CLASSIFICATION OF SHORT CIRCUIT FAULTS IN MODERN POWER SYSTEM USING MACHINE LEARNING

A Project Report Submitted

by

ADITYA KUMAR

(2001EE85)

In Partial Fulfilment
of the Requirements for the award of the degree

BACHELOR OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY PATNA

4TH DECEMBER 2023

THESIS CERTIFICATE

This is to certify that the thesis titled **CLASSIFICATION OF SHORT CIRCUIT FAULTS IN MODERN POWER SYSTEM USING MACHINE LEARNING**, submitted by **Aditya Kumar,** to the Indian Institute of Technology, Patna, for the award of the degree of **Bachelor of Technology**, is a bonafide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. S.K Parida
Supervisor
Associate Professor
Dept. of Electrical Engineering
IIT-Patna, 800 013

Place: IIT Patna

Date: 4th December 2023

ACKNOWLEDGEMENTS

I want to express my gratitude to Dr.S.K.Parida, my BTP supervisor, for his invaluable guidance throughout the project. His support in conceptualizing the project and providing essential resources to tackle the problem statement was highly appreciated.

ABSTRACT

This report explores the critical role of transmission lines in power systems for efficient power transfer from the source to the load end. Addressing the common occurrence of faults in overhead transmission lines, the study emphasizes the need for swift and accurate fault classification to ensure the uninterrupted operation of the power system. Employing MATLAB Simulink, an 11-bus system is simulated, introducing faults in various regions. The resulting dataset, comprising faulty voltages and currents at different phases and locations, serves as input for a machine-learning model. The report employs the Random Forest classifier algorithm to categorize faults into four types: single line to ground, double line, double line to ground, and symmetric faults. The findings contribute to enhancing the reliability of power systems through effective fault identification and isolation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The electrical power system consists of three critical subsystems: generation, transmission, and distribution, where transmission and distribution lines are susceptible to faults induced by environmental exposure. These faults, categorized as abnormal conditions, can occur between phases or between phases and the ground, leading to voltage drops and excessive current flow and causing damage to power system equipment. Timely fault resolution is crucial due to maintenance challenges, inconvenience to users, and financial losses. This report addresses transmission line faults, classifies them into symmetrical and unsymmetrical categories, and introduces a novel approach to data generation from a comprehensive 2-area, 11-bus system, considering multiple non-fault locations. Leveraging this data set, a Random Forest classifier achieves a high accuracy of **90%** on the training data set and **86%** percent on the testing data set, showcasing its effectiveness in fault classification.

1.1 Types of Fault

A fault within the power system refers to a deviation from the intended pathway, leading to a disruption in the flow of current. The fault in the power system is mainly categorized into two types, which are:

1. Open circuit fault (Series Fault)
2. Short circuit fault (Shunt Fault)

1.1.1 Open circuit fault

The open circuit fault mainly occurs because of the failure of one or two conductors. The open circuit fault takes place in series with the line, and because of this, it is also called the series fault. The open circuit fault is categorized as:

- Open Conductor Fault
- Two conductors Open Fault
- Three conductors Open Fault

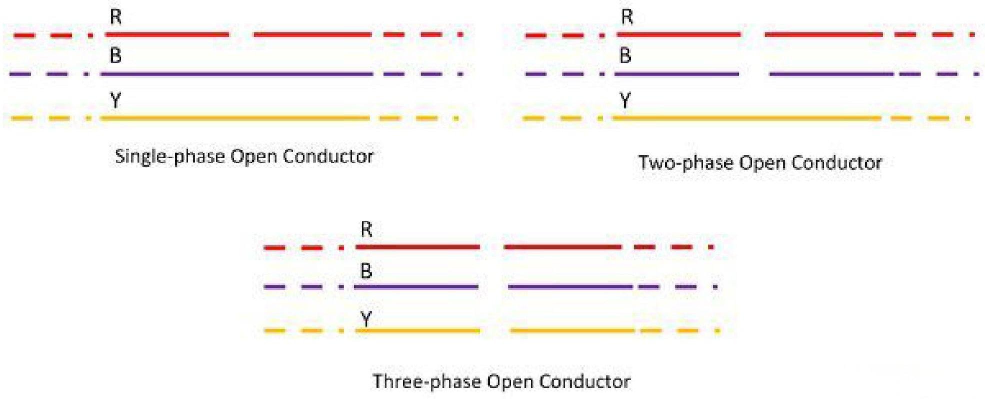


Figure 1.1: Open Circuit Faults

1.1.2 Short circuit fault

In this fault scenario, conductors from distinct phases make unintended contact, typically involving a power line, power transformer, or another circuit element. This contact results in a significant surge of current within one or two phases of the system. Short-circuit faults are categorized into symmetrical and asymmetrical faults based on their characteristics.

Symmetrical Faults: -

The faults which involve all three phases are known as the symmetrical fault. Such types of fault remain balanced even after the fault. The symmetrical faults mainly occur at the terminal of the generators.

1. **Line-Line-Line (L-L-L) Fault:** The L–L–L fault occurs rarely, but it is the most severe type of fault that involves the largest current. This large current is used for determining the rating of the circuit breaker.
2. **Line-Line-Line-Ground (L-L-L-G) Fault:** The L–L–L–G fault occurs between the three phases and the ground of the system. The probability of occurrence of such type of fault is nearly 2 to 3 percent.

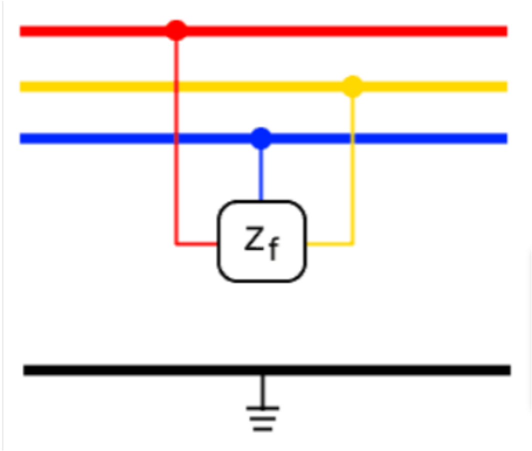


Figure 1.2: L-L-L Fault

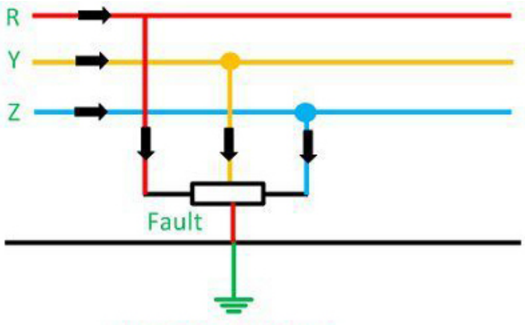


Figure 1.3: L-L-L-G Fault

Unsymmetrical Faults: -

The fault gives rise to unsymmetrical current, i.e., current differing in magnitude and phases in the three phases of the power system are known as the unsymmetrical fault. The unsymmetrical makes the system unbalanced.

1. **Single Line to Ground (S-L-G) Fault:** The single line of ground fault occurs when one conductor falls to the ground or contacts the neutral conductor. 70 – 80 percent of the fault in the power system is the single line-to-ground fault.

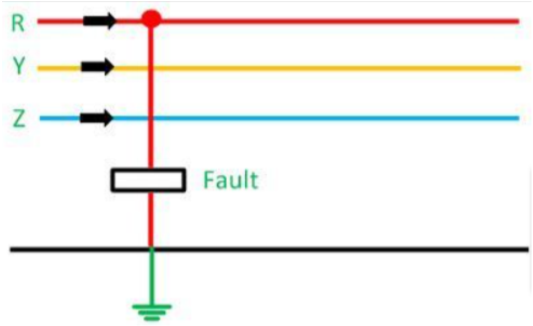


Figure 1.4: S-L-G Fault

2. **Line-Line (L-L) Fault:** A line-to-line fault occurs when two conductors are short circuited. The major cause of this type of fault is the heavy wind. The heavy wind swings the line conductors which may touch together and hence cause a short-circuit. The percentage of such types of faults is approximately 15 – 20 percent.

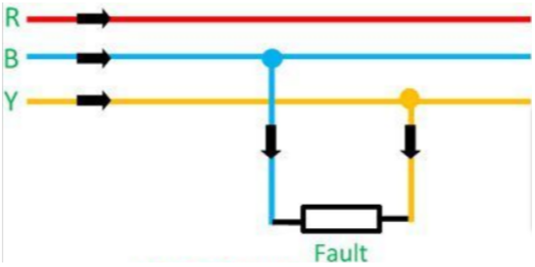


Figure 1.5: L-L Fault

3. **Line-Line-Ground (L-L-G) Fault:**In a double line-to-ground fault, the two lines come in contact with each other along with the ground. The probability of such types of faults is nearly 10 percent.

CHAPTER 2

LITERATURE REVIEW

1. In [1], the use of ANN for fault classification and detection of faults in power systems has been thoroughly investigated. The phase currents, phase voltages, and zero sequence components of a system model created with Simulink/MATLAB were used to apply the ANN algorithm. The data obtained from the different fault models was fed to the Artificial Neural Network for both training and testing purposes. The results were impressive. They can also reach better results by increasing the number of buses, creating more fault locations, and finding exact locations of fault.

2. In [2], a three-phase medium power transmission line system under study was converted into a Pi-model and simulation using the MATLAB/ Simulink® environment. Simulated values of transmission systems are used as training data for neural networks. Feed-forward BPNN algorithms were used for fault classification and detection. Performance analysis of three BPNN algorithms was done and results are presented.

Paper Name	**Publication**	**Methodology**	**Research Gap**
Fault Detection and Classification in Power System using ANN.	2nd International Conference on Intelligent Technologies (CONIT),2022	Uses 2 ANN models, one for detection and one for classification.	They can reach better results by increasing the number of buses, creating more fault locations, and finding exact locations of fault.
Fault Detection and Classification in Power Transmission Lines using BPNN.	2020 International Conference on Smart Electronics and Communication (ICOSEC)	Uses BPNN model with Levenberg-Marquardt (LM) algorithm for fault detection and classification.	The study is done on a small 3-phase transmission line for only A-G and A-B-G type faults.

Table 2.1: Summary of Literature Review

CHAPTER 3

SIMULATION WITH TWO AREA SYSTEM

In this project, we have used the famous 2 area kundur system for fault analysis at various regions. The system consists of 2 areas where each area consists of 4 buses, 2 generators, and 2 transformers. The two areas have been bridged through 3 additional buses, making it total of 11 buses in use.

The whole system is divided into 6 regions. Region 5-6, Region 6-7, Region 7-8, Region 8-9, Region 9-10, Region 10-11 of lengths 25 km, 10 km, 110 km, 110 km, 10 km, 25 km respectively.

After simulation, various faults have been observed through below waveforms:

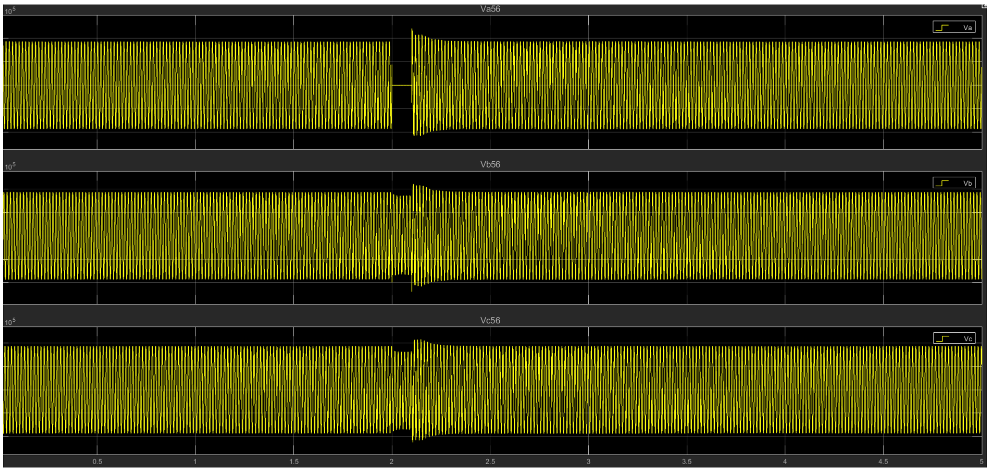


Figure 3.1: Voltage waveform after SLG Fault on line A.

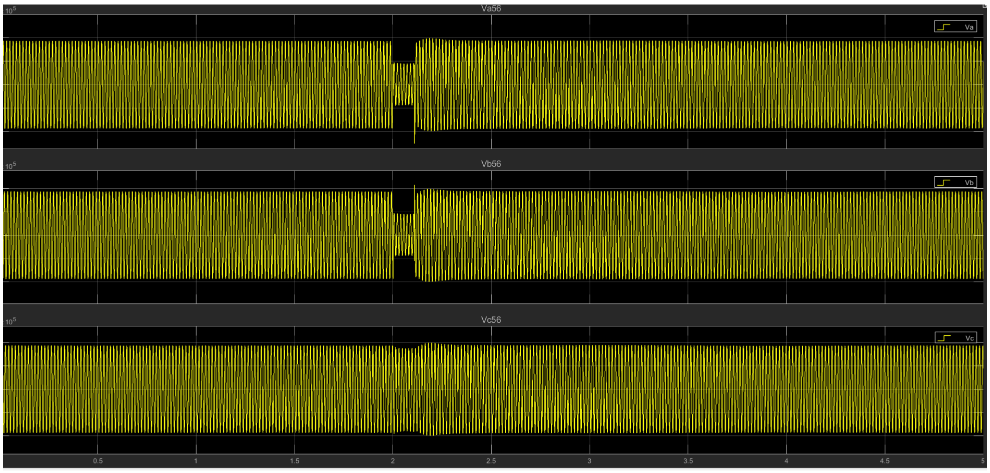


Figure 3.2: Voltage waveform after LL Fault on line A and B.

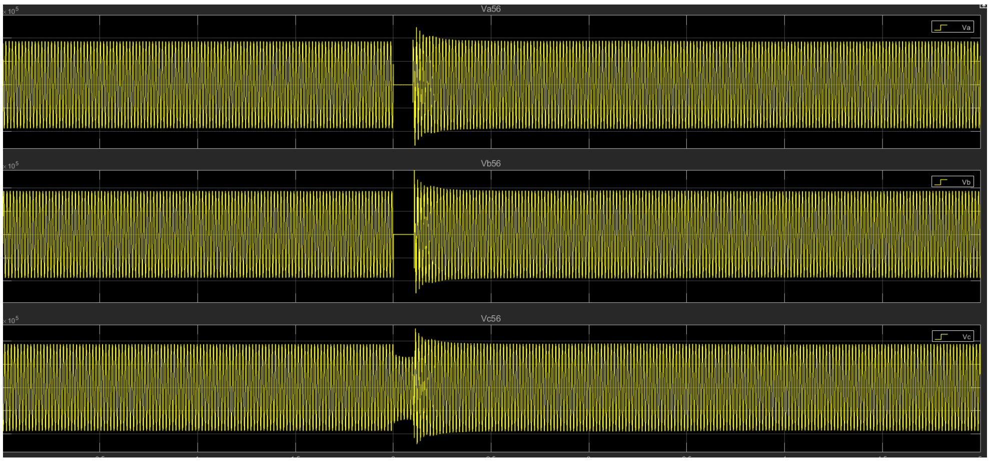


Figure 3.3: Voltage waveform after LLG Fault on line A and B.

CHAPTER 4

METHODOLOGY

Our approach will involve a methodical progression through a set of clearly outlined stages, as depicted in the accompanying flow chart [Fig 4.1]. The process begins with gathering data from diverse sources, such as sensors and monitoring devices. Subsequently, we undergo thorough data pre-processing to maintain data quality and uniformity. Next, we employ feature extraction methods to capture crucial insights from the data, and feature selection is used to streamline the data set. Leveraging this dataset, a Random Forest classifier is implemented, achieving robust accuracy on both training and testing datasets.

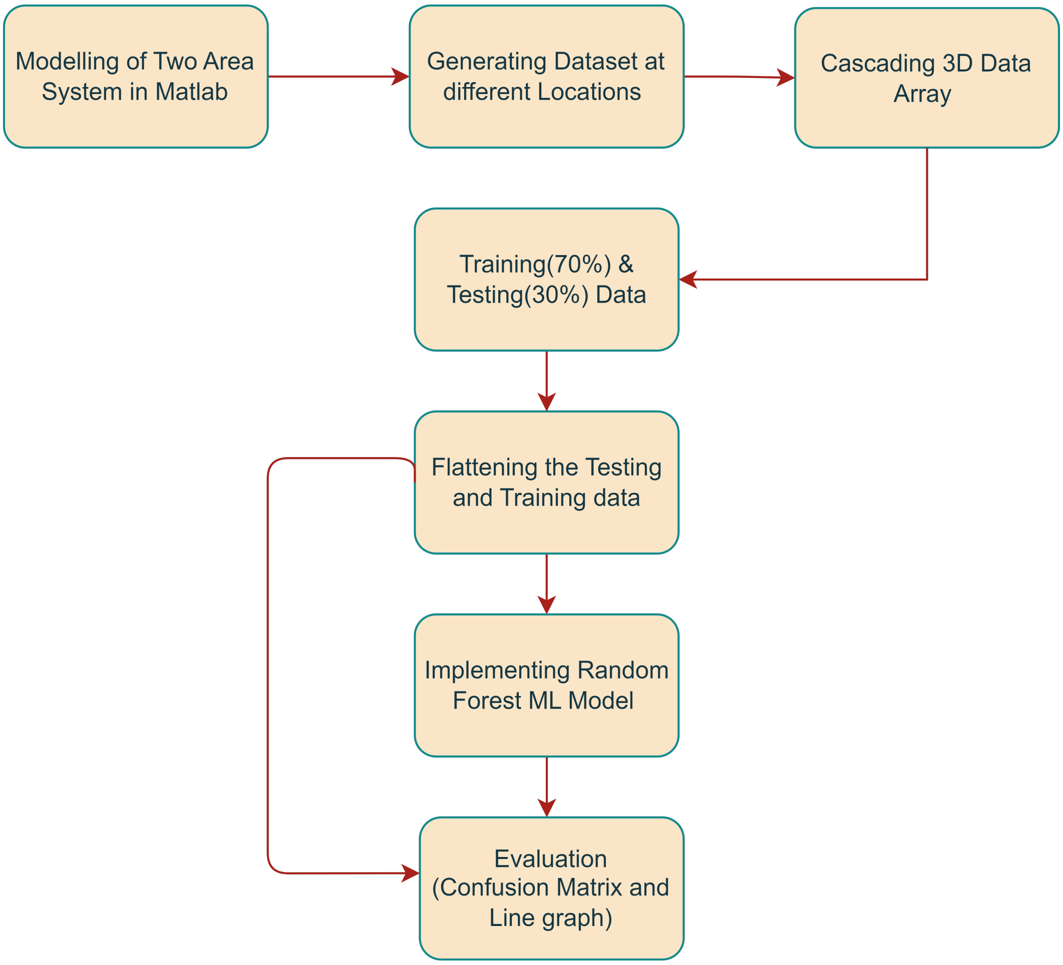

Modelling of Two Area System in Matlab
Generating Dataset at different Locations
Cascading 3D Data Array
Training(70%) & Testing(30%) Data
Flattening the Testing and Training data
Implementing Random Forest ML Model
Evaluation (Confusion Matrix and Line graph)

Figure 4.1: Flowchart of Methodology for Shunt Fault Classification Model.

4.1 Data Collection

In our research, we employed an IEEE 2 area system, consisting of 11 buses, as the foundation for our primary dataset. This system, illustrated in Figure 4.2, serves as a reliable model of an electrical power network, enabling the realistic simulation of various fault scenarios. Our data generation process was meticulous, covering all common electrical fault types encountered in power systems. Focusing on the dynamic behavior of the system during faults, we specifically collected voltage and current values from a strategically chosen subset of five buses. These buses, located in diverse regions, were selected to provide a comprehensive representation of the system's responses to faults in four distinct regions. This approach ensures the breadth and authenticity of our dataset, laying the groundwork for robust fault detection and classification in power systems.

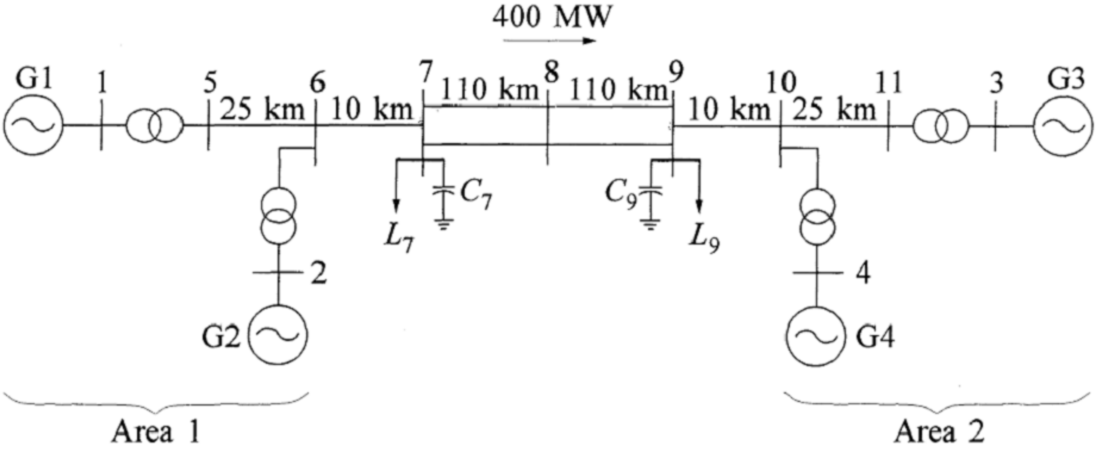


Figure 4.2: 2 Area Kundur System.

Derived from a complex 2-area electrical power system with 11 buses, 4 generators, and 4 transformers, our data set captures the intricate dynamics found in real power networks. The post-fault data collection strategy, involving diligent recording of voltage and current data after each fault event, facilitates a detailed analysis of the system's responses. The data set encompasses a wide range of fault types, including single-line-to-ground, line-to-line, line-to-line with ground, triple-line-to-line, and triple-line-to-line with ground faults. By incorporating these diverse fault types and strategically selecting fault regions, our data set becomes a faithful representation of real-world power system conditions. This comprehensive data set is pivotal to our research, providing a solid foundation for fault detection and classification studies.

4.2 Creation and Splitting of 3D array

A three-dimensional array was generated by grouping the data set based on the Series_ID. This resulting 3D array was subsequently partitioned into training and testing data sets, maintaining a ratio of 70:30, where both the training and testing data are represented as 3D arrays. The dimensions of these arrays are delineated as the total number of Series_IDs, the number of samples within one signal, and the number of features. This systematic approach ensures a structured representation of the data for the subsequent training and evaluation processes in the context of the machine learning model.

4.3 Flattening of Training and Testing data

The process of flattening the 3D arrays ‘x_train‘ and ‘x_test‘ involves reshaping them by applying a flattening operation. This operation is carried out by multiplying the number of samples within an individual signal by the total number of features present in that particular signal. The reshaping ensures that the data retains its structural integrity while conforming to a two-dimensional format, facilitating compatibility with subsequent processing steps.

4.4 Choosing the model and its hyperparameters

4.4.1 Random Forest Classifier

Random Forest is an ensemble learning algorithm widely used for both classification and regression tasks in machine learning. It operates by constructing a multitude of decision trees during training and outputting the mode (for classification) or mean prediction (for regression) of the individual trees. The key innovation lies in the randomness introduced during both the construction of each tree and the selection of features for each split. This randomness helps to mitigate overfitting and enhances the model’s generalization performance. The algorithm’s strength lies in its ability to handle high-dimensional data, capture complex relationships, and provide robust predictions by aggregating the diverse outputs of numerous decision trees.

Key terms associated with a random forest classifier:

1. **Decision Tree**: A decision tree is a flowchart-like structure where each internal node represents a decision based on a particular feature, and each leaf node represents the outcome.

2. **Bagging (bootstrap aggregating)**: creating subsets of training data through sampling with replacement.

3. **Feature Subset Selection**: Randomly selecting features for each tree to enhance diversity.

4. **Entropy and Information Gain**: Measures of impurity and attribute effectiveness in decision trees.

5. **Out-of-Bag Error**: Evaluating model performance on instances not used for training.

6. **Hyperparameters**: Pre-set parameters like the number of trees and tree depth.

7. **Node impurity**: Measure the impurity or disorder of a set of data points within a specific node of a decision tree. In the Random Forest algorithm, impurity is used as a criterion to determine how well a node separates the data into classes.

4.4.2 Key hyper-parameters of a Random Forest Classifier

1. **Number of Trees (n_estimators)**: Key Point: Determines the total number of decision trees in the forest. Impact: Higher values can enhance performance but increase computational cost.

2. **Tree Depth (max_depth)**: Key Point: Governs the maximum depth of each decision tree. Impact: Control overfitting; deeper trees capture complex patterns.

3. **Minimum Samples Split (min_samples_split)**: Key Point: Specifies the minimum samples needed to split an internal node. Impact: Influences tree structure; higher values prevent small, noisy splits.

4. **Minimum Samples Leaf (min_samples_leaf)**: Key Point: Sets the minimum

samples for a leaf node. Impact: Controls node size; crucial for preventing over-fitting.

5. **Maximum Features (max_features)**: Key Point: Limits the number of features considered for each split. Impact: Enhances model robustness by controlling feature diversity.

Our model is configured with the following hyper-parameters: n_estimators=500, max_features='log2', max_depth=10, min_samples_split= 5, min_samples_leaf= 2. These settings define the behavior of the Random Forest classifier.

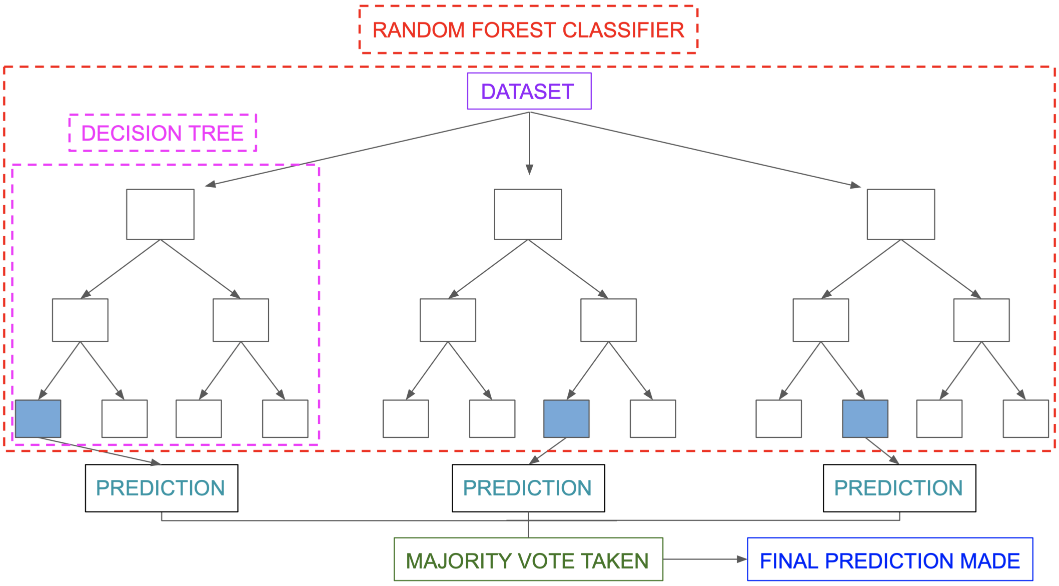
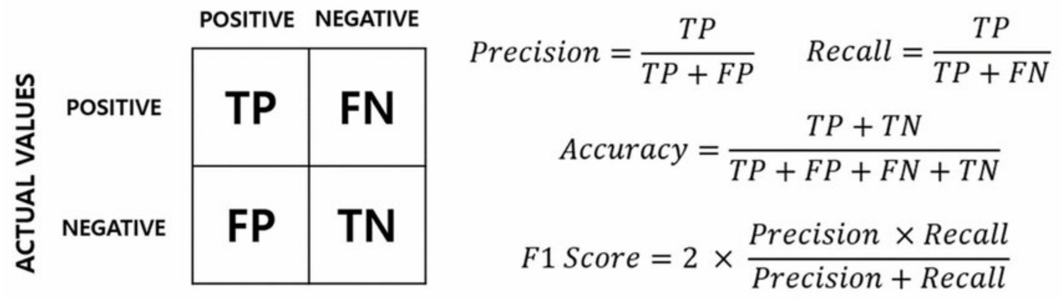


Figure 4.3: Random Forest Classfier

4.5 Prediction and Evaluation of the model

The random forest classifier is employed to make predictions on the test dataset after being trained on the corresponding training dataset. The trained classifier, configured with specified hyperparameters such as 500 decision trees, a maximum depth of 10 levels for each tree, and constraints on the minimum number of samples required to split an internal node and the minimum number of samples required to be at a leaf node, is applied to the flattened features of the test set. The predicted labels are then compared to the true labels, and performance metrics such as accuracy and a detailed classification report are computed. The accuracy score quantifies the proportion of correctly predicted labels in the test set, while the classification report provides additional metrics such as precision, recall, and F1-score for each class. This evaluation process offers insights into the model's effectiveness in generalizing from the training set to unseen data, helping to assess its overall performance and identify potential areas for improvement.



$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

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

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

Figure 4.4: Evaluation Metrics

Our model's performance has been thoroughly assessed using key metrics such as accuracy, precision, recall, and F1-score. The training accuracy stands at **90%**, showcasing the model's proficiency in learning from the training data set. In the testing phase, the model achieved an accuracy of **86%**, indicating its capability to generalize well to unseen data. The detailed classification reports for both training and testing data sets are presented in the following tables:

Class	Metrics			Support
	Precision	Recall	F1-Score	
1	1.00	1.00	1.00	3
2	1.00	0.67	0.80	3
3	0.75	1.00	0.86	6
4	1.00	0.50	0.67	2
Accuracy: 0.86				
Macro Avg: 0.94 / 0.79 / 0.83				
Weighted Avg: 0.89 / 0.86 / 0.85				

Table 4.1: Test Set Classification Report

4.6 Confusion matrix and Line plot

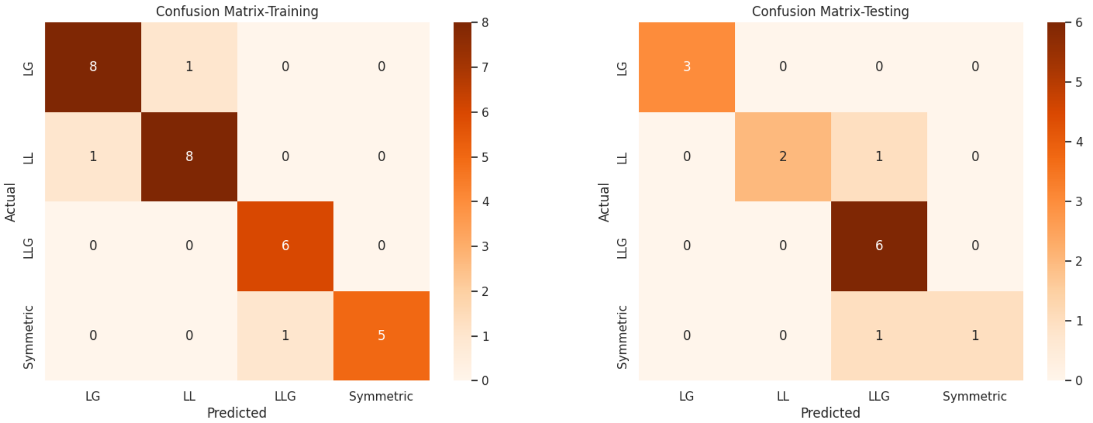
The utilization of a confusion matrix and line plot significantly enhances the understanding of model results in classification tasks. The confusion matrix provides a comprehensive breakdown of the model's predictions, detailing the true positive, true negative, false positive, and false negative instances for each class. This matrix serves as

Class	Metrics			Support
	Precision	Recall	F1-Score	
1	0.89	0.89	0.89	9
2	0.89	0.89	0.89	9
3	0.86	1.00	0.92	6
4	1.00	0.83	0.91	6
Accuracy: 0.90				
Macro Avg: 0.91 / 0.90 / 0.90				
Weighted Avg: 0.90 / 0.90 / 0.90				

Table 4.2: Training Set Classification Report

a valuable tool for assessing the model's performance, enabling the calculation of metrics such as precision, recall, and accuracy. Additionally, a line plot offers a visual representation of model predictions over a continuum, allowing for the identification of patterns and trends. By juxtaposing the confusion matrix with a line plot, one can gain a holistic insight into the model's strengths and weaknesses, facilitating informed decision-making and refinement of the classification model.

The confusion matrices and line plots for training and testing data are given below:



(a) Confusion Matrix for training

(b) Confusion Matrix for testing

Figure 4.5: Confusion Matrix

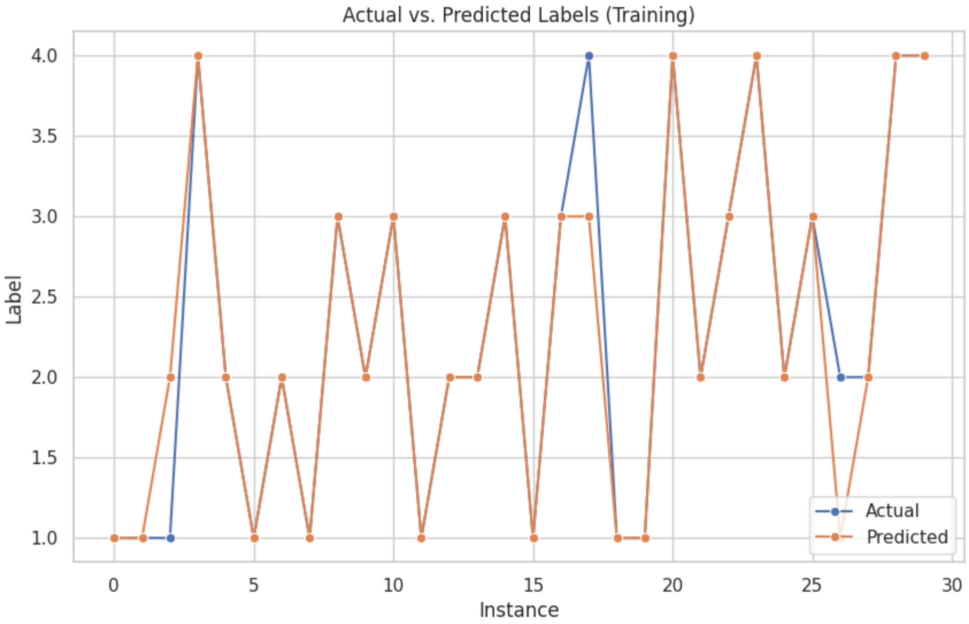


Figure 4.6: Line Plot for Training data

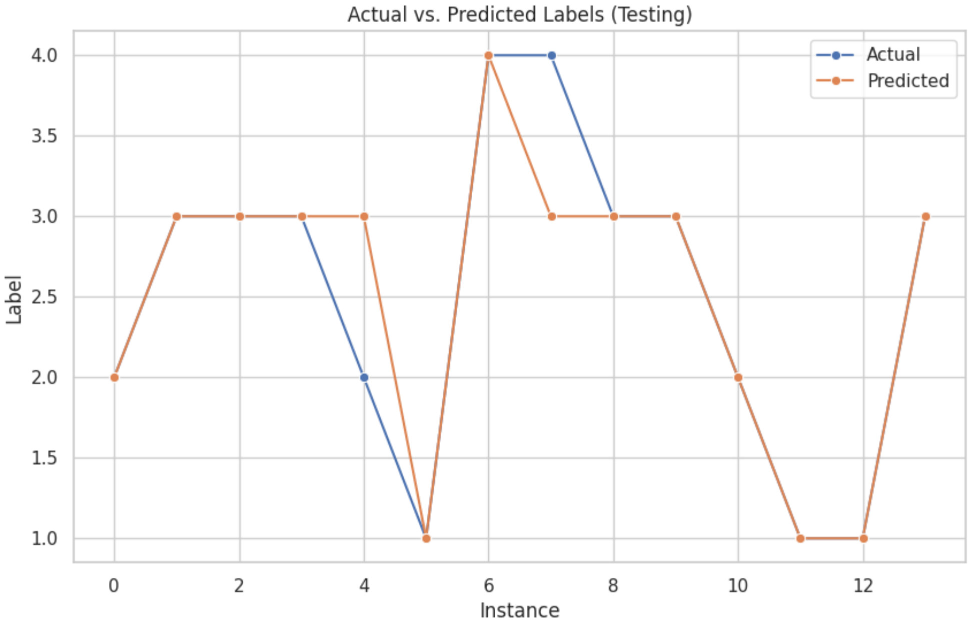


Figure 4.7: Line Plot for Testing data

CHAPTER 5

CONCLUSION AND FUTURE WORK

In conclusion, this report has delved into the intricacies of the electrical power system, focusing on its three critical subsystems: generation, transmission, and distribution. The susceptibility of transmission and distribution lines to faults, stemming from environmental exposure, was examined, emphasizing the detrimental impact of these faults on power system equipment. Notably, the report addressed the necessity of promptly addressing faults due to associated maintenance challenges, user inconvenience, and financial losses. The classification of transmission line faults, categorized as symmetrical and unsymmetrical, was discussed, with an emphasis on the challenges posed by existing literature's limited focus on specific fault locations. The report presented a novel approach to data generation from a comprehensive 2-area, 11-bus system, considering multiple non-fault locations. Leveraging this dataset, a Random Forest classifier was employed, achieving a 90% accuracy on the training dataset and 86% accuracy on the testing dataset. This robust performance underscores the effectiveness of the proposed methodology in fault classification. Overall, this report contributes valuable insights to the field of power system fault analysis and underscores the importance of comprehensive data generation for enhanced model accuracy.

As for future work, to enhance accuracy and further classify symmetric faults into triple-line faults and triple-line-to-ground faults, future work should involve incorporating sequence currents derived from the desired fault location. Analyzing positive-sequence, negative-sequence, and zero-sequence currents becomes crucial for precise fault differentiation. Specifically, distinguishing between triple-line faults and triple-line-to-ground faults can be facilitated by leveraging the absence or presence of zero-sequence current components. Additionally, future efforts should extend the analysis to include source location detection, providing a more comprehensive understanding of fault dynamics. Furthermore, expanding the scope to a more complex power system, such as the 32-bus system, will contribute valuable insights into the scalability and generalizability of fault classification methodologies. Overall, these future endeavors aim to bolster the accuracy and applicability of fault classification in power systems.

REFERENCES

[1] S. M. Chopdar and A. Koshti, “Fault detection and classification in power system using artificial neural network,” in *2022 2nd International Conference on Intelligent Technologies (CONIT)*, 2022, pp. 1–6.

[2] O. N. Teja, M. S. Ramakrishna, G. Bhavana, and K. Sireesha, “Fault detection and classification in power transmission lines using back propagation neural networks,” in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 2020, pp. 1150–1156.