techniques for industrial fault diagnosis.

In the realm of artificial intelligence (AI) applications within Industry 4.0, particularly focusing on its utilization in maintenance processes. (Rojek et al., 2023)focuses on the use of AI methods, particularly artificial neural networks (ANN), to enhance the supervision of machine failures and support their repair. It addresses the challenges associated with unbalanced training data in real industrial settings and emphasizes the limitations of using supervised machine learning models in such scenarios. The study also proposed future research directions to enhance AI-based maintenance solutions' predictive accuracy and utility in industry, emphasizing the practical challenges that need further investigation.

A study by (Breviglieri et al., 2021b), explored within an in-depth literature review centered on the application of deep learning models for predicting smart grid stability, with a specific emphasis on the Decentral Smart Grid Control (DSGC) system. The study highlights the challenges of integrating renewable energy sources into smart grids and underscores the significance of stability analysis in networked control systems. Acknowledging some limitations, such as the need for more generalization and extension of the analysis to larger grids with more than 10 users, the paper provides valuable perspectives on the complexities of smart grid stability prediction.

A novel unsupervised analysis method for anomaly detection in industrial machinery by (Carratu et al., 2023) using electrical current values and power grid parameters. The framework combines machine learning algorithms and traditional analysis, with a focus on optimizing performance and execution time. It includes a technique for analyzing temporal dynamics based on short-time Fourier transform (STFT) to enhance detection accuracy. Results show exceptional performance, with zero false positives across all datasets tested and less than 4% undetected outlier events, surpassing expert evaluations and other existing methodologies. However, the paper acknowledges dependence on specific features, highlighting a potential limitation for future research with diverse anomaly types.

Fault prediction and location methods are crucial for ensuring the reliability and continuity of energy provision in power systems. (Dashti et al., 2021)delve deep into this realm, exploring both simple weather-based predictions and complex algorithms like support vector machines. They dissect fault location methods for different network types, considering distributed generation, communication quirks, and even measurement timing. While acknowledging limitations like sensor costs and data demands, they paint a promising picture of these methods safeguarding our vital electricity arteries.

A structured approach by (Marcelino et al., 2021) proposed to broaden the application of machine learning models beyond ANNs, illustrating its effectiveness through a case study and its potential for network-level Pavement Management Systems (PMS). This method involves gathering data from sources such as the Long-Term Pavement Performance (LTPP) database, employing imputation techniques for preprocessing, and developing models for 5 and 10-year predictions. The approach is commended for its thoroughness, utilization of the LTPP database, and its potential to enhance PMS predictive capabilities. However, challenges persist, including a historical reliance on ANNs and issues related to data availability and quality.

# CHAPTER 3

# METHODOLOGY

## 3.1 INTRODUCTION

In this chapter, the methodology for the study is outlined. The analysis uses operational and historical data sourced from Kaggle (kaggle, 2024), a widely recognized platform for hosting and sharing datasets across various domains. Following the structured and iterative approach of the Cross Industry Standard Process for Data Mining (CRISP-DM) model(IBM Corporation, 2021), the main objectives are first elucidated. The dataset is then explored and prepared, employing Python for preprocessing tasks such as filtering, training, normalization, validation, and testing being performed to ensure its suitability for modeling. Subsequently, the XGBoost and Random Forest algorithms are applied for classification and In the Evaluation phase, the performance of each algorithm is assessed based on several metrics including Accuracy, Confusion Matrix, AUROC (Area Under the Receiver Operating Characteristics), FPR (False Positive Rate), and TPR (True Positive Rate).

## 3.2 STUDY FRAMEWORK

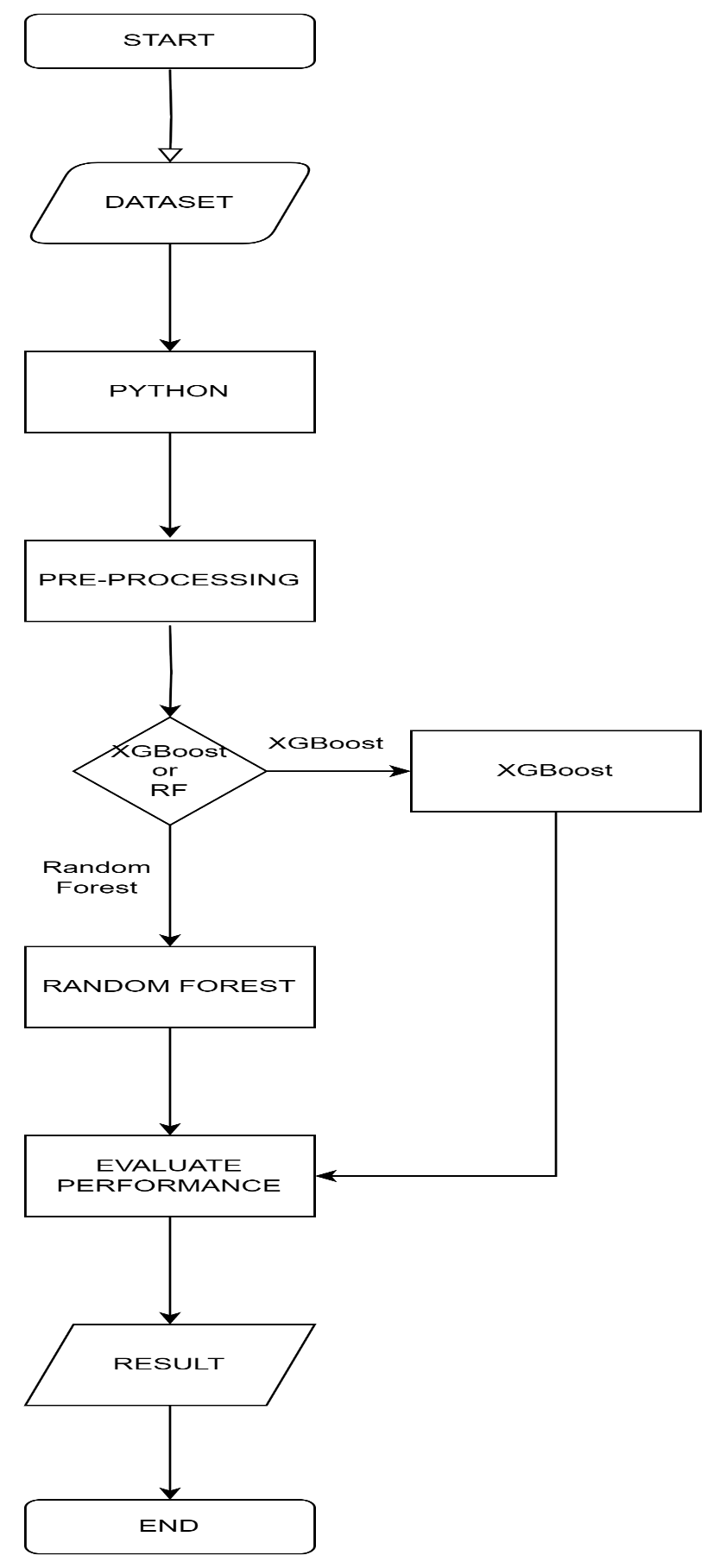


Figure 8: Study Framework

### 3.2.1 DATASET

The "Distributed Transformer Monitoring" dataset was collected via Internet of Things (IoT) devices, the dataset spans from June 25th, 2019, to April 14th, 2020, with updates recorded at 15-minute intervals(Sreshta, 2020). It consists of 20,352 rows and 17 columns, with each row representing a unique observation and each column denoting a specific feature or attribute. The dataset encompasses both numerical and categorical variables, providing comprehensive insights into transformer health and performance.

Table : Dataset Parameters

|  |  |
| --- | --- |
| **Parameter** | **Description** |
| **Independent Parameters** | |
| VL1 | Phase Line 1 |
| VL2 | Phase Line 2 |
| VL3 | Phase Line 3 |
| IL1 | Current Line 1 |
| IL2 | Current Line 2 |
| IL3 | Current Line 3 |
| VL12 | Voltage line 1 2 |
| VL23 | Voltage line 2 3 |
| VL31 | Voltage line 3 1 |
| INUT | Neutral Current |
| OTI | Oil Temperature Indicator |
| WTI | Winding Temperature Indicator |
| ATI | Ambient Temperature Indicator |
| OLI | Oil Level Indicator |
| OTI\_A | Oil Temperature Indicator Alarm |
| OTI\_T | Oil Temperature Indicator |
| **Dependent Parameter** | |
| MOG\_A | Magnetic Oil Guage Alarm |

### 3.2.2 MACHINE LEARNING TOOL

This refers to the environment where our experiment will be carried out.

**PYTHON**

**Python** is a high-level, general-purpose programming language known for its readability and simplicity. Developed by Guido van Rossum and first released in 1991, Python supports multiple programming paradigms, including procedural, object-oriented, and functional programming(Python Software Foundation, 2024). It is widely used in various fields such as web development, data analysis, artificial intelligence, scientific computing, and more(W3schools, 2024). Python offers several relevant libraries:

**Scikit-learn:** This library provides a comprehensive set of machine learning and statistical algorithms to build predictive models, analyze and visualize data, and implement computational statistics(Python Software Foundation, 2024). It includes a suite of supervised and unsupervised machine learning algorithms.

**SciPy**: This library is used for scientific and technical computing. It provides modules for optimization, integration, interpolation, eigenvalue problems, algebraic equations, and other tasks(Python Software Foundation, 2024). It is particularly useful for signal processing, analysis, and algorithm development.

**Matplotlib**: Python’s built-in plotting library is a powerful tool for visualizing data and results. It provides functions for creating a variety of plots, including line, bar, scatter, histogram, and other types of plots(Python Software Foundation, 2024; W3schools, 2024). It also supports 3D plotting for visualizing multivariate data.

### 3.2.3 PRE-PROCESSING DATA

Pre-processing data is a critical step in transforming raw data into a machine-readable format, which is essential for effective utilization by machine learning models (Abbasi, 2021). In this project, this phase is crucial. It involves several steps including data cleaning to remove noise and inconsistencies, normalization to scale numerical features, handling missing values through imputation or removal, and feature selection to identify the most relevant attributes. These steps ensured the data's accuracy and uniformity, enhancing the models' ability to learn and make accurate predictions.

Using Python, this project employed libraries such as Pandas for data manipulation, NumPy for numerical operations, and Scikit-learn for various pre-processing techniques. These tools provide comprehensive functionalities to handle the entire workflow, ensuring the data is in the best possible shape for analysis and model training. This preparation is key to developing reliable and efficient transformer failure prediction models and maintenance practices.

### 3.2.4 MODEL IMPLEMENTATION

To implement the models, Python was used due to its extensive ecosystem of libraries. For XGBoost, **xgboost** library was used, which offers a highly efficient and flexible implementation of the gradient boosting framework. For Random Forest, **scikit-learn** library was utilized, known for its robust and user-friendly ensemble methods. These libraries were selected for their performance, ease of use, and comprehensive documentation.

The next step involved hyperparameter tuning to optimize the models' performance. For XGBoost, key hyperparameters include the learning rate, maximum depth of the trees, and the number of boosting rounds. For Random Forest, important hyperparameters are the number of trees, maximum depth, and the number of features considered for splitting at each node. The project employed grid search and cross-validation techniques to systematically explore different hyperparameter combinations and identify the optimal settings. After tuning, the models were trained using the pre-processed data, which was split into training and testing sets, with 80% allocated for training and 20% for testing. This preparation ensures that the models are trained effectively and can generalize well to unseen data, aiming to contribute to the development of reliable and efficient predictive maintenance strategies for transformers.

### 3.2.5 PERFORMANCE EVALUATION

This is the process of using different performance metrics to evaluate the different machine learning algorithms(Abbasi, 2021). The performance metrics used in this research are Confusion Matrix, Accuracy, Error Rate, AUROC (Area Under the Receiver Operating Characteristics), and F1 Score.  The metrics used to evaluate the performance of machine learning algorithm is very important because the choice of metrics affect the performance of machine learning algorithm.

#### 3.2.4.1 CONFUSION MATRIX

A confusion matrix or error matrix is a machine learning performance metrics used for statistical classification, it consists of table layout that allows visualization of an algorithm based on the accuracy and correctness of the model(Abbasi, 2021). It is use to evaluate the result of the predicted model with the class outcome to see the number of the classes that were correctly classified. Key terms used in confusion matrix include(Abbasi, 2021; Mohammed, 2017):

1. **True Positive (TP):** This is when the actual value was 1(True) and the predicted value is also 1(True).
2. **True Negative (TN):** This when the actual value was 0(False) and the predicted value is 0(False).
3. **False Positive (FP):** This is when the actual value was 0(False) and the predicted value is 1(True).
4. **False Negative (FN):** This is when the actual value was 1(True) and the predicted value is 0(False).

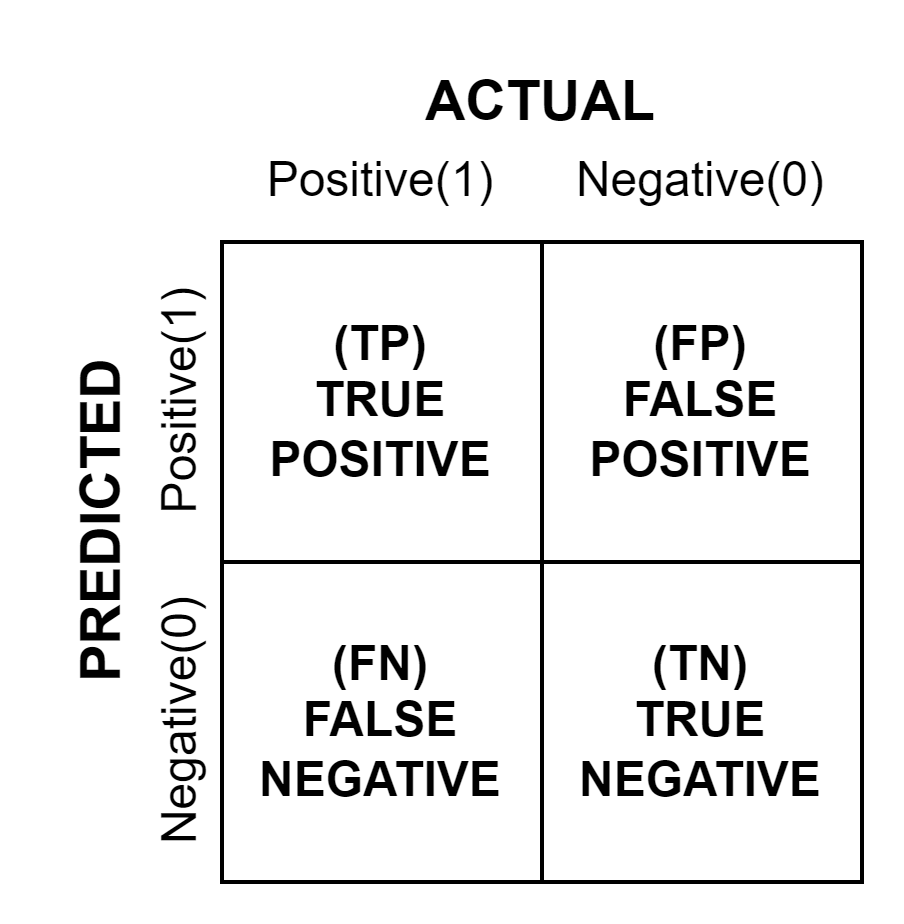


Figure 9: Confusion Matrix

#### 3.2.4.2 ACCURACY

It is the number of correct predictions divided by the total number of datasets(Abbasi, 2021; Mohammed, 2017). The higher the value the more reliable the model is.

|  |  |  |
| --- | --- | --- |
|  |  | **(7)** |

#### 3.2.4.3 PRECISION

It is the ratio of total number of predictive positive to total number of predicted positive(Abbasi, 2021).

|  |  |  |
| --- | --- | --- |
|  |  | **(8)** |

#### 3.2.4.4 RECALL

It is the number of correct positive results divided by the number of allrelevant samples (all samples that should have been identified as positive)(Abbasi, 2021; Mohammed, 2017).

|  |  |  |
| --- | --- | --- |
|  |  | **(9)** |

#### 3.2.4.5 Specificity

Specificity is the ratio of correctly predicted negative observations to all actual negatives. It gives us an idea of how well our model can find all the negative instances. The formula is Abbasi, 2021):

|  |  |  |
| --- | --- | --- |
|  |  | **(10)** |

#### 3.2.4.6 F1-SCORE

The F1 score is a better choice to evaluate the performance of imbalanced datasets. Higher the value of F1 the better the performance of the model(Abbasi, 2021). The value of the F1 score is between ‘0’ and ‘1’.

|  |  |  |
| --- | --- | --- |
|  |  | **(11)** |

#### 3.2.4.7 AUROC

For checking or visualizing the performance of the multi - class classification problem, AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve is use (Abbasi, 2021). It is one of the most important evaluation metrics for checking any classification model’s performance. It is also written as AUROC (Area Under the Receiver Operating Characteristics).

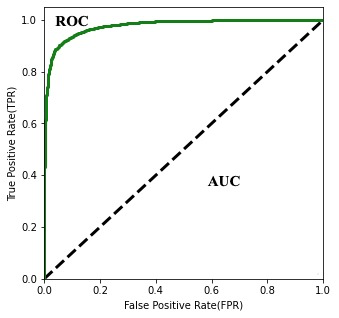


Figure 10: ROC curve (Abbasi, 2021)

# CHAPTER FOUR

# RESULT AND DISCUSSION

## 4.0 INTRODUCTION

This section presents the result and discussion of findings generated from the study. Section 4.1 shows the data analysis and model training, while 4.2 presented the performance evaluation of the ML algorithms.

## 4.1 DATA ANALYSIS AND MODEL IMPLEMENTATION

A transformer dataset containing 21174 instances and 11 attributes was extracted from a CSV file and loaded into a Jupyter Notebook environment. After data collection, an exploratory data analysis (EDA) is carried out on the dataset to  
evaluate and classify the data's key features by means of visualizations, then data cleaning and preparation is carried out before the models are implemented.

### DESCRIPTIVE STATISTICS OF DATA

The dataset comprises 21,174 observations, detailing various parameters relevant to transformer failure prediction. Key features include oil temperature indicators (OTI, ATI, OLI), winding temperature indicator (WTI), and electrical characteristics such as voltages (VL1, VL2, VL3) and currents (IL1, IL2, IL3). Descriptive statistics reveal that OTI has a mean of 30.18°C with a standard deviation of 11.96°C, ranging from 0 to 250°C, while ATI averages 27.74°C with a standard deviation of 5.75°C. OLI shows a mean value of 69.66 with a wider spread, indicating variability in oil levels. Electrical parameters, such as VL1, VL2, and VL3, show close mean values around 240V, with standard deviations near 9V, reflecting consistent voltage levels. Currents IL1, IL2, and IL3 display higher variability, with means around 70A, 56A, and 80A respectively. Additionally, the dataset includes binary variables like OTI\_A and OTI\_T, and measured values for MOG\_A, showcasing diverse data points crucial for predictive modeling.

Table : Descriptive Statistics of Data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **Mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| OTI | 21174 | 30.17786 | 11.961138 | 0 | 26 | 30 | 34 | 250 |
| WTI | 21174 | 0.259564 | 0.438406 | 0 | 0 | 0 | 1 | 1 |
| ATI | 21174 | 27.740059 | 5.750776 | 0 | 24 | 28 | 32 | 44 |
| OLI | 21174 | 69.661519 | 27.79258 | 36 | 40 | 64 | 100 | 100 |
| OTI\_A | 21174 | 0.00477 | 0.068902 | 0 | 0 | 0 | 0 | 1 |
| OTI\_T | 21174 | 0.00222 | 0.047063 | 0 | 0 | 0 | 0 | 1 |
| MOG\_A | 21174 | 0.101681 | 0.302236 | 0 | 0 | 0 | 0 | 1 |
| VL1 | 20652 | 241.023455 | 9.392606 | 0 | 235.8 | 242.4 | 247.3 | 261.2 |
| VL2 | 20652 | 240.490538 | 9.784313 | 0 | 235.5 | 241.9 | 246.4 | 261.3 |
| VL3 | 20652 | 239.923107 | 8.712857 | 0 | 235.5 | 241 | 245.1 | 261.3 |
| IL1 | 20652 | 70.56517 | 42.963096 | 0 | 43.6 | 67.5 | 98.4 | 224.1 |
| IL2 | 20652 | 56.522187 | 41.311167 | 0 | 28.7 | 48.7 | 80.3 | 253.6 |
| IL3 | 20652 | 79.705825 | 45.816941 | 0 | 53.1 | 77.7 | 111.9 | 247.3 |
| VL12 | 20652 | 363.387391 | 140.051283 | 0 | 397.4 | 416.2 | 427.1 | 446.5 |
| VL23 | 20652 | 362.494984 | 139.488304 | 0 | 398.2 | 415.3 | 424.9 | 444.8 |
| VL31 | 20652 | 363.871165 | 140.200657 | 0 | 399 | 416.5 | 427.3 | 447.3 |
| INUT | 20652 | 25.170497 | 15.705378 | 0 | 15.3 | 24.8 | 35.1 | 145.8 |

### CORRELATION MARTRIX

To further understand the data, the correlations between the features were checked to ensure that the correlation between them is not too high, making them suitable for machine learning algorithms and avoiding overfitting or underfitting the models. Figure 11 shows the heatmap of the features in the dataset. Notably, the variables VL1, VL2, VL3, and IL1, IL2, IL3 exhibit strong positive correlations with each other, indicating that they capture similar information about the system's state, which suggests potential redundancy. Moderate correlations are observed between OTI, ATI, and CI1 with other features, hinting at their unique contributions to the dataset. Understanding these correlations is crucial for effective feature selection and engineering, ensuring that the models, particularly XGBoost and Random Forest, do not suffer from multicollinearity, which could lead to overfitting. By strategically selecting and possibly reducing features, the project aims to improve the robustness and accuracy of the predictive models.

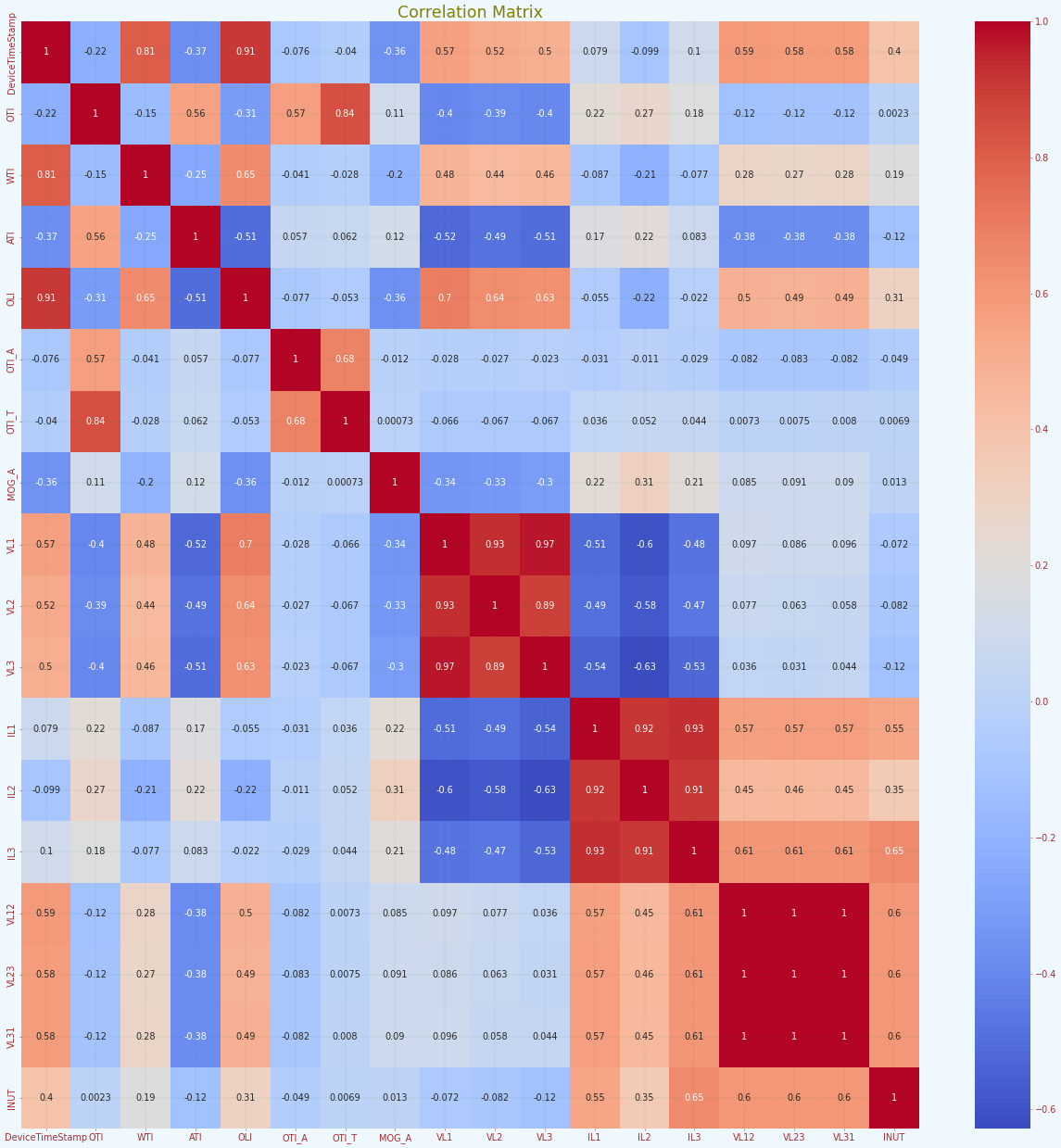
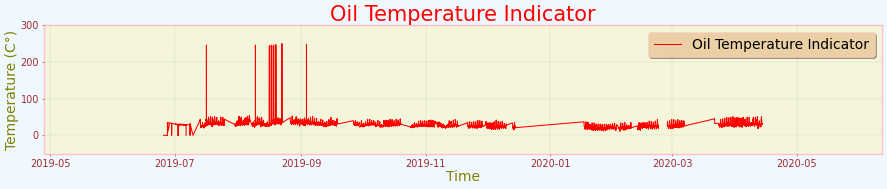
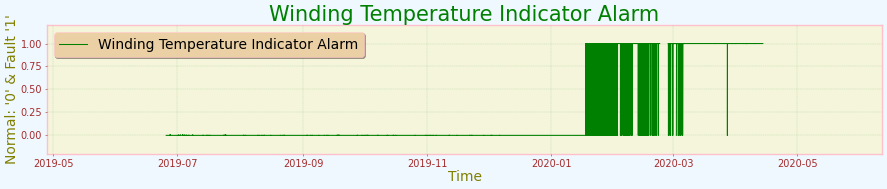
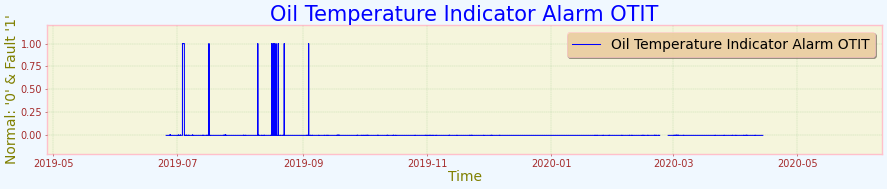
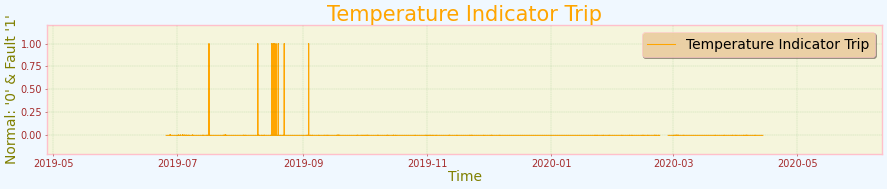
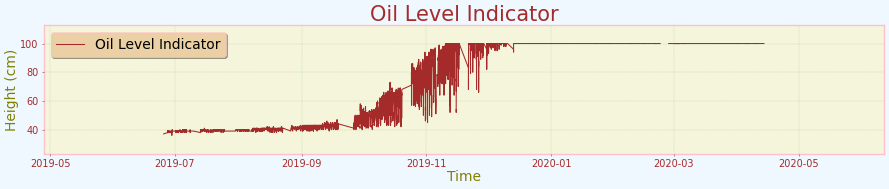
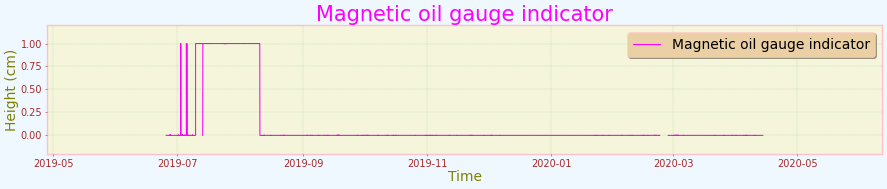


Figure 11: Correlation Matrix

### EXPLORATORY DATA ANALYSIS

Visualizations relating to the distribution of the data to be used for the modeling are provided below.

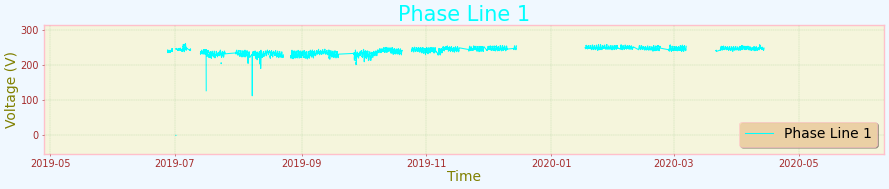
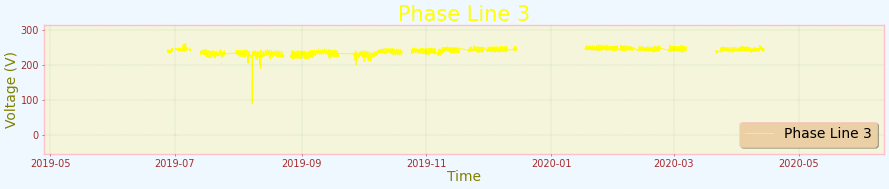
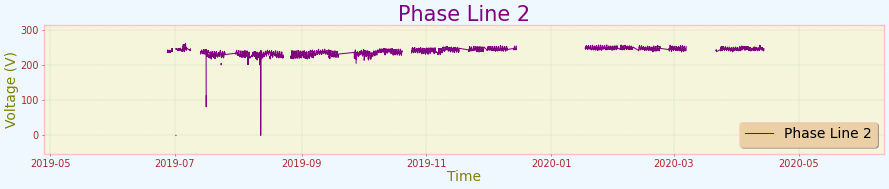
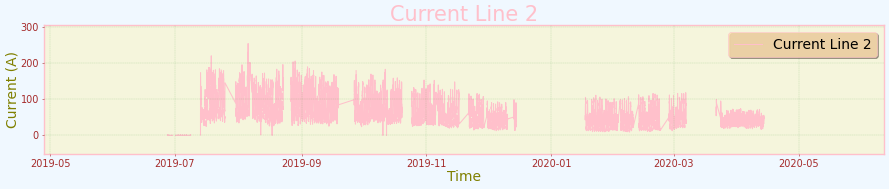
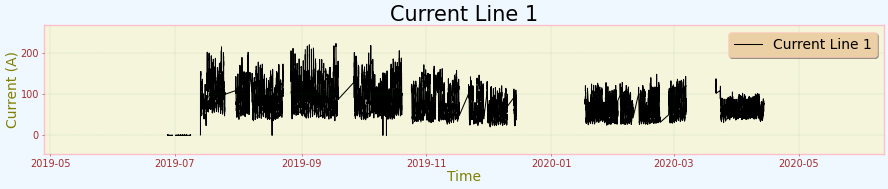
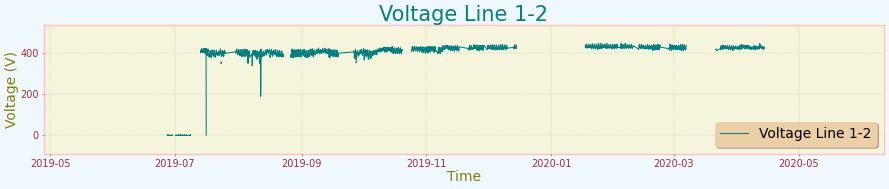
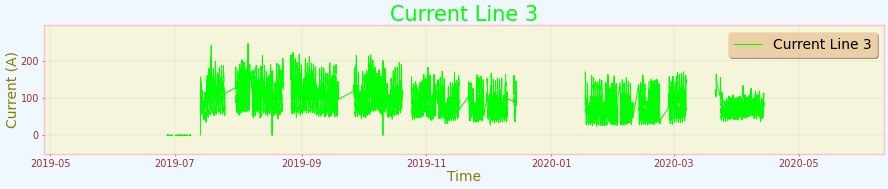
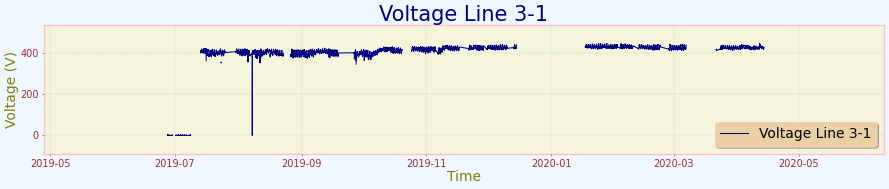
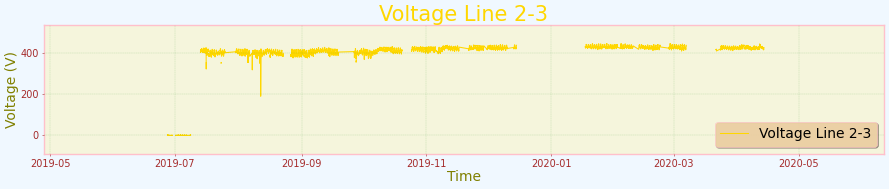
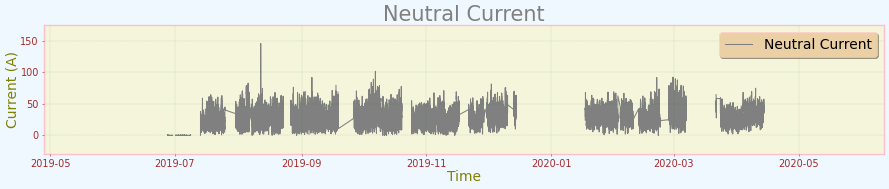
     

Figure 12: Data Visualization

### DATA PREPARATION

In preparing the data for machine learning, it began by examining the data types and checking for any missing values to ensure the dataset’s integrity. Then the data was split into training (80%) and test (20%) subsets to facilitate robust and unbiased model evaluation. The training set was used for fitting the models and tuning hyperparameters, while the test set provided an independent assessment of model performance. To address the variance in attribute scales, normalized the data using Scikit-Learn’s RobustScaler, which is less sensitive to extreme values, ensuring that all features were scaled between 0 and 1 for optimal model performance(Python Software Foundation, 2024).

### MODEL TRAINING AND IMPLEMENTATION

The model training process involved standardizing the data, tuning hyperparameters for both XGBoost and Random Forest classifiers using grid search, and fitting the best models on the training data. Predictions were then made on the test data, and the accuracy of the models was evaluated to ensure they were well-tuned and capable of making accurate predictions on unseen data.

**RANDOM FOREST ALGORITHM**

Table 3:Random Forest Performance

|  |  |  |  |
| --- | --- | --- | --- |
| **RANDOM FOREST PERFOMRANCE SUMMARY AT DIFFERENT THRESHOLD** | | | |
| **Measure** | **Value at 0.5** | **Value at 0.7** | **Value at 0.9** |
| **Accuracy** | 98.43% | 97.99% | 91.74% |
| **F1 score** | 98.45% | 97.94% | 89.01% |
| **Recall** | 98.43% | 97.99% | 91.74% |
| **precision** | 98.49% | 97.95% | 92.43% |
| **Roc** | 99.68% | 99.68% | 99.68% |
| **MCC** | 99.55% | 88.46% | 39.84% |

The Random Forest model demonstrates strong performance at lower thresholds, achieving an accuracy of 98.43%, F1 score of 98.45%, recall of 98.43%, and precision of 98.49% at a threshold of 0.5. These metrics slightly decrease to 97.99% accuracy, 97.94% F1 score, 97.99% recall, and 97.95% precision at a threshold of 0.7. However, a substantial performance drop is observed at a threshold of 0.9, with accuracy falling to 91.74%, F1 score to 89.01%, recall to 91.74%, and precision to 92.43%.

While the ROC AUC consistently maintains a high value of 99.68% across all thresholds, indicating good overall discriminative power, the MCC (Matthews Correlation Coefficient) undergoes a dramatic decline from 99.55% at 0.5 to 88.46% at 0.7 and further to 39.84% at 0.9. This suggests a significant reduction in the model's predictive ability at higher thresholds.

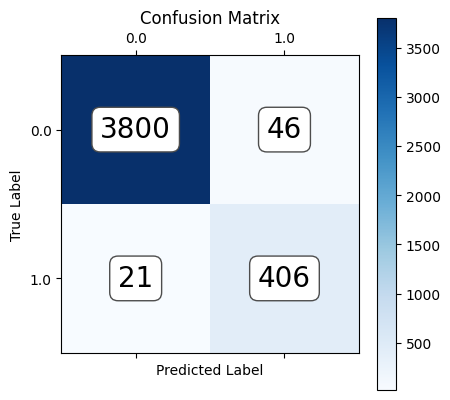
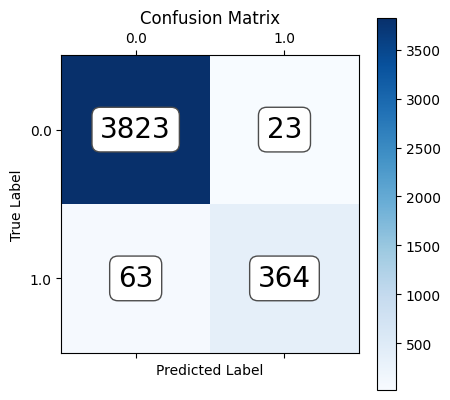
 

Figure 13: RF CM at 0.9 Figure 14: RF CM at 0.7

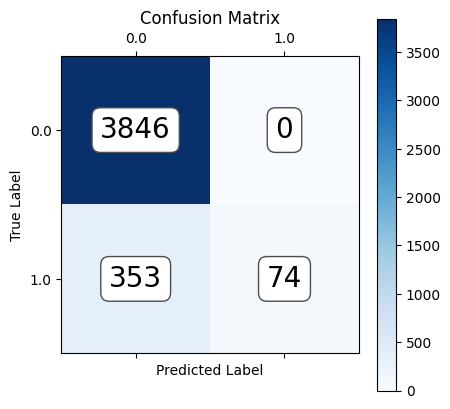
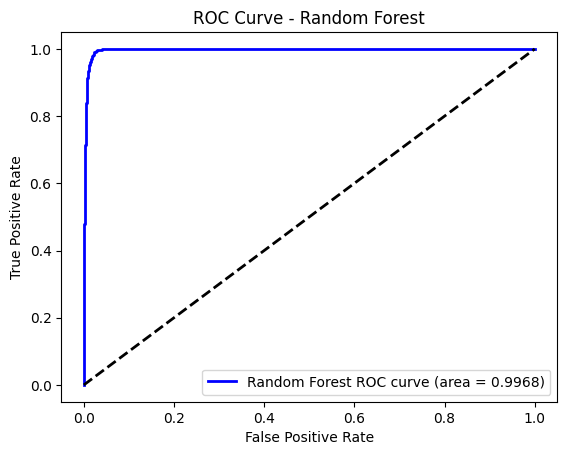
 

Figure 15: RF CM at 0.9 Figure 16: RF ROC

**XGBOOST ALGORITHM**

Table :XGBOOST Performance

|  |  |  |  |
| --- | --- | --- | --- |
| **XGBOOST PERFOMRANCE SUMMARY AT DIFFERENT THRESHOLD** | | | |
| **Measure** | **Value at 0.5** | **Value at 0.7** | **Value at 0.9** |
| **Accuracy** | 98.5% | 98.36% | 97.52% |
| **F1 score** | 98.51% | 98.34% | 97.40% |
| **Recall** | 98.50% | 98.36% | 97.52% |
| **precision** | 98.53% | 98.34% | 97.50% |
| **Roc** | 99.74% | 99.74% | 99.74% |
| **MCC** | 91.83% | 90.76% | 85.50% |

The XGBoost model demonstrates consistent performance across different thresholds, maintaining high accuracy, F1 score, recall, and precision levels. At a threshold of 0.5, the model achieves an accuracy of 98.5%, F1 score of 98.51%, recall of 98.50%, and precision of 98.53%. While these metrics slightly decline to 98.36% accuracy, 98.34% F1 score, 98.36% recall, and 98.34% precision at a threshold of 0.7, and further to 97.52% accuracy, 97.40% F1 score, 97.52% recall, and 97.50% precision at a threshold of 0.9, the overall performance remains strong.

The ROC AUC consistently stays at 99.74% across all thresholds, indicating excellent discriminative power. However, the MCC (Matthews Correlation Coefficient) shows a gradual decrease from 91.83% at 0.5 to 90.76% at 0.7 and 85.50% at 0.9, suggesting a slight reduction in the model's predictive ability as the threshold increases.

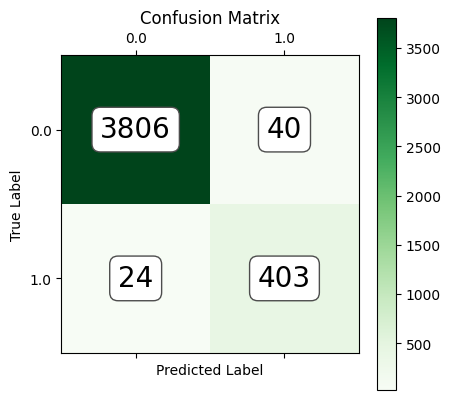
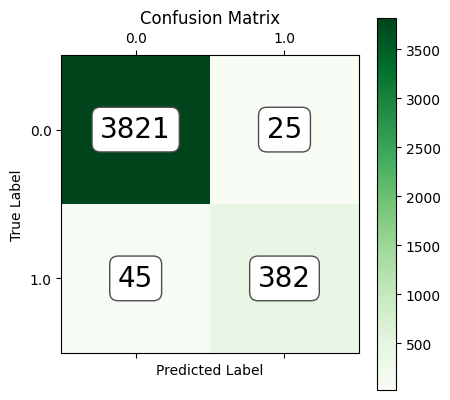
 

Figure 17: XGBoost CM at 0.5 Figure 18: XGBoost CM at 0.7

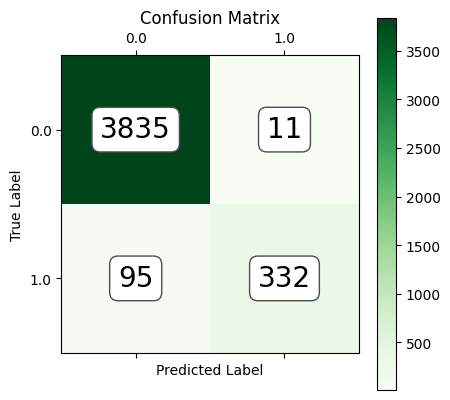
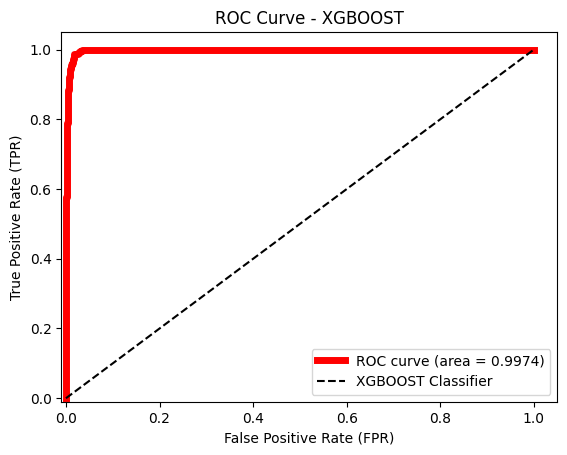
 

Figure 19: XGBoost CM at 0.9 Figure 20: XGBoost ROC

## 4.2 MODEL COMPARISON

Accuracy, Confusion Matrix, AUROC (Area Under the Receiver Operating Characteristics), Precision, Recall and MCC (Matthews Correlation Coefficient) are metrics used to evaluate the performance of both models.

Table 5: Model Comparison

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Measure** | **Random Forest (0.5)** | **Random Forest (0.7)** | **Random Forest (0.9)** | **XGBoost (0.5)** | **XGBoost (0.7)** | **XGBoost (0.9)** |
| **Accuracy** | 98.43 | 97.99 | 91.74 | 98.5 | 98.36 | 97.52 |
| **F1 score** | 98.45 | 97.94 | 89.01 | 98.51 | 98.34 | 97.4 |
| **Recall** | 98.43 | 97.99 | 91.74 | 98.5 | 98.36 | 97.52 |
| **Precision** | 98.49 | 97.95 | 92.43 | 98.53 | 98.34 | 97.5 |
| **ROC AUC** | 99.68 | 99.68 | 99.68 | 99.74 | 99.74 | 99.74 |
| **MCC** | 99.55 | 88.46 | 39.84 | 91.83 | 90.76 | 85.5 |

Both Random Forest and XGBoost models achieve high accuracy, F1 score, recall, and precision at a threshold of 0.5. However, the performance of the Random Forest model drops significantly at higher thresholds (0.7 and 0.9), while XGBoost maintains a more consistent performance across all thresholds. This is reflected in the MCC metric, which shows a much sharper decline for Random Forest compared to XGBoost.

Overall, XGBoost appears to be a more robust model, as it is less sensitive to the choice of the threshold and delivers consistently good performance across different thresholds. However, if a high true positive rate is critical and a higher false positive rate is acceptable, then the Random Forest model might be a good choice at a lower threshold. The ROC comparison of the models.

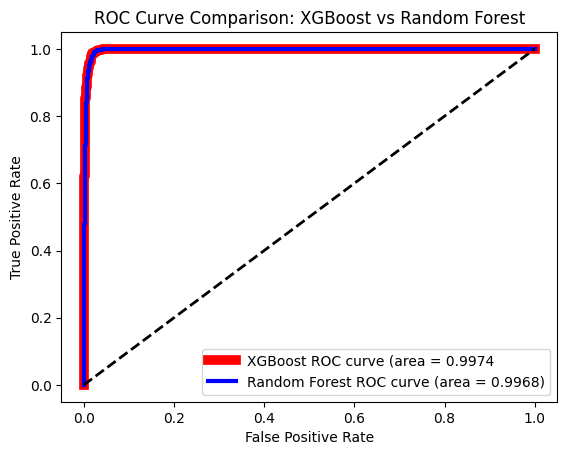


Figure 21: ROC Curve Comparison

# CHAPTER FIVE

# CONCLUSION AND RECOMENDATIONS

## 5.1 CONCLUSION

This study has demonstrated the feasibility and effectiveness of utilizing machine learning algorithms, specifically Random Forest and XGBoost, for predicting transformer failures using operational and historical data. The analysis revealed the superior performance of XGBoost across all evaluated metrics. While both models achieved high accuracy, precision, recall, and F1-score, XGBoost consistently outperformed Random Forest, particularly at higher thresholds. The significantly higher MCC value for XGBoost reinforces its superior predictive capability.

These findings suggest that XGBoost is a more robust and reliable model for transformer failure prediction. By effectively leveraging its strengths, utilities can significantly enhance grid reliability, optimize maintenance schedules, and reduce operational costs associated with unplanned outages. Future research should explore additional algorithms and incorporate a broader range of data sources to further enhance predictive model robustness and improve electrical grid operations.

## 5.2 RECOMMENDATION

Based on the findings of this research, the following recommendations are proposed:

1. **Implementation of XGBoost for Predictive Maintenance**: Given its superior performance, organizations should consider implementing the XGBoost algorithm for predictive maintenance of transformers. This can lead to improved reliability and reduced downtime, ultimately enhancing operational efficiency.
2. **Further Research on Hybrid Models**: Future research could explore the development of hybrid models that combine the strengths of both Random Forest and XGBoost. Such models may leverage the interpretability of Random Forest while benefiting from the predictive power of XGBoost.
3. **Real-World Application and Testing**: It is recommended that the algorithms be tested in real-world scenarios to validate their effectiveness in diverse operational conditions. This practical application can provide valuable insights into their performance and adaptability.
4. **Continuous Data Collection and Model Updating**: Organizations should establish a continuous data collection framework to ensure that the predictive models remain relevant and accurate over time. Regular updates to the models can help in adapting to changing operational conditions and emerging failure patterns.
5. **Training and Capacity Building**: To maximize the benefits of predictive maintenance, it is essential to invest in training personnel on the use of machine learning algorithms and data analytics. This will empower teams to make informed decisions based on predictive insights.

In conclusion, the comparative analysis of Random Forest and XGBoost algorithms for transformer failure prediction has provided valuable insights that can significantly enhance predictive maintenance strategies. By adopting the recommendations outlined in this chapter, organizations can improve their operational reliability and efficiency in managing transformer assets.

## 5.3 LIMITATIONS

This study has several limitations that may affect the generalizability and applicability of the findings. Firstly, the research relied on a specific dataset that may not cover all possible transformer failure scenarios, potentially limiting the robustness of the predictive models. Additionally, data quality and completeness could impact prediction accuracy, as missing or erroneous data may lead to biased results. Lastly, the study focused solely on XGBoost and Random Forest algorithms, potentially overlooking other effective machine learning techniques. These limitations suggest that further research with a broader dataset and additional algorithms is necessary to validate and enhance predictive maintenance strategies for transformers.

# REFERENCES

Abbasi, J. A. (2021). *Predictive Maintenance in Industrial Machinery using Machine Learning*.

Breviglieri, P., Erdem, T., & Eken, S. (2021a). Predicting Smart Grid Stability with Optimized Deep Models. *SN Computer Science*, *2*(2). https://doi.org/10.1007/s42979-021-00463-5

Breviglieri, P., Erdem, T., & Eken, S. (2021b). Predicting Smart Grid Stability with Optimized Deep Models. *SN Computer Science*, *2*(2). https://doi.org/10.1007/s42979-021-00463-5

British Standards Institution, T. (2018). *BSI Standards Publication*.

Carratu, M., Gallo, V., Iacono, S. Dello, Sommella, P., Bartolini, A., Grasso, F., Ciani, L., & Patrizi, G. (2023). A Novel Methodology for Unsupervised Anomaly Detection in Industrial Electrical Systems. *IEEE Transactions on Instrumentation and Measurement*, *72*. https://doi.org/10.1109/TIM.2023.3318684

Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. da P., Basto, J. P., & Alcalá, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers and Industrial Engineering*, *137*. https://doi.org/10.1016/j.cie.2019.106024

Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang, C. H., & Liu, R. (2019). XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System. *IEEE Access*, *7*, 13149–13158. https://doi.org/10.1109/ACCESS.2019.2893448

Çinar, Z. M., Nuhu, A. A., Zeeshan, Q., Korhan, O., Asmael, M., & Safaei, B. (2020). Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0. *Sustainability (Switzerland)*, *12*(19). https://doi.org/10.3390/su12198211

Coandǎ, P., Avram, M., & Constantin, V. (2020). A state of the art of predictive maintenance techniques. *IOP Conference Series: Materials Science and Engineering*, *997*(1). https://doi.org/10.1088/1757-899X/997/1/012039

Dashti, R., Daisy, M., Mirshekali, H., Shaker, H. R., & Hosseini Aliabadi, M. (2021). A survey of fault prediction and location methods in electrical energy distribution networks. *Measurement: Journal of the International Measurement Confederation*, *184*. https://doi.org/10.1016/j.measurement.2021.109947

Hussain, M. R., Refaat, S. S., & Abu-Rub, H. (2021). Overview and Partial Discharge Analysis of Power Transformers: A Literature Review. In *IEEE Access* (Vol. 9, pp. 64587–64605). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2021.3075288

IBM Corporation. (2021). *CRISP-DM Help Overview*. https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3). https://doi.org/10.1007/s12525-021-00475-2

kaggle. (2024). *kaggle webpage* . https://www.kaggle.com/

Khan, K., Ahmad, W., Amin, M. N., Ahmad, A., Nazar, S., & Alabdullah, A. A. (2022). Compressive Strength Estimation of Steel-Fiber-Reinforced Concrete and Raw Material Interactions Using Advanced Algorithms. *Polymers*, *14*(15). https://doi.org/10.3390/polym14153065

Liu, T., Song, Y., Zhu, L., & Hill, D. J. (2022). *Stability and Control of Power Grids*. https://doi.org/10.1146/annurev-control-042820

Marcelino, P., de Lurdes Antunes, M., Fortunato, E., & Gomes, M. C. (2021). Machine learning approach for pavement performance prediction. *International Journal of Pavement Engineering*, *22*(3), 341–354. https://doi.org/10.1080/10298436.2019.1609673

MathWorks, I. (2024). *matlab doc*. https://www.mathworks.com/help/matlab/

Mohammed, S. (2017, November 11). *Performance Metrics for Classification problems in Machine Learning*. Medium.Com. https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Rojek, I., Jasiulewicz-Kaczmarek, M., Piechowski, M., & Mikołajewski, D. (2023). An Artificial Intelligence Approach for Improving Maintenance to Supervise Machine Failures and Support Their Repair. *Applied Sciences (Switzerland)*, *13*(8). https://doi.org/10.3390/app13084971

Sreshta, P. (2020). *Distributed Transformer Monitoring*. Distributed Transformer Monitoring. https://www.kaggle.com/datasets/sreshta140/ai-transformer-monitoring

Tianjin da xue, Zhongguo dian ji gong cheng xue hui (Beijing, C., Guo jia dian wang gong si (China), IEEE Power & Energy Society, Institution of Engineering and Technology, International Council on Large Electric Systems, Institute of Electrical and Electronics Engineers, & International Conference on Electricity Distribution. Chinese National Committee, organizer. (2018a). *2018 China International Conference on Electricity Distribution : proceedings : 17-19 September 2018, Tianjin, China*.

Tianjin da xue, Zhongguo dian ji gong cheng xue hui (Beijing, C., Guo jia dian wang gong si (China), IEEE Power & Energy Society, Institution of Engineering and Technology, International Council on Large Electric Systems, Institute of Electrical and Electronics Engineers, & International Conference on Electricity Distribution. Chinese National Committee, organizer. (2018b). *2018 China International Conference on Electricity Distribution : proceedings : 17-19 September 2018, Tianjin, China*.

Wang, T., Li, Q., Yang, J., Xie, T., Wu, P., & Liang, J. (2023). Transformer Fault Diagnosis Method Based on Incomplete Data and TPE-XGBoost. *Applied Sciences (Switzerland)*, *13*(13). https://doi.org/10.3390/app13137539

Zhang, R., Li, B., & Jiao, B. (2019). Application of XGboost Algorithm in Bearing Fault Diagnosis. *IOP Conference Series: Materials Science and Engineering*, *490*(7). <https://doi.org/10.1088/1757-899X/490/7/072062>

Python Software Foundation. (2024). *Scikit-learn Documentation*. Https://Www.Python.Org/Doc/Essays/Blurb/. https://www.python.org/doc/

W3schools. (2024). *W3schools*. https://www.w3schools.com/python/python\_intro.asp