**Name:** Abdulrahman AbdulKareem

**Title:** Comparative Analysis of XGBoost and Random Forest Algorithms for Transformer Failure Prediction using Grid Stability Data

# 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. Instead of directly targeting transformers, it focuses on the untapped potential of electrical grid stability data. This treasure trove of information holds hidden clues about the health of individual components like transformers. By employing powerful machine learning algorithms such as Random Forest and Extreme Gradient Boosting (XGBoost), this research deciphers these clues and identifies patterns that might foreshadow transformer failures. Analyzing metrics like voltage fluctuations, frequency deviations, and load imbalances could allow us to 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 power grids. 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 PdM, 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 the project is to demonstrate the feasibility and advantages of utilizing grid stability data for indirect transformer failure prediction via a comparative analysis of Random Forest and Extreme Gradient Boosting (XGBoost) algorithms, ultimately aiming to establish a reliable PdM strategy for electrical grids. The objectives of the project are:

1. Collect and analyze comprehensive transformer grid stability 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 and create a robust predictive model.

## 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 unexplored realm of indirect prediction using grid stability data, an abundant, readily available source. 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 grid resilience.

## SCOPE OF STUDY

The potential of grid stability data for indirectly predicting transformer failures is explored within this research, paving the way for a proactive and cost-effective PdM strategy. Through meticulous comparison of the performance of Random Forest and XGBoost algorithms, the hidden insights within grid metrics like voltage fluctuations, frequency deviations, and load imbalances are sought to be unlocked. This endeavor aims to enable anticipation of transformer failures before they disrupt power supply, ultimately focusing on harnessing readily available data instead of expensive and sparse sensor coverage. The potential outcome is a more resilient, efficient, and optimized grid(Çinar et al., 2020; Janiesch et al., 2021).

## 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).

**Electrical Grid Stability Data**

This refers to the diverse set of data collected from the electrical grid, including voltage levels, current flows, frequency variations, and other operational parameters(Breviglieri et al., 2021a).

**Indirect Prediction of Transformer Failures**

Rather than directly monitoring the transformers, the project aims to use the patterns and anomalies observed in the electrical grid stability data to infer potential issues with transformers. By identifying correlations and trends in the grid 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 PdM research for transformers, focusing on indirect sensor-based approaches and analyzing previous research utilizing AI algorithms for transformer failure prediction and identify gaps in knowledge.

**Chapter 3:** Methodology

This chapter describes the data collection process for both grid stability metrics and transformer failure records. Explaining the preprocessing techniques used to ensure data quality and prepare it for analysis. Also, specifying the evaluation metrics to be used for comparing the performance of the two 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 grid stability data.

**Chapter 5:** Conclusion and Recommendation

This chapter summarizes the key findings of the project, emphasizing the feasibility and advantages of using grid stability data for indirect transformer failure prediction. Also to draw conclusions about the effectiveness of Random Forest and XGBoost in this specific context, making recommendations for future PdM applications.

# 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.

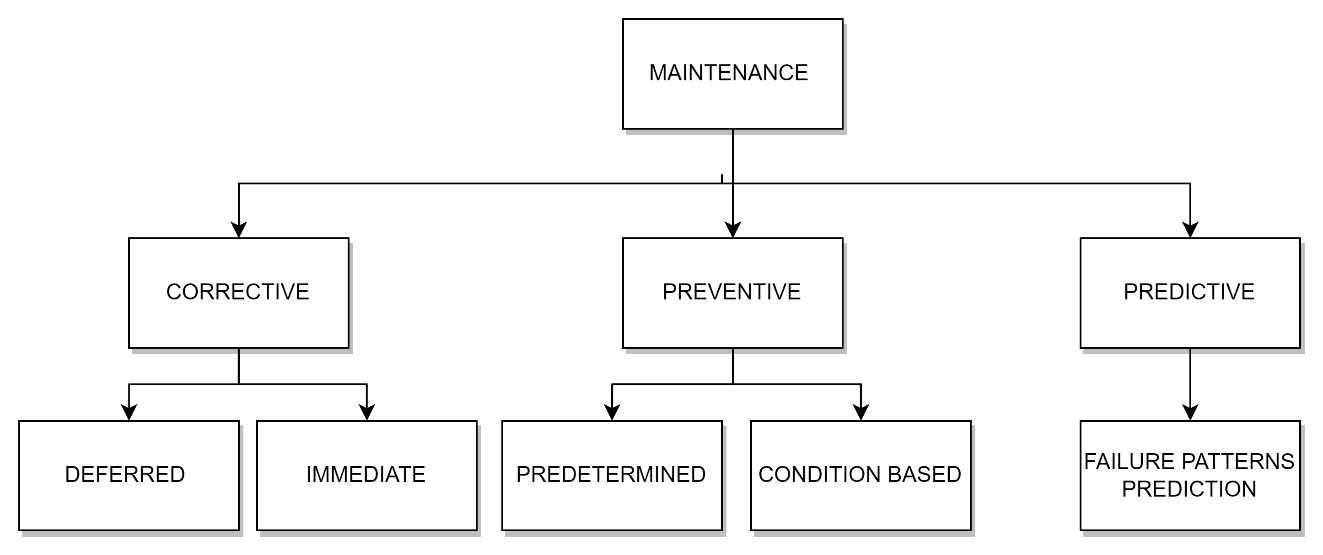


Figure 1: Types of maintenance according to EN 13306 standard

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**.

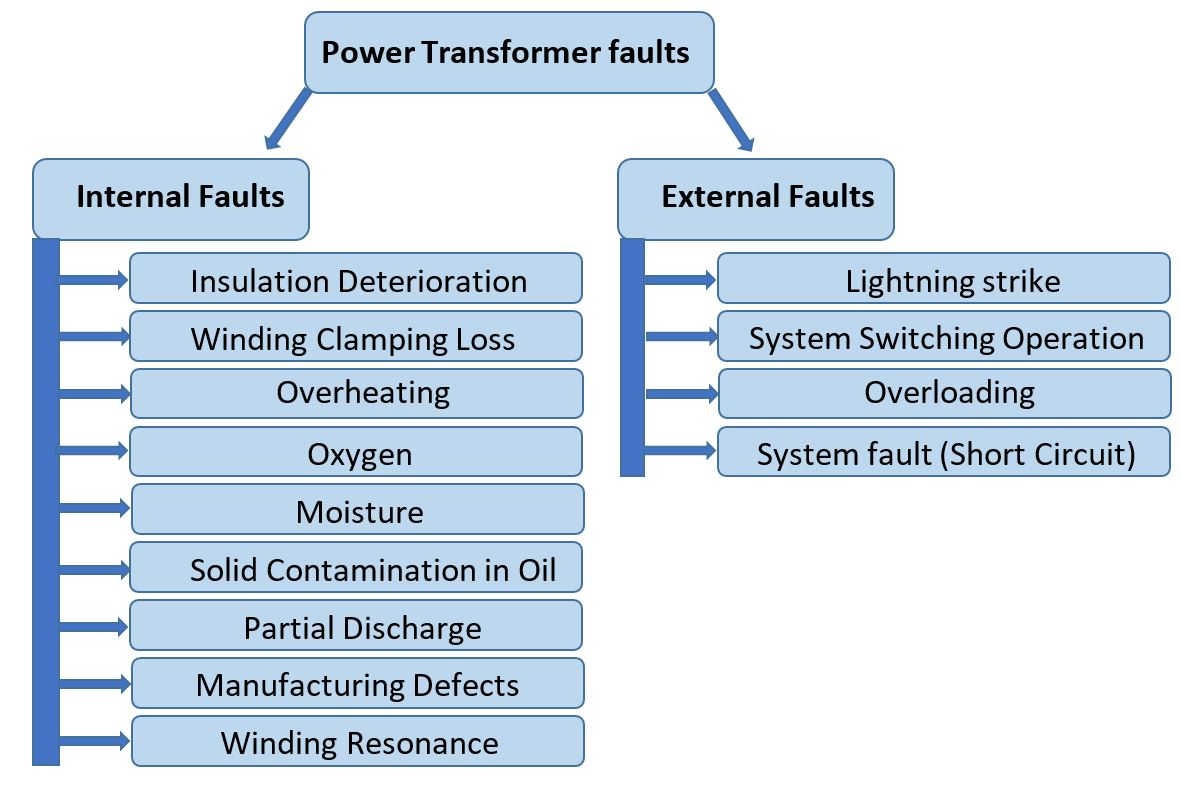


Figure 2: Electrical Faults of Transformer(Hussain et al., 2021)

**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 grid stability 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 power grid(Tianjin da xue et al., 2018b).

### PRIDICTIVE MAINTENANCE AND ELECTRICAL GRID STABILITY DATA

Electrical grid stability data refers to the information that reflects the condition of the electrical grid, including parameters such as voltage, current, and frequency. This data is crucial as it ensures the grid’s safe and efficient operation, and guarantees the delivery of power as required. By monitoring this data, utilities and grid operators 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.

**Voltage Data:** This type of data includes measurements of voltage levels at various points within the electrical grid. Monitoring voltage levels is critical for ensuring proper equipment function and power quality(Liu et al., 2022). Variations in voltage can indicate issues such as overloading or equipment malfunctions, affecting the reliability of electricity supply.

**Frequency Data:** Frequency data involves measurements of the frequency of alternating current (AC) within the grid. Grid frequency is typically maintained at a constant value, and deviations can signal imbalances in supply and demand(Liu et al., 2022). Frequency data helps operators regulate grid stability and address issues like overloading or generation shortages promptly.

**Load Data:** Load data pertains to measurements of electricity demand or load levels across the grid. Understanding load patterns helps operators anticipate peak demand periods, plan for capacity requirements, and optimize resource allocation(Liu et al., 2022). It provides valuable insights into consumer behavior and usage trends, guiding effective grid management strategies.

**Power Quality Data:** Power quality data encompasses measurements of various parameters related to the quality and reliability of electrical power(Liu et al., 2022). This includes factors like voltage harmonics, voltage unbalance, and transient voltage fluctuations. Monitoring power quality ensures that electricity meets acceptable standards and helps identify issues affecting grid performance.

Electrical grid stability data can provide valuable insights for predictive maintenance programs. By monitoring stability 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:

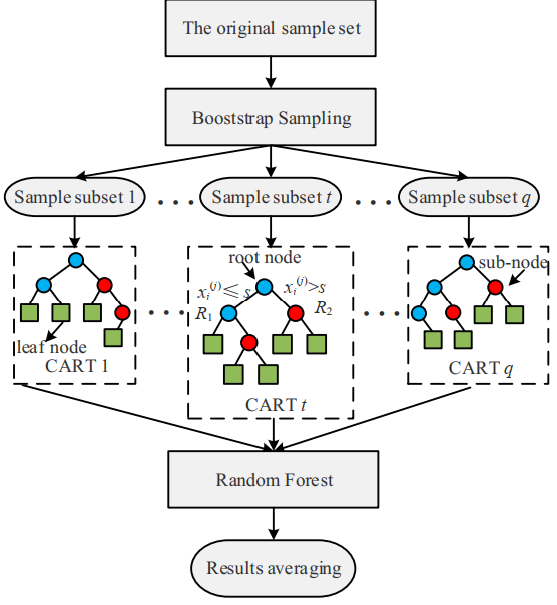


Figure 6: Random Forest Algorithm(Wang et al., 2023)

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:

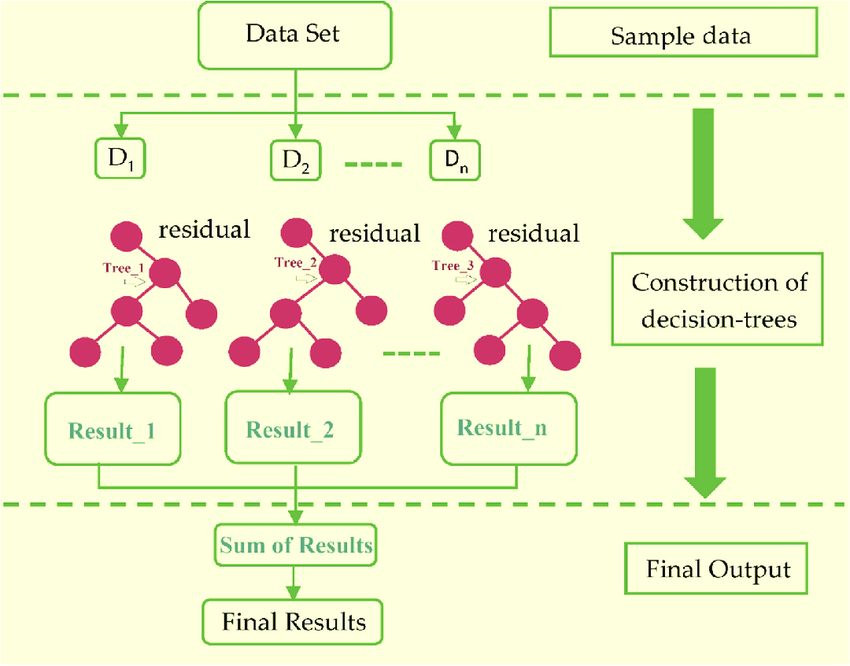


Figure 7: Extreme Gradient Boosting Algorithm(Khan et al., 2022)

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.

Insight are drawn from 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, insights are drawn from a study by Rojek et al., 2023). The study 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.

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, findings are derived from a study by (Breviglieri et al., 2021b).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.

(Marcelino et al., 2021) propose a structured approach 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 electrical grid stability data sourced from Kaggle, 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 business objectives are first elucidated. The dataset is then explored and prepared, employing MATLAB 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

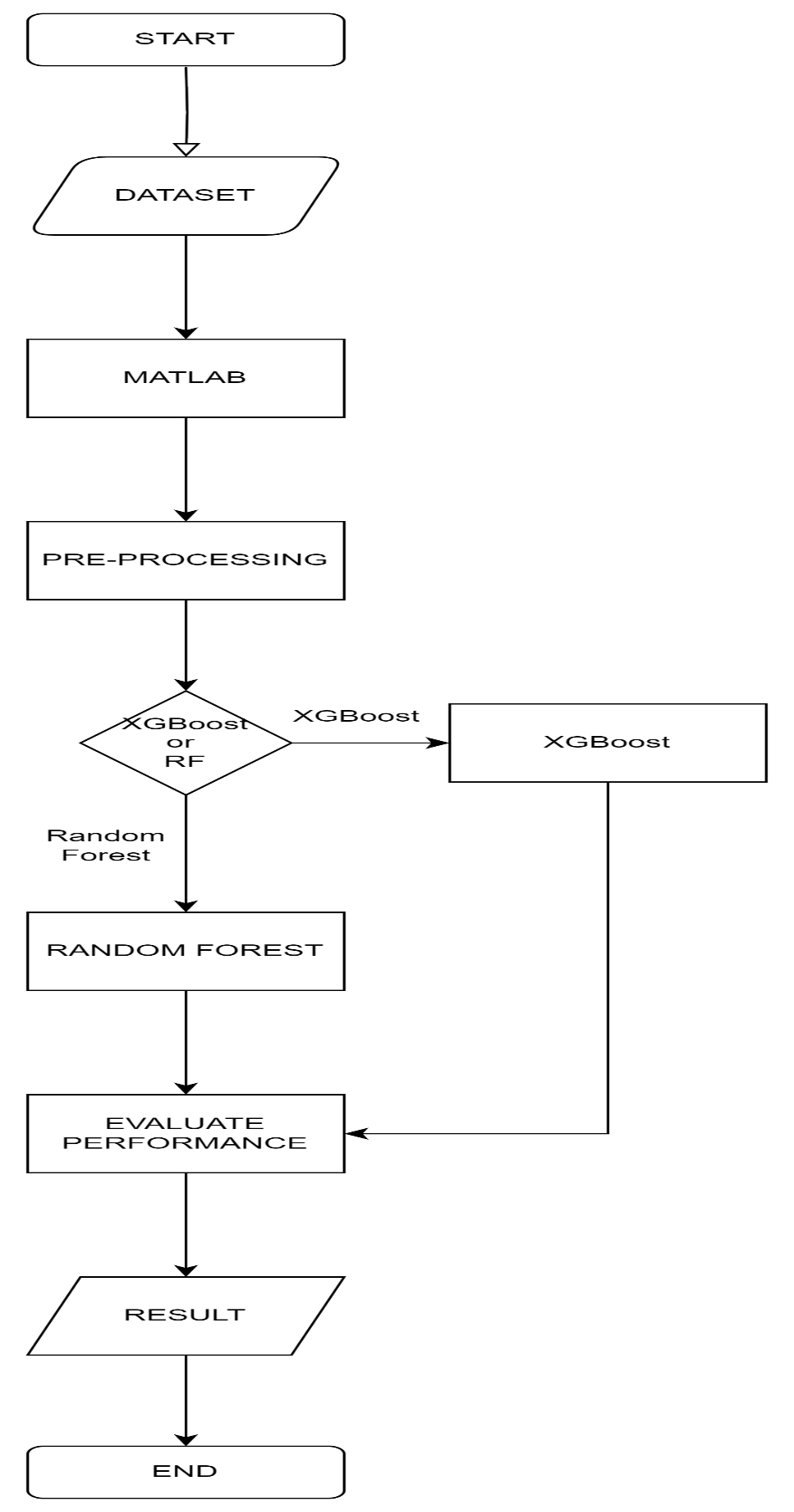


Figure 8: Study Framework

### 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. It consists of 19,352 rows and 10 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.

|  |  |
| --- | --- |
| **Parameter** | **Description** |
| **Grid stability data overview** | |
| 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 |
| **Internal transformer fault overview** | |
| 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 Trip |
| MOG\_A | Magnetic Oil Gauge Alarm |

Table 1: Dataset parameters

## 3.2.2 MACHINE LEARNING TOOL

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

**MATLAB**

MATLAB, an abbreviation for MATrix LABoratory, is a proprietary multi-paradigm programming language and numeric computing environment developed by MathWorks(MathWorks, 2024). It is designed for engineers and scientists, and is widely used for numerical computation, visualization, and algorithm development. It provides an extensive range of tools and functionalities that are particularly useful for matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages. MATLAB offers several relevant libraries(MathWorks, 2024)

1. **Statistics and Machine Learning Toolbox:** This toolbox provides a comprehensive set of machines learning and statistical algorithms to build predictive models, analyze and visualize data, and implement computational statistics(MathWorks, 2024). It includes a suite of supervised and unsupervised machine learning algorithms.
2. **Signal Processing Toolbox:** This toolbox is used for signal processing, analysis, and algorithm development(MathWorks, 2024). It provides functions and apps to generate, measure, transform, filter, and visualize signals.
3. **MATLAB Plotting Library:** MATLAB’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. It also supports 3D plotting for visualizing multivariate data(MathWorks, 2024).  it can be used to plot the performance metrics of algorithms.

### 3.2.3 PRE-PROCESSING DATA

Pre-processing data is the process of transforming or preparing data so that it can be readable by the machine(Abbasi, 2021; MathWorks, 2024). Data does not necessarily mean large volume of datasets with rows and columns, it could be in different format like videos, audio, images and many more. Machines can’t understand plain texts, videos, images and audios the way humans do, such data has to go through series of processes so that the machine can understand representing these data with 0s and 1s which is known as machine language(Abbasi, 2021).

#### 3.2.3.1 LOAD/IMPORT DATASET

MATLAB is straightforward using built-in functions like readtable for tabular data or imread for image data(MathWorks, 2024). MATLAB also supports importing data from various file formats such as CSV, Excel, HDF5, and more(MathWorks, 2024). After loading your dataset, you notice the columns of the data is either nominal or numeric and we can further proceed by checking for missing values in the dataset.

#### 3.2.3.2 MISSING VALUE

It is always common to have missing values in dataset which occur during data collection or human error. Handling missing values in MATLAB involves techniques like data imputation, where missing values are replaced with estimated values based on surrounding data points or statistical methods. MATLAB provides functions like fillmissing for automatic missing value imputation, ensuring data integrity and quality for subsequent analysis(MathWorks, 2024). Another method for replacing missing values is applying -1 or 0 for numeric values or NULL for string values.

#### 3.2.3.3 TRAIN, VALIDATE AND TEST

The training, validation, and testing of machine learning models such as XGBoost and Random Forest are conducted in a structured process crucial for the development of accurate predictive models. During training, labeled data is processed by these algorithms to learn patterns and optimize predictive performance(Abbasi, 2021; Breviglieri et al., 2021b). Model performance is assessed and hyperparameters are fine-tuned using validation data, ensuring robustness and generalization. Subsequently, testing data is utilized for an impartial evaluation of the predictive capabilities of the models, with MATLAB offering tools for computing various performance metrics, thereby facilitating thorough model evaluation and refinement(MathWorks, 2024).

### 3.2.4 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).

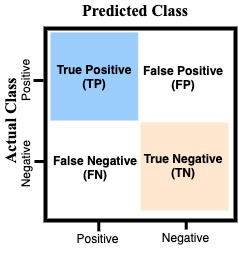


Figure 9: Confusion Matrix(Abbasi, 2021)

#### 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)** |

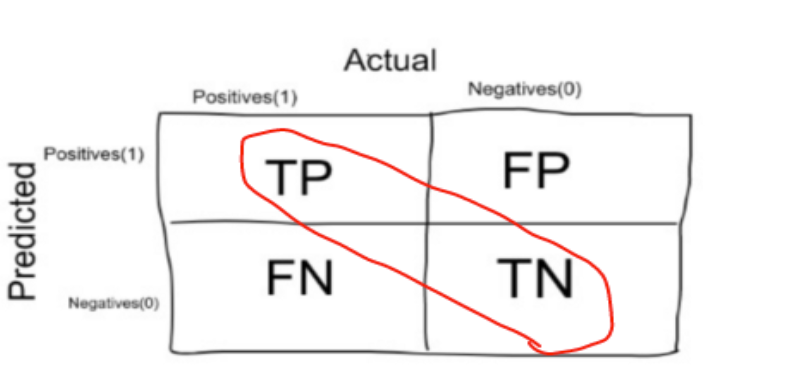


Figure 10: Accuracy Matrix(Mohammed, 2017)

#### 3.2.4.3 PRECISION

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

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

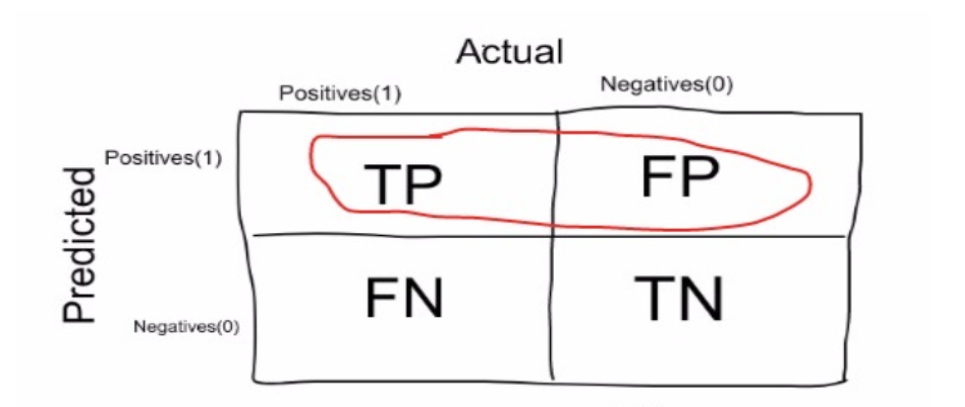


Figure 11: Precision Matrix(Mohammed, 2017)

#### 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) |

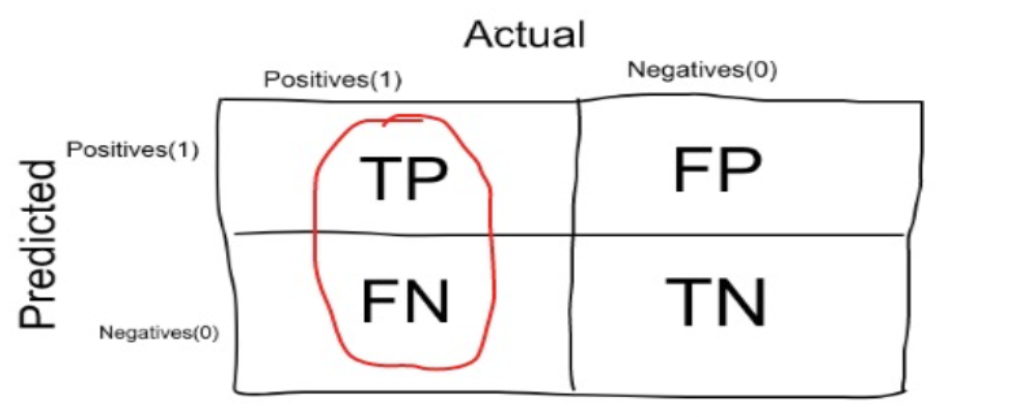


Figure 12: Recall(Mohammed, 2017)

#### 3.2.4.5 Specificity

Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative)(Abbasi, 2021).

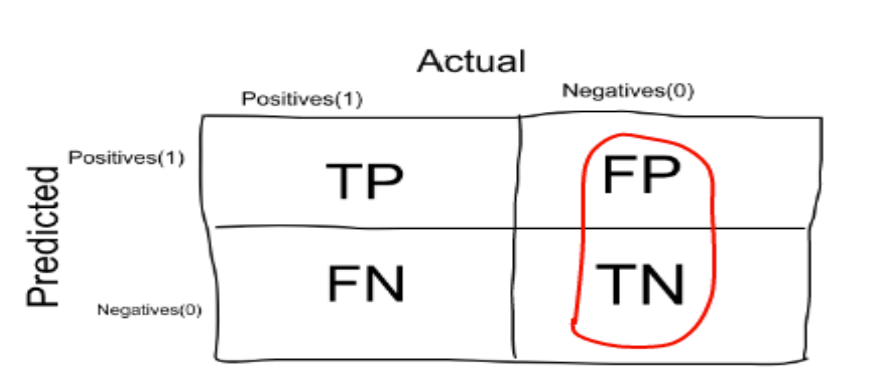


Figure 13: Specificity Matrix(Mohammed, 2017)

#### 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’.

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

#### 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).

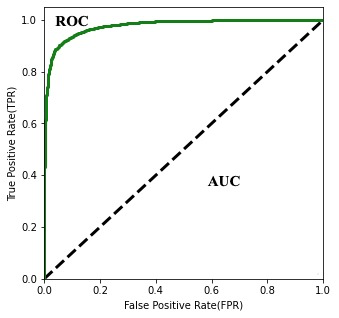


Figure 14: ROC curve (Abbasi, 2021)