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

AbdulRahman Opeyemi AbdulKareem

Department of Electrical and Computer Engineering

Kwara State University Malete, Kwara,Nigeria

[Remaxi135@gmail.com](mailto:Remaxi135@gmail.com)

***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.**

***Keywords-******Predictive Maintenance (PdM), Transformer Failures, Machine Learning, XGBoost, Random Forest, Operational Data, Historical Data, Predictive Models, Data Quality, Maintenance Strategies, Electrical Grids, Failure Prediction, Utility Companies, Operational Reliability.***

**INTRODUCTION**

For decades, the heart of any power grid, transformers, hum with the vital energy that fuels our world, as they underpin the functioning of homes, industries, and critical infrastructure. Among the various components of power systems, electrical transformers play a crucial role in the transmission and distribution of electricity. These devices are responsible for stepping down high-voltage electricity to safer levels for consumer use. 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).

In recent years, the emergence of predictive maintenance (PdM) has transformed the landscape of equipment management. PdM leverages advanced data analytics, machine learning, and real-time monitoring to anticipate equipment failures before they occur(Carvalho et al., 2019). By analyzing historical and operational data, PdM enables timely interventions, optimizing maintenance schedules and enhancing the overall resilience of power systems. This proactive approach not only minimizes the risk of unexpected failures but also extends the lifespan of critical infrastructure.

The application of machine learning algorithms in predictive maintenance has gained significant attention in the field of electrical engineering. 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.

Among these algorithms, Extreme Gradient Boosting (XGBoost) and Random Forest have emerged as powerful tools for failure prediction. XGBoost is known for its high performance and efficiency in handling large datasets, while Random Forest offers robustness and ease of interpretation. Both algorithms have shown promise in various domains, but their comparative effectiveness in predicting transformer failures remains underexplored.

This study aims to fill this gap by conducting a comprehensive analysis of XGBoost and Random Forest algorithms in the context of transformer maintenance. By utilizing operational and historical data, the research evaluates the predictive capabilities of both algorithms, focusing on key performance metrics such as accuracy, precision, and recall. The findings of this study are expected to provide valuable insights into the strengths and weaknesses of each algorithm, contributing to the development of more effective predictive maintenance strategies for transformers.

By adopting advanced analytical techniques, the study aspires to safeguard critical infrastructure and optimize maintenance practices in power systems, ultimately contributing to a more resilient and efficient energy landscape. The findings of this research will serve as a foundation for future studies aimed at improving predictive models and maintenance strategies, ensuring the continued reliability of electrical transformers in an increasingly complex power grid.

**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**.

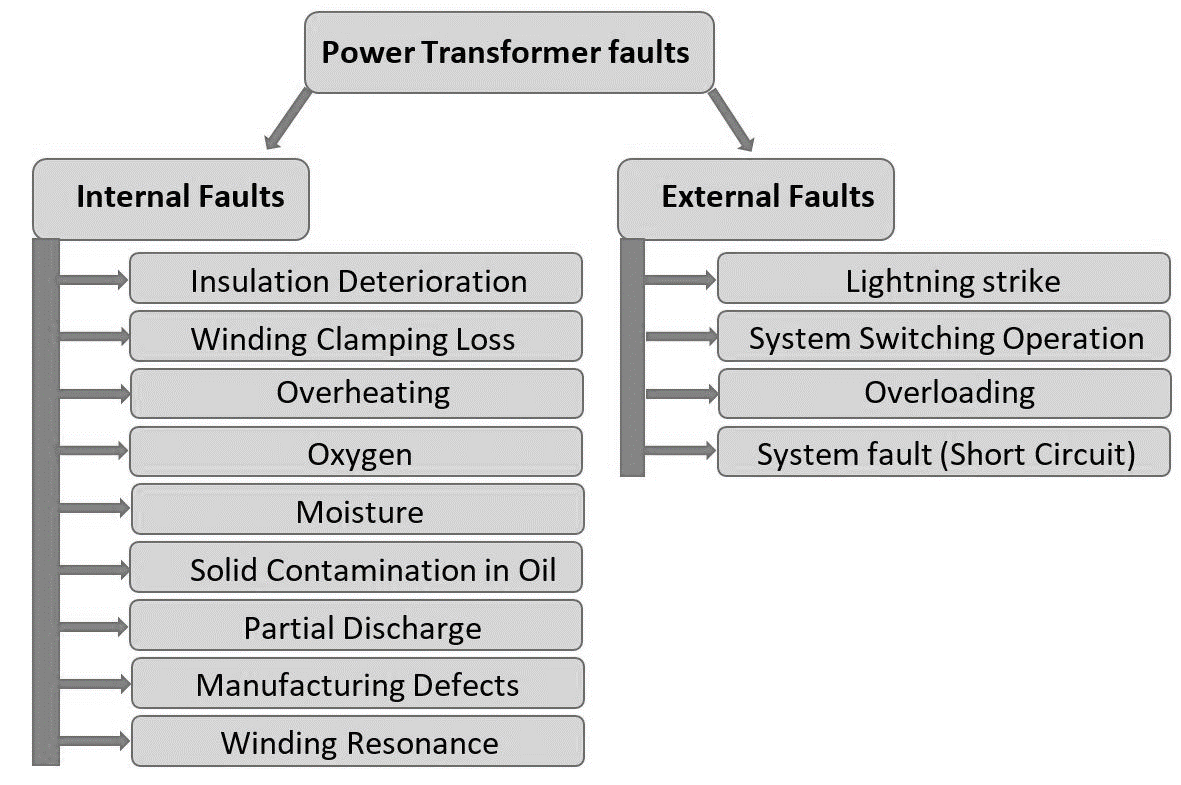


Figure 1: Electrical Faults of Transformer(Hussain et al., 2021)

**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 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 2: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 3: 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 4: 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:

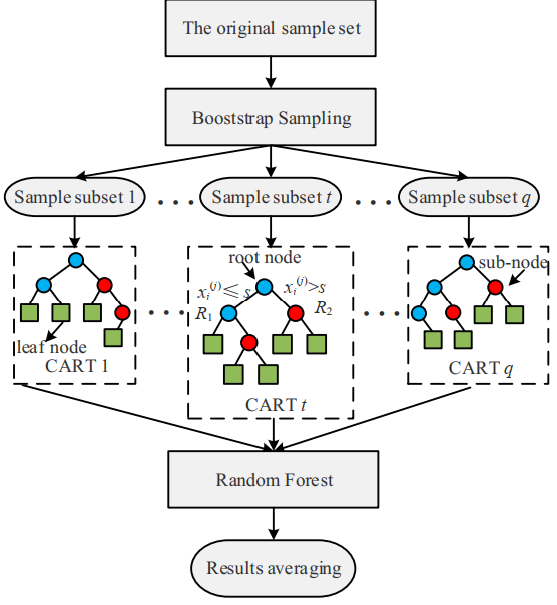


Figure 5: Random Forest Algorithm(Wang et al., 2023)

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:

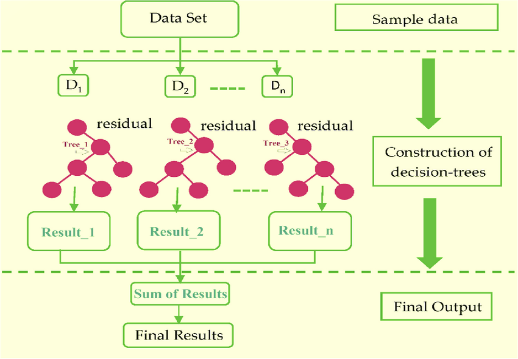


Figure 6: Extreme Gradient Boosting Algorithm(Khan et al., 2022)

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.

**RELATED WORKS**

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.

**METHODOLOGY**

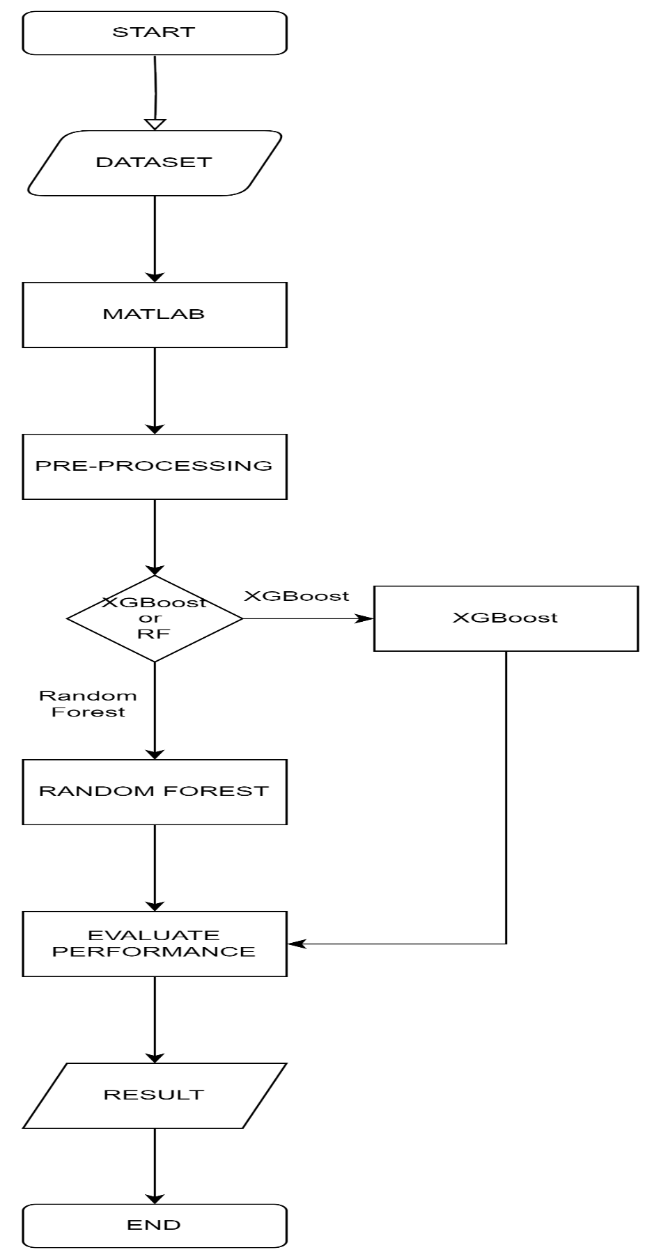


Figure 7: Study Framework

**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 19,352 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.