KWARA STATE UNIVERSITY, MALETE,
FACULTY OF ENGINEERING AND TECHNOLOGY,
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING.

TOPIC:

COMPARATIVE ANALYSIS OF XGBOOST AND RANDOM FOREST ALGORITHMS FOR TRANSFORMER FAILURE PREDICTION USING GRID STABILITY DATA

PRESENTED BY: **ABDULRAHMAN OPEYEMI ABDULKAREEM**
19/67EC/00903

SUPERVISED BY: **DR BILKISU JIMADA-OJUOLAPE**

MARCH 2024

INTRODUCTION

- Transformer failures can cause power outages, equipment damage, and safety hazards. Therefore, it is important to develop reliable and accurate methods to predict and prevent them (Tianjin da xue et al., 2018).
- Preventive or reactive repairs prove inefficient, causing unnecessary downtime and compromised power supply.(Tianjin da xue et al., 2018).
- This study aim 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.

SIGNIFICANCE OF STUDY

By comparing powerful algorithms like Random Forest and XGBoost in terms of accuracy, efficiency, and interpretability, will identify the most suitable algorithm for the task.

Successful implementation could significantly:

✓ Reduce downtime

✓ Lower maintenance costs and

✓Enhance grid resilience.

LITERATURE REVIEW

S/N	AUTHOR/YEAR	TITTLE	METHODOLOGY	STRENGHT	LIMITATION
1	Carratu et al. (2023)	A Novel Methodology for Unsupervised Anomaly Detection in Industrial Electrical Systems	Utilized unsupervised machine learning framework that leverages electrical current values and power grid parameters, employing the short-time Fourier transform for temporal analysis.	Exceptional performance with zero false positives and high overall accuracy	The technique's reliance on specific features may limit its applicability to varied anomaly types.
2	(Wang et al., 2023)	Transformer Fault Diagnosis Method Based on Incomplete Data and TPE-XGBoost	This approach utilizes Bayesian optimization to fine-tune XGBoost hyperparameters.	TPE-XGBoost effectively diagnoses transformer faults using incomplete data.	Diagnostic accuracy reduces when missing data exceeds 20%, requiring further improvement for high missing data rates (>30%)

LITERATURE REVIEW

S/N	AUTHOR/YEAR	TITTLE	METHODOLOGY	STRENGHT	LIMITATION
1	Rojek et al. (2023)	An Artificial Intelligence Approach for Improving Maintenance to Supervise Machine Failures and Support Their Repair	The study utilizes artificial neural networks (ANNs), a supervised machine learning technique, to predict machine failures and support maintenance processes within Industry 4.0 settings.	addresses the challenge of unbalanced data in real-world industrial applications	The study solely focuses on ANNs, limiting the exploration of other potentially suitable machine learning algorithms
2	Breviglieri et al. (2021b)	Predicting Smart Grid Stability with Optimized Deep Models	The study reviews the use of deep learning models for predicting stability in smart grids, focusing on the Decentral Smart Grid Control (DSGC) system.	Highlights challenges of using renewable energy sources and importance of stability analysis in networked control systems.	The need for testing with larger and more diverse grids.

LITERATURE REVIEW

S/N	AUTHOR/YEAR	TITTLE	METHODOLOGY	STRENGHT	LIMITATION
1	Chen et al. (2019)	XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System	This study proposes a method using the XGBoost algorithm to predict transient stability in power systems.	Leverages XGBoost's ability to handle missing values and avoid data normalization, simplifying the process.	Calls for further empirical validation and real-world application to confirm the effectiveness of the proposed method.
2	Marcelino et al. (2021)	Machine learning approach for pavement performance prediction.	Proposed a structured approach that involves gathering data from the LTPP database, employing imputation techniques for preprocessing, and developing models for 5 and 10-year predictions	Stands out for its forward-thinking approach, offering predictive models for both 5 and 10-year forecasts	The study faces challenges such as a historical reliance on ANNs and issues related to data availability and quality.

METHODOLOGY

In the study:

- It employs a structured methodology using the CRISP-DM model (illustrated in the flowchart) (IBM Corporation, 2021) .
- Leverages a public dataset from Kaggle to analyze electrical grid stability (kaggle, 2024) .
- Data preprocessing is done to ensure the data's readiness for modeling.
- XGBoost and Random Forest algorithms are employed.
- The performance of the models is evaluated using metrics to ensure the robustness of our predictive models.

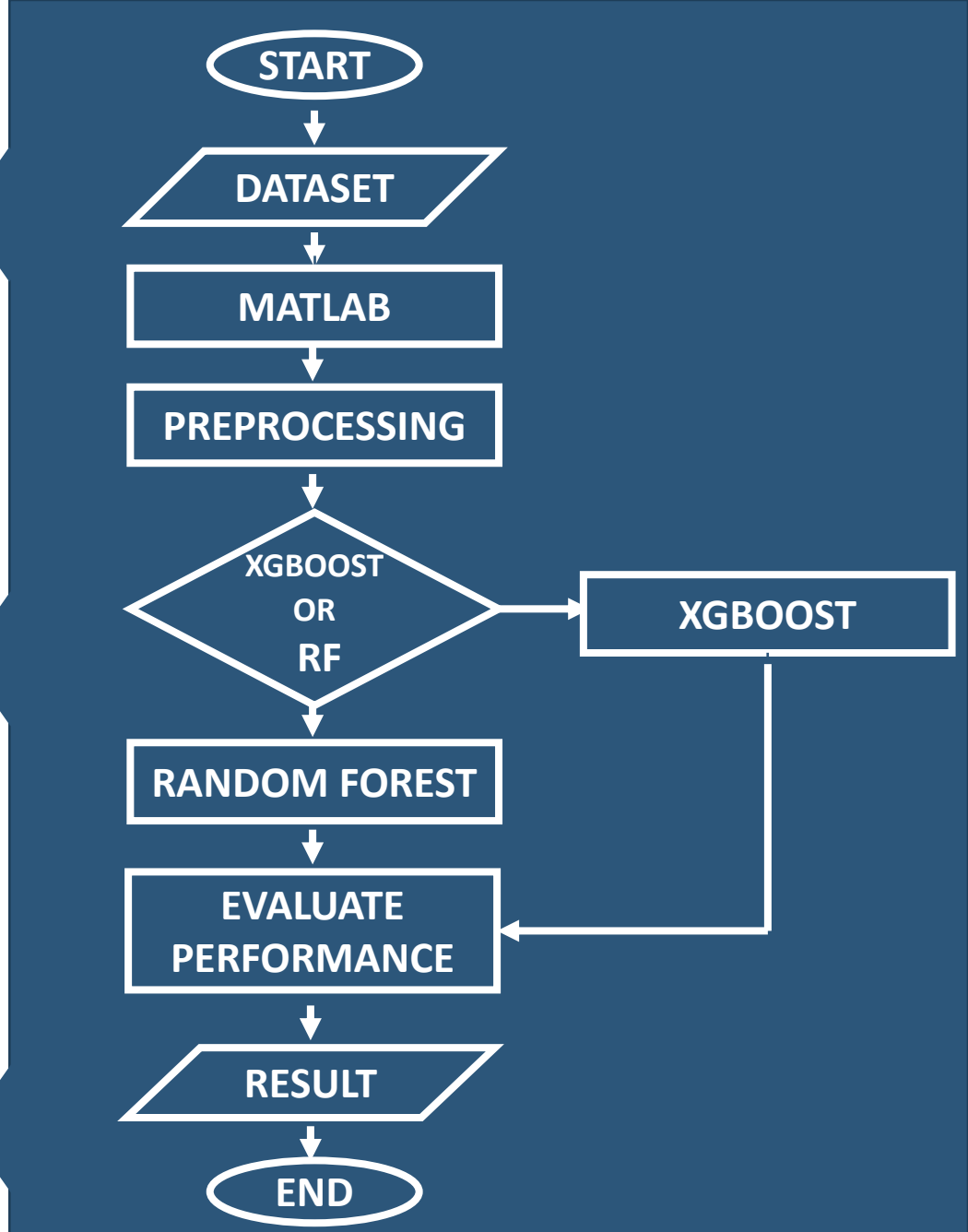

Fig1: STUDY FRAMEWORK

METHODOLOGY

Data acquisition

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 11 columns, with each row representing a unique observation and each column denoting a specific parameter or attribute.

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

Table1: DATASET PARAMETERS

METHODOLOGY

Preprocessing

Using MATLAB, data quality were enhanced before applying the machine learning algorithms. Steps include:

- ❑Removing duplicates, outliers, and missing values.
- ❑ Creating new features based on domain knowledge, such as power, resistance.
- ❑Selecting the most relevant features using correlation analysis and mutual information.

METHODOLOGY: ALGORITHM

RANDOM FOREST ALGORITHM

Random Forest uses bagging, which builds the trees independently and randomly and combines them using majority voting or averaging. It shines in handling imbalanced datasets, where the occurrence of transformer failures might be significantly lower compared to healthy operations.

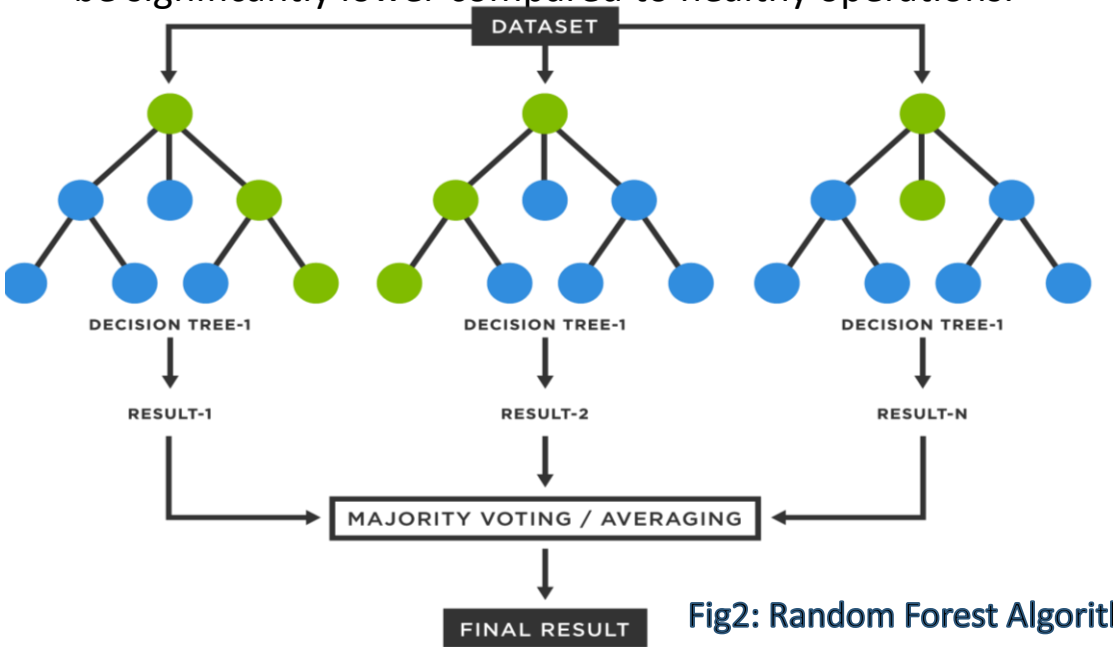

Fig2: Random Forest Algorithm

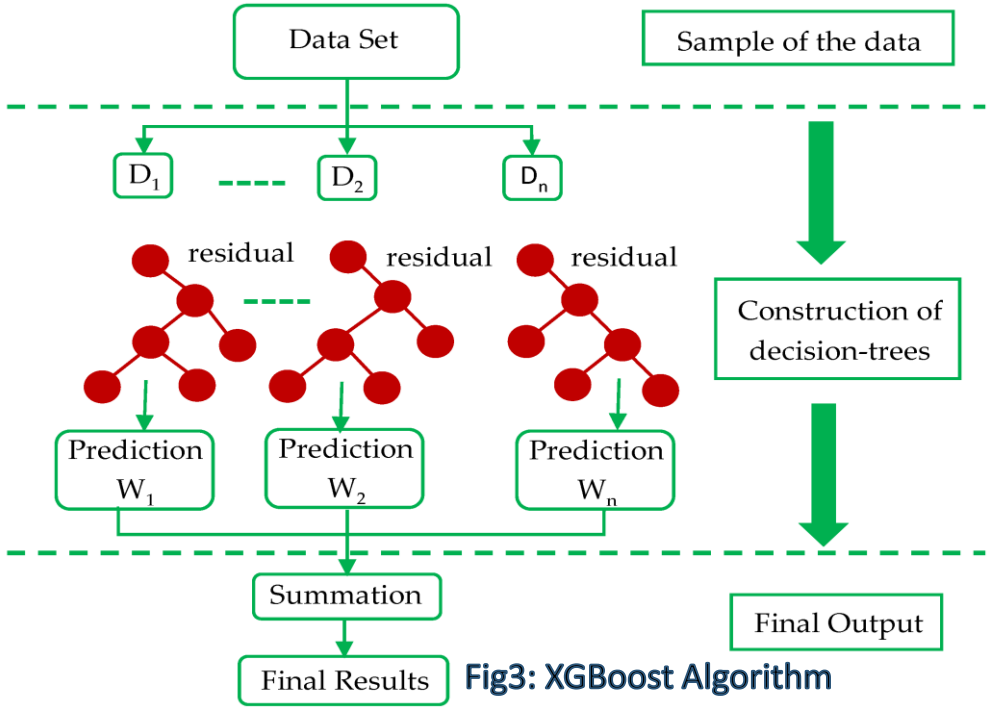

Fig3: XGBoost Algorithm

XGBoost combines multiple, weaker decision trees in a sequential way, focusing on correcting errors made by the previous ones. It excels at tackling intricate problems and preventing overfitting, ensuring generalizability beyond the training data.

XGBOOST ALGORITHM

METHODOLOGY: MODEL TRAINING

Steps for training XGBoost and Random Forest models in MATLAB include:

- ❑Load preprocessed data.
- ❑Split data into training and testing sets by utilizing the **traintestsplit** function
- ❑Train Xgboost leveraging the **Xgboost for MATLAB** toolbox with key parameters.
- ❑Train Random Forest leveraging the built-in **fitctree** function with specific parameters.
- ❑Utilize trained models to predict transformer failures on the testing set.
- ❑Hyperparameter tuning for further optimization of models.

PERFORMANCE EVALUATION

To compare the performance of XGBoost and Random Forest algorithms, six metrics were used:

- ❖Confusion matrix ,
- ❖Accuracy,
- ❖ Precision,
- ❖Recall,
- ❖ F1-score, and
- ❖ Receiver Operating Characteristics (ROC) curve

These metrics measure different aspects of the prediction quality, such as the overall correctness, the balance between false positives and false negatives, and the trade-off between sensitivity and specificity

PERFORMANCE EVALUATION

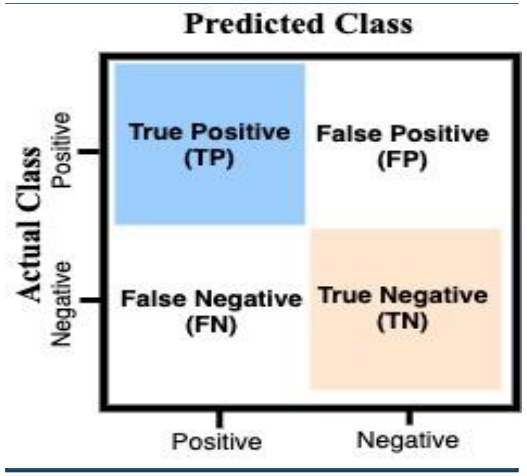

Fig4: CONFUSION MATRIX

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 (Abbasi, 2021).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Fig5: ACCURACY

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

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Fig6: F1-SCORE

The 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’

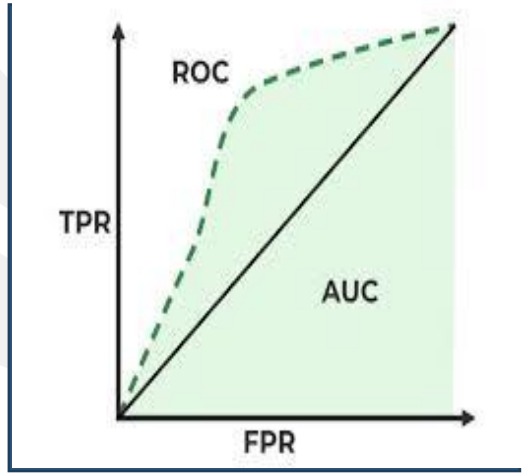

Fig7: ROC CURVE

For checking or visualizing the performance of the multi - class classification problem, AUC (Area Under the Curve) or ROC (Receiver Operating Characteristics) curve is use (Abbasi, 2021)

CONCLUSION

In summary, a quantitative and supervised approach is use to compare the performance of XGBoost and Random Forest algorithms for transformer failure prediction using grid stability data. Various data preprocessing, algorithm tuning, and performance comparison steps were performed to answer research questions and test hypotheses, which are:

- How do XGBoost and Random Forest algorithms differ in terms of accuracy, precision, recall, F1-score, ROC curve, and confusion matrix for transformer failure prediction?
- What are the advantages and disadvantages of each algorithm for this task?
- Which of the algorithm is better for predictive maintenance of a transformer?

REFERENCES

- 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
- 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
- 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.
- 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
- IBM Corporation. (2021). *CRISP-DM Help Overview*. https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview
- Sreshta, P. (2020). *Distributed Transformer Monitoring*. Distributed Transformer Monitoring. https://www.kaggle.com/datasets/sreshta140/ai-transformer-monitoring
- kaggle. (2024). *kaggle webpage* . https://www.kaggle.com/
- 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
- 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
- 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
- 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

THANK YOU FOR LISTENING!