
FAULT CLASSIFICATION IN TRANSMISSION LINE USING ANN, EXTREME GRADIENT BOOSTING AND RANDOM FOREST MACHINE LEARNING MODELS.

Joshua Alberah^{1,*}, G.a. Olarinoye¹, A.s. Abubakar¹, A.b Kunya¹, S.h Sulaiman¹

¹Ahmadu Bello University, Zaria

Email: alberahjosh@yahoo.com*, oyeguzeh@yahoo.com, adams4real56@gmail.com, abkunya@kiu.ac.ug,
4sulaiman2hs@gmail.com

***Corresponding author:** Joshua Alberah

DOI: <https://doi.org/10.67203/sp.ICESID.2026>

ABSTRACT

The rapid advancement of power systems and smart grids necessitates the development of robust fault classification techniques to ensure reliable and resilient electrical power delivery. This paper presents a machine learning-based methodologies for fault classification in power transmission lines. The study explores extreme gradient boosting (XGBoost), random forest (RF) and Artificial Neural Network (ANN) machine learning models for the fault classification. Aimed to classify no-fault, single line-ground fault, double line fault, double line-ground fault, three-line fault and three line-ground faults. The machine learning models were evaluated based on key performance metrics such as accuracy, precision, recall, and F1-score. The scheme was validated on Kaggle pre-fault and post-fault line voltage and current signals. The XGBoost perform with an accuracy of 84.47 %, precision of 84.40 %, recall of 82.30 % and F1-score of 81.79 %, the RF with an accuracy of 83.93 %, precision of 84.68 %, recall of 83.93 % and F1-score of 84.13 % and the ANN with an accuracy of 86.35%, precision of 90.73 %, recall of 83.81 % and F1-score of 82.54 % respectively. This study represents a significant contribution to advancing machine learning techniques in improving the efficiency and efficacy of fault management in high-voltage transmission lines.

Keywords: ANN; Classification; Extreme Gradient Boosting; Machine Learning; Random Forest

1. INTRODUCTION

Transmission lines form a critical part of power systems, ensuring the reliable and efficient transfer of electrical energy across vast distances. However, faults in transmission lines, such as short circuits, line-to-ground faults, or equipment failures, can lead to severe power outages, equipment damage, and significant financial losses. The ability to classify faults quickly and accurately is therefore essential to maintain the stability and safety of power systems. Traditional fault detection techniques often rely on predetermined thresholds and expert knowledge, but these methods can be inadequate for addressing the increasing complexity and dynamic nature of modern power systems (Vikram et al., 2023). Machine learning (ML) techniques have shown great promise in enhancing fault classification in transmission lines. By leveraging vast amounts of real-time sensor data and historical fault records, ML models can automatically learn the underlying patterns and characteristics of various fault types without the need for explicit programming (Sappati & Bhadra, 2018). One recent example is the work of (Anwar et al., 2025), who investigated fault classification in transmission lines using several machine-learning techniques, including Random Forest (RF), K-Nearest Neighbours (KNN), and Long Short-Term Memory (LSTM) networks. Their principal contribution was an RF–LSTM Tuned KNN ensemble, which combines different learning approaches to improve fault classification performance. Another relevant contribution is the study by (Ghaly, 2025), which proposed a Multi-Target Ensemble Classifier (MTEC) for transmission-line fault classification under limited-data conditions. Ghaly's study first evaluated conventional ML algorithms using a much larger dataset and subsequently reduced the training set with the remaining samples used for testing. Despite the promising results reported in previous studies, emphasis on overall accuracy as the principal performance indicator but could not uncover the performance of the methods in recognition of individual fault classes. This remains as a gap which has been addressed in this research work.

2. METHODOLOGY

This section describes the data and approach used to achieve the classification.

2.1 Kaggle Dataset

The dataset from Kaggle repository contains line currents and voltages from a system with four 11 kV generators and transformers at the transmission line's midpoint that were recorded under various fault scenarios. The dataset consists of a multi-class dataset with 7861 rows and 10 columns for identifying particular fault types. The dataset comprises electrical measurements representing normal and faulted transmission-line operating conditions. Eight electrical features were selected for classification, namely phase-A current (I_a), phase-B current (I_b), phase-C current (I_c), zero-sequence current (I_0), phase-A voltage (V_a), phase-B voltage (V_b), phase-C voltage (V_c), and voltage variation (ΔV). The target variable comprises six classes: no fault (N), line-to-ground (LG), line-to-line (LL), double-line-to-ground (LLG), three-phase (LLL), and three-phase-to-ground (LLLG). The data were subsequently divided into 80% training and 20% testing subsets. (Anwar et al., 2025).

2.2 Machine Learning Models for Fault Classification

This subsection offers the machine learning models used for the fault classification. These are Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Artificial Neural Network (ANN).

2.2.1 Random Forest (RF)

For classification and regression problems, random forest is an ensemble learning technique that works well with complicated datasets that have high dimensionality and non-linear features. It was utilized for classification in this paper. To increase generalization performance and decrease overfitting, it builds many decision trees, each trained on a fraction of training data and a random feature selection. Using the average for regression and majority voting for classification, the random forest combines the predictions of each individual tree to get a final prediction. Additionally, it is resistant to noise and outliers and offers insightful information about the significance of each feature for deciphering the underlying patterns. For the classification issue, the random forest uses the Gini impurity $I(t)$ in equation (1) (Song et al., 2023).

$$I(t) = 1 - \sum_{i=1}^c p(i|t)^2 \quad (1)$$

where $p(i|t)$ is the probability of training cases of a class i at node t , c is the number of classes. A node's total impurity is the weighted sum of its impurities across all classes. The hyperparameters used are: number of decision tree 100, minimum leaf size 5, number of predictors sampled at each split 4, training/test split 80%/20%, temporal window length 20, noise level 0.05, number of raw input features 8, engineered features per signal 5 and total engineered predictors 40.

2.2.2 Extreme Gradient Boosting (XGBoost)

A boosting approach called Extreme Gradient Boosting (XGBoost) expands the model by adding several Classification and Regression Trees (CART). Constant prediction to reduce the error between the predicted and actual values is the goal of XGBoost training. Fitting the residuals from the prior forecast is the main goal of this approach when a new decision tree is introduced. The ultimate result of the predictive model is then generated by combining or aggregating all of the decision trees' anticipated outputs (Zhang, 2023).

Given the assumption that the label associated with transmission line fault data set x is y , the XGBoost model's fault prediction output can be written as follows:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (2)$$

Where F represents all possible CART trees, and $f_k(x_i)$ represents the predicted value of the i_{th} sample on the k_{th} tree. This model consists of K CART trees.

A decision tree is a simple algorithm that divides data into subsets according to feature values. Each node in the tree structure reflects a feature-based decision. The choice was made by:

$$Decision = \underset{c}{\operatorname{argmax}} \left(\sum_{i=1}^N I(y_i = c) \right) \quad (3)$$

where c is the class label and I is an indicator function. Decision trees require less pre-processing of the input and are simple to understand. However, they are sensitive to slight changes in the data and are prone to overfitting, especially with deep trees (Meng et al., 2023). The hyperparameters used are: number of boosting cycles/trees 100, maximum tree splits 15, minimum leaf size 7, learning rate 0.05, train/test split 80%/20%, temporal window 20 samples, gaussian noise level 0.05, raw input features 8 and engineered features 40.

2.2.3 Artificial Neural Network (ANN)

An artificial neural network (ANN) is a massively parallel computing system with many connections that mimics the function of biological neurons. They are composed of multiple processing components connected in a specific manner to provide a specific result. They become more effective in practical applications as a result. Furthermore, it outperformed other artificial intelligence in fault identification in transmission lines due to its capacity to handle incomplete and corrupt data, solve naturally nonlinear problems without the need for prior knowledge functions relating the problem variables, and process information in a parallel, distributed manner (Peter et al., 2018). The ANN architecture used in this work is the Convolutional Neural Network–Bidirectional Long Short-Term Memory (CNN–BiLSTM) with the following hyperparameters: input 8, sequence length 20 samples, 1-D Convolution filter 32, kernel/filter size 3, BiLSTM units 64, dropout 0.30, fully connected layer 64, output fully connected layer 6, classification classes 6, maximum epochs 50, mini-batch size 32, initial learning rate 0.001, training/testing split 80%/20%. An artificial neural network with a feedforward multilayer perceptron structure is shown in Figure 1.

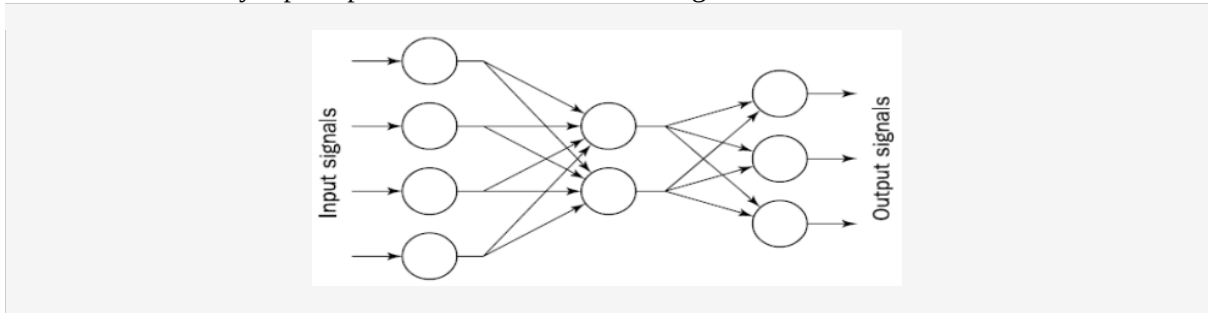


Figure 1: Feedforward Multilayer Perceptron ANN's Basic Structure.

2.3 Assessment Criteria

The models are evaluated using the well-used statistical concepts, including accuracy, precision, recall, and F1-score.

2.3.1 Accuracy

The evaluation criteria's mathematical expression, which includes accuracy is the ratio of the correctly predicted instances to the total number of instances, is provided by equation (4)

$$A_{cc} = \frac{(TP + TN)}{TP + TN + FP + FN} \times 100\% \quad (4)$$

The symbols TP, TN, FP, and FN denote TP = faulty correctly predicted as faulty; TN = not faulty correctly predicted as not faulty; FP = not faulty incorrectly predicted as faulty; FN = faulty incorrectly predicted as not faulty.

2.3.2 Precision

The Precision(P), as determined by Equation (5), is the ratio of the correctly predicted positive instances to the total number of predicted positives (Goni et al., 2023).

$$P = \frac{TP}{(TP + FP)} \quad (5)$$

2.3.3 Recall

Recall, also known as Sensitivity or True Positive Rate, is the ratio of correctly predicted positive instances to actual positives (Goni et al., 2023). The formula for recall is as follows:

$$R = \frac{TP}{(TP + FN)} \quad (6)$$

2.3.4 F1 – Score

The formula for calculating the F1-score is the harmonic mean of precision and recall (Goni et al., 2023).

$$F1 - score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (7)$$

3. RESULTS AND DISCUSSION

Table 1 shows the evaluation of XGBoost, RF and ANN. The XGBoost perform with an accuracy of 84.47 %, precision of 84.40 %, recall of 82.30 % and F1-score of 81.79 %, the RF with an accuracy of 83.93 %, precision of 84.68 %, recall of 83.93 % and F1-score of 84.13 % and the ANN with an accuracy of 86.35%, precision of 90.73 %, recall of 83.81 % and F1-score of 82.54 %.

Table 1: Evaluation of Classification Models on Kaggle Data

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
XGBoost	84.47	84.40	82.30	81.79
RF	83.93	84.68	83.93	84.13
ANN	86.35	90.73	83.81	82.54

Table 2 shows a comparison of the present study with that of (Ghaly, 2025). The work of Ghaly 2025 presented an overall model performance accuracy of 82.91%, precision of 88.32%, recall of 84.44% and F1-Score of 86.34%. This is depicted in Table 2.

Table 2: Comparison of XGBoost, RF and ANN With The Work of Ghaly 2025.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
XGBoost	84.47	84.40	82.30	81.79
RF	83.93	84.68	83.93	84.13
ANN	86.35	90.73	83.81	82.54
Ghaly 2025	82.91	88.32	84.44	86.34

Figure 2 shows the true class and predicted class of the no-fault, single line-ground fault, double line fault, double line-ground fault, triple-line fault and triple line-ground faults mapped as 0,1,2,3,4 and 5 respectively. The classified faults are the values on the diagonal on the confusion matrix. Whereas the remaining entries are where the model made wrong classification. Figure 2 represent XGBoost confusion matrix which classified correctly the no-fault, single line-ground fault, double line fault, double line-ground fault, triple-line fault and triple line-ground faults with 96.62%, 77.07%, 92.0%, 100%, 83.64% and 44.49% respectively.

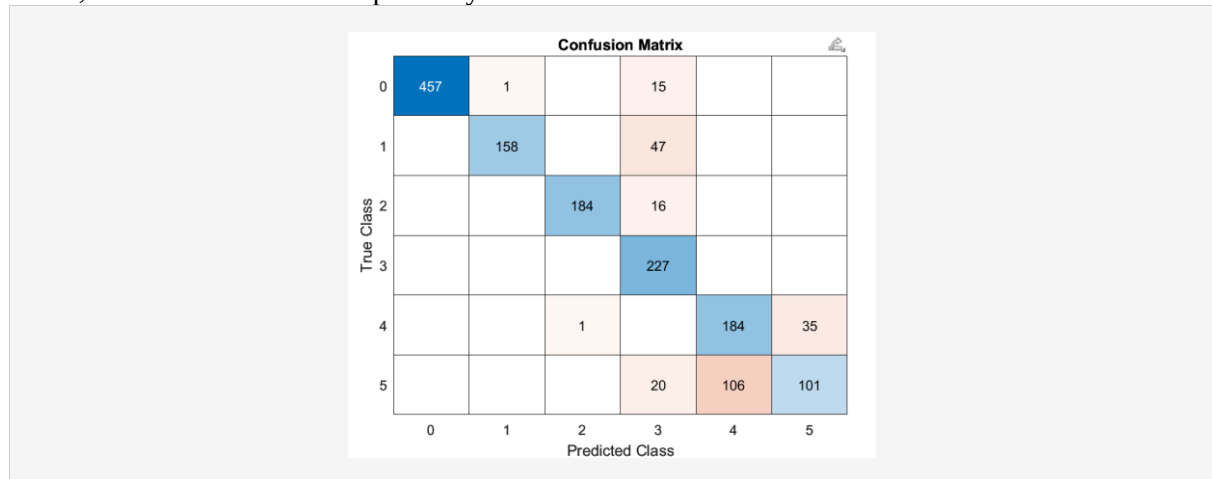


Figure 2: XGBoost confusion matrix

Figure 3 represent RF confusion matrix which detected and classified correctly the no-fault, single line-ground fault, double line fault, double line-ground fault, triple-line fault and triple line-ground faults with 96.62, 86.35%, 93.0%, 82.82%, 80.45% and 64.32% respectively.

Confusion Matrix						
True Class	0	1	2	3	4	5
	457		5	10		1
		177		22		6
			186	14		
				188		39
			1		177	42
Predicted Class	0	1	2	3	4	5

Figure 3: RF confusion matrix

Figure 4 shows the ANN confusion matrix which classified correctly the no-fault, single line-ground fault, double line fault, double line-ground fault, triple-line fault and triple line-ground faults with 99.79%, 100%, 100%, 95.47%, 67.51% and 27.16% respectively.

Confusion Matrix						
True Class	0	1	2	3	4	5
	465	1				
		211				
			208			
		10	1	232		
					133	64
Predicted Class	0	1	2	3	4	5

Figure 4: ANN Confusion Matrix

4. CONCLUSION

Supervised machine learning techniques were employed to classify no-fault, single line-ground fault, double line fault, double line-ground fault, triple-line fault and triple line-ground faults. The XGBoost perform with an accuracy of 84.47 %, precision of 84.40 %, recall of 82.30 % and F1-score of 81.79 %, the RF with an accuracy of 83.93 %, precision of 84.68 %, recall of 83.93 % and F1-score of 84.13 % and the ANN with an accuracy of 86.35%, precision of 90.73 %, recall of 83.81 % and F1-score of 82.54 % respectively. Slightly lower values were observed for triple-line faults and triple line-ground faults on the confusion matrix respectively this is as a result of identical overlapping fault signatures. The use of advanced ML models that can optimize separating margins in overlapping space is recommended. The proposed machine-learning-based approach therefore provides a pathway toward more intelligent monitoring and protection of transmission networks, supporting the development of a more resilient and sustainable power system.

5. REFERENCES

Anwar, T., Mu, C., Yousaf, M. Z., Khan, W., Khalid, S., Hourani, A. O., & Zaitsev, I. (2025). *Robust fault detection and classification in power transmission lines via ensemble machine learning models*. 1–18.

- Ghaly, A. El. (2025). Hybrid ML Algorithm for Fault Classification in Transmission Lines Using Multi-Target Ensemble Classifier with Limited Data.
- Goni, M. O. F., Nahiduzzaman, M., Anower, M. S., Rahman, M. M., Islam, M. R., Ahsan, M., Haider, J., & Shahjalal, M. (2023). Fast and Accurate Fault Detection and Classification in Transmission Lines using Extreme Learning Machine. *E-Prime - Advances in Electrical Engineering, Electronics and Energy*, 3(January), 100107. <https://doi.org/10.1016/j.prime.2023.100107>
- Meng, L., Bai, B., Zhang, W., Liu, L., & Zhang, C. (2023). Research on a Decision Tree Classification Algorithm Based on Granular Matrices. *Electronics (Switzerland)*, 12(21), 1–14. <https://doi.org/10.3390/electronics12214470>
- Peter, O., Oluwaseun, S., & Ayokunle, A. (2018). *Fault Identification System for Electric Power Transmission Lines Using Artificial Neural Networks*. 9(2), 678–685.
- Sappati, K. K., & Bhadra, S. (2018). Piezoelectric polymer and paper substrates: A review. *Sensors (Switzerland)*, 18(11). <https://doi.org/10.3390/s18113605>
- Song, Z., Wang, J., Wang, M., Yang, B., Cui, L., Zhang, Q., Xiang, G., & Lin, R. (2023). Hydraulic Systems Fault Diagnosis Based on Random Forests Recursive Feature Elimination and XGBoost. *Journal of Physics: Conference Series*, 2569(1). <https://doi.org/10.1088/1742-6596/2569/1/012056>
- Vikram, S., Singh, P., Prasad, T., Kamila, S., & Agnihotri, P. (2023). *Fault Detection and Classification using Wavelet and ANN in DFIG and TCSC Connected Transmission Line*. 1–23.
- Zhang, K. (2023). Transmission Line Fault Diagnosis Method Based on SDA-ISSA-XGBoost under Meteorological Factors. *Journal of Physics: Conference Series*, 2666(1). <https://doi.org/10.1088/1742-6596/2666/1/012006>