Fault Classification in Transmission Lines using Machine Learning

Nakul Dhanani and Suryansh Dixit
B.Tech., Indian Institute of Technology (BHU), Varanasi, 221005
Email: nakul.dhanani.cd.eee23@itbhu.ac.in,
suryansh.dixit.cd.eee23@itbhu.ac.in

***Abstract*—Faults in power transmission lines or grids can arise from various sources such as adverse weather conditions, animal interference, or equipment failure. Timely and accurate detection of these faults is important for maintaining a reliable and continuous power supply. This exploratory project aims to develop a model capable of detecting and classifying faults in a three-phase power supply system based on input values of currents and voltages provided by the user. We explore both conventional and modern fault detection techniques. Traditional physical methods include the Terminal Method, which utilizes bridge techniques with resistors to detect faults from one or both cable ends, and Drone Aerial Photography, used primarily for visual inspection of insulators. On the technical side, advanced approaches include Voltage/Current Threshold Monitoring and the application of machine learning algorithms for automated fault classification. In this project, we employ various machine learning models such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), XGBoost, Decision Trees, and Random Forest to identify and categorize the types of faults, if any, in the transmission line. The results aim to improve the fault detection process, making it more accurate and less reliant on manual inspection.**

I. Introduction

Faults in transmission lines can be broadly classified into two categories: open circuit faults and short circuit faults.

A. Open Circuit Faults

Open-circuit faults occur for many reasons, including environmental disturbances such as conductor breakage due to wind, ice, or falling trees. The open circuit faults can be of three types:

1) Single Line Fault: A single line fault is generated when one conductor is broken or disconnected. The single line fault is depicted in Figure 1(a). One of the three phases gets disconnected(open circuited) during the fault. The figure depicts the fault in R-Line of transmission system. This fault causes unbalanced load conditions which can damage sensitive equipment or cause malfunctions.

2) Double Line Fault: A single line fault is generated when one conductor is broken or disconnected. The single line fault is depicted in Figure 1(a). One of the three phases gets disconnected(open circuited) during the fault. The figure depicts the fault in R-Line of transmission system. This fault causes unbalanced load conditions which can damage sensitive equipment or cause malfunctions.

3) Triple Line Fault: A triple line fault occurs when all the three phases become open circuited. This results in complete power interruption and no power reaches the load. But in this case as no current will flow through any of the lines. Thus the chance of overheating or overcurrent is zero. But this type of fault causes sudden loss of power and can cause system crashes, data loss, or production loss. The figure 1(c) depicts this fault.

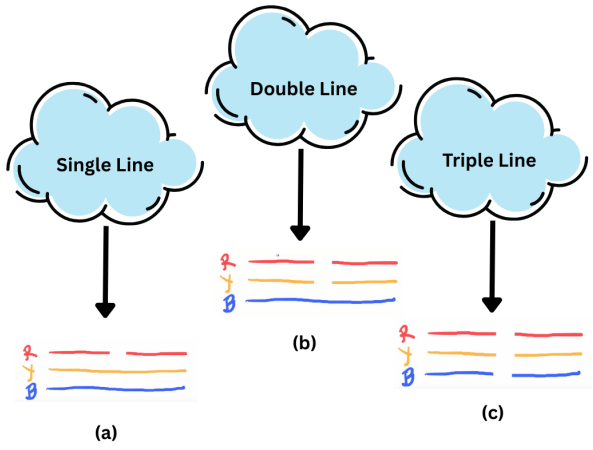

Fig. 1. Illustration of single line, double line, and triple line open-circuit faults in a 3-phase power system.

B. Short Circuit Faults

Short circuit faults are defined as the faults which occur when a low-resistance path is created between two or more conductors(lines) or between conductors and ground. This leads to a sudden surge of current, which can damage equipments and disrupt power flow. The waveforms so obtained during faults can be classified into two categories: 1. Symmetrical faults and 2. Asymmetrical faults.

1) Symmetrical Faults:

- L-L-L Fault – This fault arises when all the three phases(conductors) get short circuited. The system remains symmetrical, but a large fault current flows. The waveform thus generated is symmetrical about time axis, resulting in a zero dc voltage and current output. This

causes an insulation failure of all the three phases. The circuit and the waveforms of voltages and currents are shown in the figure 2(a).

- L-L-L-G Fault – In this case, all the conductors are simultaneously gets connected with each other and with ground. This causes a very high current to flow from them, even higher then LLL fault due to grounding. This fault is Very rare but severe. Figure 2(b) depicts the circuit and waveform conditions.

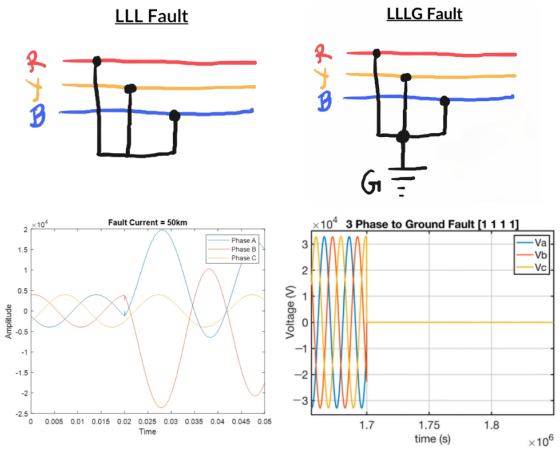

Fig. 2. Illustration of single line, double line, and triple line open-circuit faults in a 3-phase power system.

2) Unsymmetrical Faults:

- L-G Fault – One conductor gets short circuited(either R, Y, or B) to the ground. This is the most common fault in transmission lines. The main cause of this fault is the fallen conductors due to strong winds, tree contact, storm, lightening, etc. This can be omitted by use of shackle insulators. This causes an imbalance in the phases and creates zero sequence currents. The L-G fault scenario is depicted in the figure 3(a).
- L-L Fault – This type of fault is observed when two conductors get short circuited with each other. This causes a very high current to flow from the lines as voltage difference is very high between the two phases. This can be a consequence of strong winds, falling branches or debris, poor insulation or damaged equipment, etc. This is depicted in figure 3(b).
- L-L-G Fault – The fault occurs when two lines get connected with each other and also with the ground. This is rare to happen but is possible in some cases. The circuit diagram and current waveform are shown in the figure 3(c).

This paper proposes an efficient method for fault detection and classification in transmission lines using machine learning models. But first, we will look into the other works done in the related field.

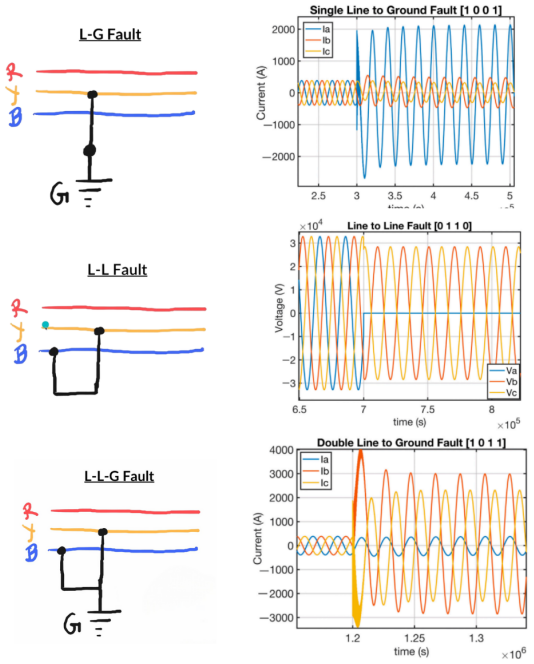

Fig. 3. Illustration of single line, double line, and triple line open-circuit faults in a 3-phase power system.

II. Related Works

This section introduces some of the prominent works done in this field.

Shameem Hasan discusses the generation of synthetic data using matlab(simulink) and then applying Neural Network Topology for classification of data.

Yanhui Xi. discusses the SA-MobileNetV3 methodology for fault detection and classification. This paper introduces the SA(shuffle attention) module in MobileNetV3 which can effectively fuse the importance of pixels in different channels and in different locations at the same channel.

Ei Phyo Thwe suggests that the ANN that was trained and tested using various sets of field data, which was obtained from the simulation of faults at various fault scenarios (fault types, fault locations and fault resistance) of 230 kV, 193.2 km in length "Mansan-Shwesaryan, Mandalay Region, Myanmar" transmission line using a computer program based on MATLAB/Simulink was a successful attempt in classification as well as the detection of faults. Simulation results confirm that the proposed method can efficiently be used for accurate fault classification on the transmission line. In this paper, we discussed various types of faults in transmission lines and presented a machine learning-based approach to detect and classify these faults. This can help reduce downtime, ensure safety, and enhance the efficiency of power delivery systems.

Ozan Turanlı uses a deep learning architecture that was based on a one-dimensional convolutional neural network (CNN) was utilized in this study. Accuracy, specificity, recall,

precision, F1 score, ROC curves, and AUC were employed as performance criteria for the suggested model. Not only synthetic but also real data were used in this study. They developed a **non linear** Convolutional Neural Network(CNN) for classification.

In this study, Md.Omaer Faruq Goni develops a spontaneous fault detection (FD) and fault classification (FC) system based on ML has been proposed. MATLAB Simulink was employed to simulate two different TLs and to generate normal and fault data (Per unit voltage and current) of ten different types. TL-1 consisted of a single generator and a single load whereas TL-2 consisted of two generators and three loads. Upon normalizing the data, an extreme learning machine (ELM) algorithm was used as the classifier. Two different ELM models were developed for FD and FC purposes through training.

III. Proposed Work

We provide a comprehensive and detailed analysis of different Machine learning models for detection and classification of faults.

A. Datasets

The datasets which we used are the "ClassData.csv" and "Detect.csv". The links are provided in the ending of this section.

1) Detect.csv: The "Detect.csv" dataset consists of line voltages and phase voltages under different fault conditions as features for detection of fault. The output label column consists of the two conditons(fault or no fault). A "0" designates a no fault condition and "1" a fault. This dataset consists of a total of 12001 rows of data including the both faulty and non-faulty conditions. The pie-chart(figure 4) depicts the amount of data present for faulty and non-faulty states.

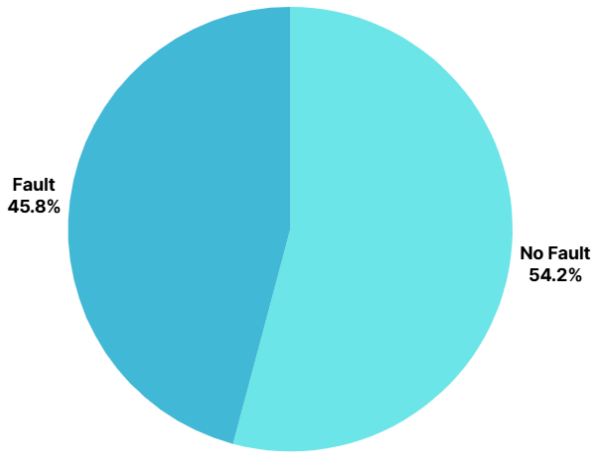

Fig. 4. Pie Chart for "Detect.csv" dataset providing insights of amount of faulty and non-faulty rows present in the dataset

The head of the dataset is shown in the figure 5. The dataset have 6505 non-faulty values and 5505 faulty values for detection of a fault. The dataset consists of the symmetric-fault

Output (S)	Ia	Ib	Ic	Va	Vb	Vc
0	-170.472196	9.219613	161.252583	0.054490	-0.659921	0.605431
0	-122.235754	6.168667	116.067087	0.102000	-0.628612	0.526202
0	-90.161474	3.813632	86.347841	0.141026	-0.605277	0.464251
0	-79.904916	2.398803	77.506112	0.156272	-0.602235	0.445963
0	-63.885255	0.590667	63.294587	0.180451	-0.591501	0.411050

Fig. 5. The figure depicts the first 5 rows (head) of the dataset representing the line currents and voltatges along with the output label assigned as fault or no-fault

detection only. This is because this is a real world dataset and it is easy to capture the short-circuit faults in data as compared to open circuit faults. The description of the dataset is provided in the figure 6.

	Output (S)	Ia	Ib	Ic	Va	Vb	Vc
count	12001.000000	12001.000000	12001.000000	12001.000000	12001.000000	12001.000000	12001.000000
mean	0.457962	6.709369	-26.557793	22.353043	0.010517	-0.015498	0.004980
std	0.498250	377.158470	357.458613	302.052809	0.346221	0.357644	0.349272
min	0.000000	-883.542316	-900.526951	-883.357762	-0.620748	-0.659921	-0.612709
25%	0.000000	-64.348986	-51.421937	-54.562257	-0.237610	-0.313721	-0.278951
50%	0.000000	-3.239788	4.711283	-0.399419	0.002465	-0.007192	0.008381
75%	1.000000	53.823453	69.637787	45.274542	0.285078	0.248681	0.289681
max	1.000000	885.738571	889.868884	901.274261	0.609864	0.627875	0.608243

Fig. 6. The figure depicts the description of the dataset including mean, standard deviation, minimum and maximum values in the data.

2) ClassData.csv: The "ClassData.csv" dataset consists of the 6 possible conditions of faulty and non-faulty conditions. The faulty conditions consists of the 5 possibilities of short circut faults L-G, L-L, L-L-G, L-L-L and L-L-L-G. The header of the dataset is shown in the figure 7. The dataset consists of

G	C	B	A	Ia	Ib	Ic	Va	Vb	Vc
1	0	0	1	-151.291812	-9.677452	85.800162	0.400750	-0.132935	-0.267815
1	0	0	1	-336.186183	-76.283262	18.328897	0.312732	-0.123633	-0.189099
1	0	0	1	-502.891583	-174.648023	-80.924663	0.265728	-0.114301	-0.151428
1	0	0	1	-593.941905	-217.703359	-124.891924	0.235511	-0.104940	-0.130570
1	0	0	1	-643.663617	-224.159427	-132.282815	0.209537	-0.095554	-0.113983

Fig. 7. The figure shows the head of the dataset "ClassData.csv" representing the features and output labels.

7861 rows with three line currents and line voltages under different faulty and non faulty conditions. The labels are presented as the fault or no-fault in Ground(G), Line-A(A), Line-B(B) and Line-C(C). For fault "1" and "0" for no fault is assigned as label.

The Description and pie-chart for the dataset is provided in figure 8 and figure 9. The relationship between Line Voltages and currents in L-L-L-G faults are shown in figure 10. The relationship between Line Voltages and currents in L-L-G faults are shown in figure 11.

Other graphs can be found in the Google Colab Notebook whose link is provided after references.

The fault can be easily depicted from the figures 10 and 11. The faulty region shows dissimilarity with the normal characterstics of current and voltage.

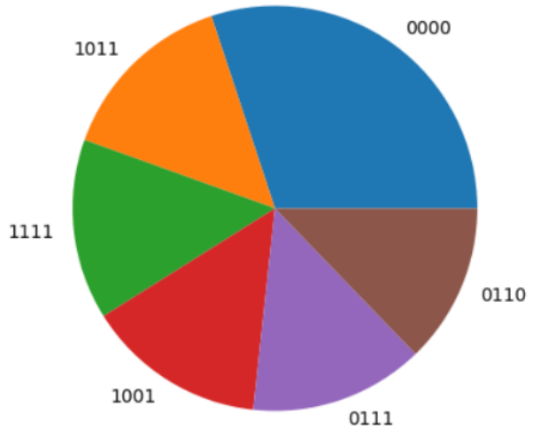

Fig. 8. The figure represents the Pie-Chart depicting the Data Distribution for different fault possibilities including no-fault condition.

	G	C	B	A	Ia	Ib	Ic	Va	Vb	Vc
count	7861.000000	7861.000000	7861.000000	7861.000000	7861.000000	7861.000000	7861.000000	7861.000000	7861.000000	7861.000000
mean	0.432006	0.411271	0.555527	0.571429	13.721194	-44.845268	34.392394	-0.007667	0.001152	0.006515
std	0.495387	0.492095	0.496939	0.494903	464.741671	439.269195	371.107412	0.289150	0.313437	0.307897
min	0.000000	0.000000	0.000000	0.000000	-883.542316	-900.526951	-883.357762	-0.620748	-0.608016	-0.612709
25%	0.000000	0.000000	0.000000	0.000000	-119.802518	-271.845947	-61.034219	-0.130287	-0.159507	-0.215977
50%	0.000000	0.000000	1.000000	1.000000	2.042805	5.513317	-4.326711	-0.005290	0.001620	0.009281
75%	1.000000	1.000000	1.000000	1.000000	227.246377	91.194282	49.115141	0.111627	0.153507	0.239973
max	1.000000	1.000000	1.000000	1.000000	885.738571	889.868884	901.274261	0.595342	0.627875	0.600179

Fig. 9. The figure depicts the description of the dataset including mean, standard deviation, minimum and maximum values in the data.

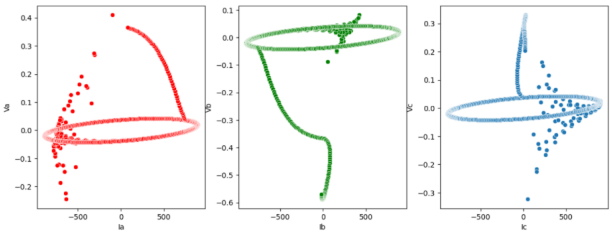

Fig. 10. The V-I charactersitcs during LLLG fault.

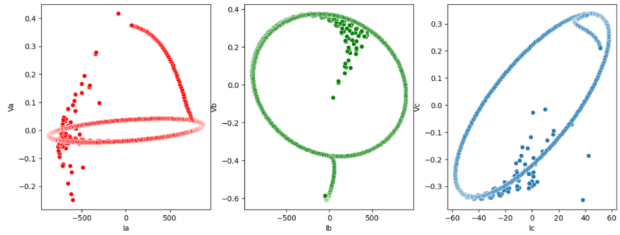

Fig. 11. The V-I charactersitcs during LLG fault.

B. Machine Learning Models

Here, we present our two works including the fault detection algoritms and fault classification algorithms. The algorithms used are as follows:

1) Support Vector Machine: Support Vector Machine (SVM) is a supervised learning algorithm that is well-suited for classification tasks such as identifying different types of faults in transmission lines. According to our datasets, phase currents and voltages serve as input features, while the type of fault is the class label.

Let the training data be $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the feature vector formed from line current and voltage measurements, and $y_i \in \{0, 1\}$ denotes the fault class label (for binary classification- fault or no fault). SVM finds a hyperplane defined by $\mathbf{w}^\top \mathbf{x} + b = 0$ that separates the fault classes with the maximum margin.

The optimization problem is given by:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad \forall i.$$

For multiclass fault classification (e.g., LL, LG, LLG, etc.), one-vs-rest classifier is employed, where multiple binary classifiers are trained.

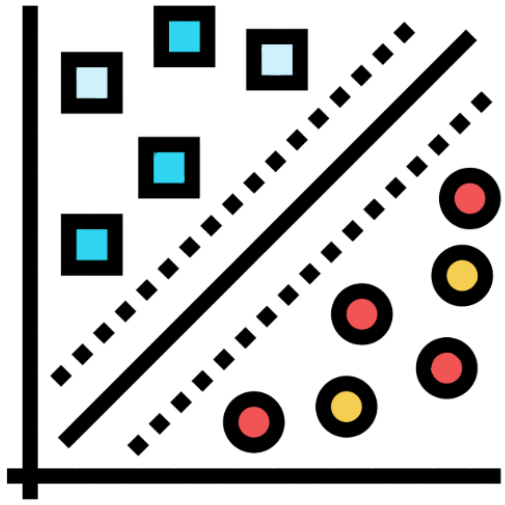

Fig. 12. Support Vector Machine (one rest classifier) Topology

2) Decision Tree: Decision Tree algorithm is used for both classification tasks and regression tasks. In our context of fault classification, the input feature vector $\mathbf{x}_i$ includes line voltages and currents, and the goal is to detect and classify the fault y_i.

The decision tree partitions the feature space in a recursive manner by selecting the feature and threshold that separates the classes at each node as optimally as possible. The best split is calculated by maximizing a purity measure for example,

the information gain (based on entropy) or **Gini index**. For example, Gini impurity for a node is defined as:

$$G = 1 - \sum_{k=1}^{K} p_k^2,$$

where p_k is the proportion of samples of class k in the node. The process continues until a stopping criterion is met, such as maximum depth or minimum number of samples per node.

Decision Trees are interpretable and efficient, but may overfit on noisy data. Ensemble methods like Random Forest and XGBoost address this limitation.

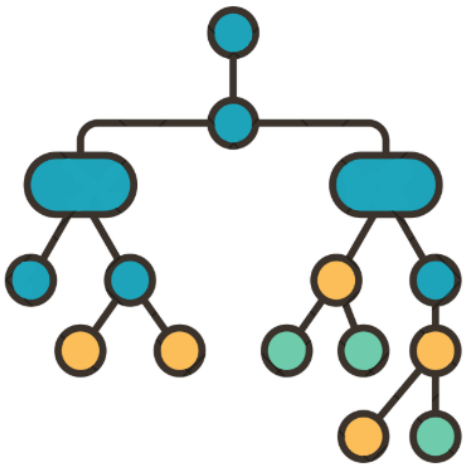

Fig. 13. Decision Tree Classifier Topology

3) K-Nearest Neighbour Classifier: The K-Nearest Neighbour classifer chooses an optimum value of k which is the number of nearest neighbours around a point (which has to be classified) and based on majority count of one of the class in the neighbour, the point is classified as of that class. Given a test point $\mathbf{x}$, KNN algorithm computes distances to all training points (typically using Euclidean distance):

$$d(\mathbf{x}, \mathbf{x}_i) = \|\mathbf{x} - \mathbf{x}_i\|_2.$$

and then based on the optimal value of k, the classifer classifies it to the class with majority in the neighbour.

4) Ensemble Techniques: These includes XGBoost and Random Forest Algorithm. The Random Forest creates multiple decision trees and takes an average out of them which helps in improving accuracy and reducing overfitting. The final prediction is then made by majority voting which is the most common class among all trees. For classification, each tree T_j outputs a predicted class:

$$\hat{y}_i^{(j)} = T_j(\mathbf{x}_i)$$

The final prediction is obtained by majority voting across M trees:

$$\hat{y}_i = \text{mode}\left\{T_1(\mathbf{x}_i), T_2(\mathbf{x}_i), \ldots, T_M(\mathbf{x}_i)\right\}$$

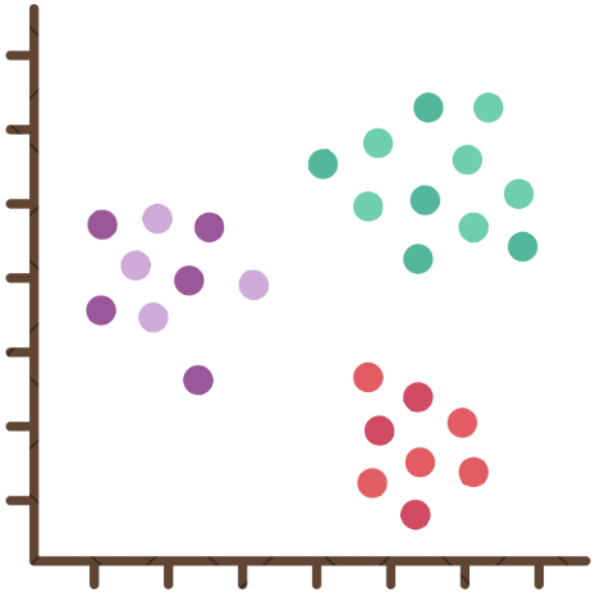

Fig. 14. KNN Classifier Topology

Extreme Gradient Boosting Algorithm or XGBoost Algorithm starts with an initial prediction building trees sequentially in such a way that each of the tree learns from the errors of the previous trees. Given a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the model predicts:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(\mathbf{x}_i), \quad f_t \in \mathcal{F}$$

where $\mathcal{F}$ is the space of regression trees.

The objective function to minimize is:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

where l is the loss function (e.g., cross-entropy), and $\Omega(f_t)$ is the regularization term:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Each new tree f_t is fit to the **negative gradients** of the loss function, effectively learning the residuals of previous predictions.

IV. Results

We had applied all of the algorithms mentioned above on both of the data sets to dectect as well as to classify the faults. The best results for detection is observed by SVM Algorithm with a testing accuracy of **99.98%**. SVM is effective in this application due to its ability to handle high-dimensional feature spaces and its robustness. The confusion matrics of all of the algorithms are attached in figure 15, 16, 17, 18 and 19.

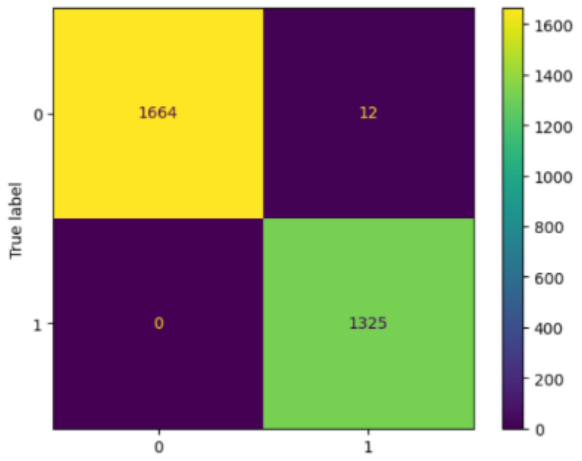

Fig. 15. SVM Confusion Matrix

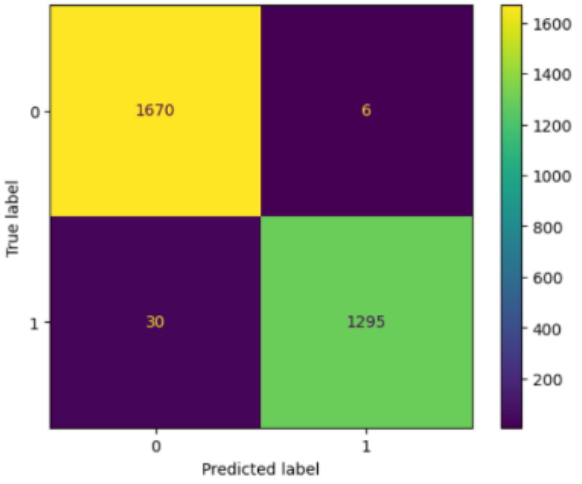

Fig. 16. Decision Tree Confusion Matrix

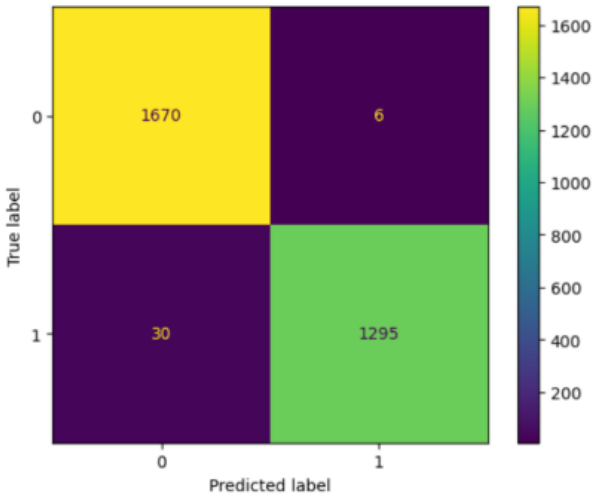

Fig. 17. KNN Classifier Confusion Matrix

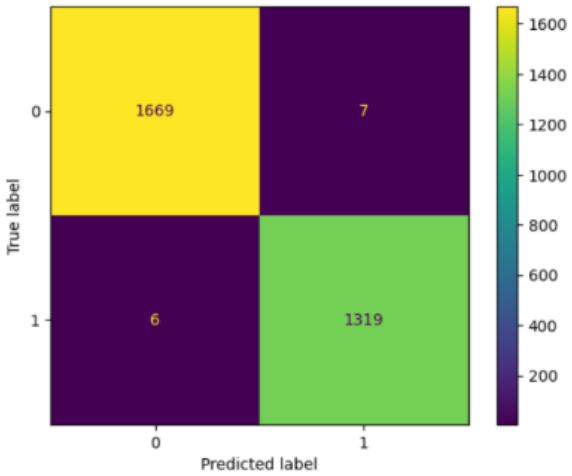

Fig. 18. XGBoost Confusion Matrix

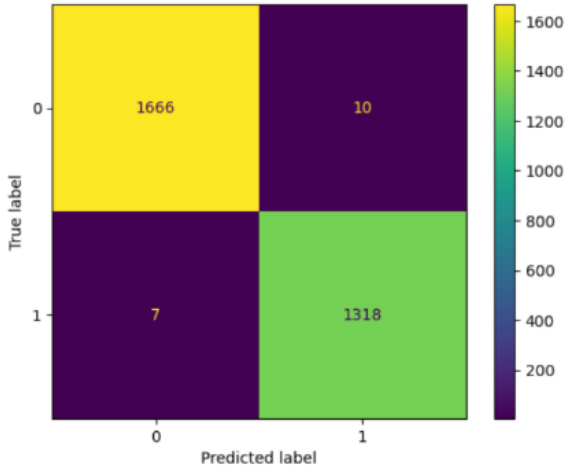

Fig. 19. Random Forest Confusion Matrix

The table depicts the testing accuracy of the five algorithm which we used for detection.

TABLE I
CONFUSION MATRIX VALUES OF DIFFERENT CLASSIFIERS FOR FAULT DETECTION

Model	TP	TN	FP	FN
SVM	1325	1664	12	0
Decision Trees	1295	1670	6	30
KNN	1295	1670	6	30
XGBoost	1319	1669	7	6
Random Forest	1318	1666	10	7

TABLE II
PERFORMANCE METRICS OF DIFFERENT CLASSIFIERS FOR FAULT DETECTION

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	99.98	99.10	100.00	99.55
DT	98.88	99.54	97.74	98.63
KNN	99.53	99.54	97.74	98.63
XGB	99.56	99.47	99.55	99.51
RF	99.43	99.25	99.47	99.36

A similar analysis has been conducted for classification of the faults. This was a multiclass classification problems with labels GCB and A. So four confusion matrices are formed for each label column for each algorithm as shown in the figures 20, 21, 22, 23 and 24.

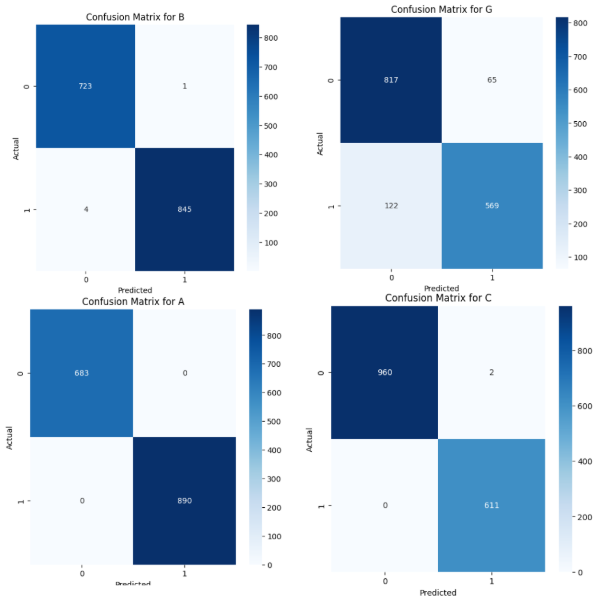

Fig. 20. SVM Confusion Matrix

TABLE III
ACCURACY METRICS FOR EACH LABEL (G, C, B, A) - SVM

Label	TP	TN	FP	FN	Total	Accuracy (%)
B	845	723	1	4	1573	99.68
G	569	817	65	122	1573	88.13
A	890	683	0	0	1573	100.00
C	611	960	2	0	1573	99.87

TABLE IV
ACCURACY METRICS FOR EACH LABEL (G, C, B, A) - DECISION TREE

Label	TP	TN	FP	FN	Total	Accuracy (%)
G	610	811	71	81	1573	90.43
C	610	961	1	1	1573	99.87
B	844	723	1	5	1573	99.62
A	890	683	0	0	1573	100.00

TABLE V
ACCURACY METRICS FOR EACH LABEL (G, C, B, A) - KNN

Label	TP	TN	FP	FN	Total	Accuracy (%)
G	539	738	144	152	1573	80.85
C	609	960	2	2	1573	99.75
B	843	724	0	6	1573	99.62
A	889	682	1	1	1573	99.87

TABLE VI
ACCURACY METRICS FOR EACH LABEL (G, C, B, A) - XGBOOST

Label	TP	TN	FP	FN	Total	Accuracy (%)
G	545	751	131	146	1573	82.40
C	611	961	1	0	1573	99.94
B	846	724	0	3	1573	99.81
A	890	683	0	0	1573	100.00

TABLE VII
ACCURACY METRICS FOR EACH LABEL (G, C, B, A) - RANDOM FOREST

Label	TP	TN	FP	FN	Total	Accuracy (%)
G	591	801	81	100	1573	88.57
C	610	961	1	1	1573	99.87
B	847	723	1	2	1573	99.81
A	889	683	0	1	1573	99.94

TABLE VIII
CLASSIFICATION TESTING ACCURACIES OF DIFFERENT MACHINE LEARNING ALGORITHMS

Algorithm	Testing Accuracy (%)
Support Vector Machine (SVM)	87.73
Decision Tree	89.95
XGBoost	82.10
Random Forest	88.36
K-Nearest Neighbors (KNN)	80.54

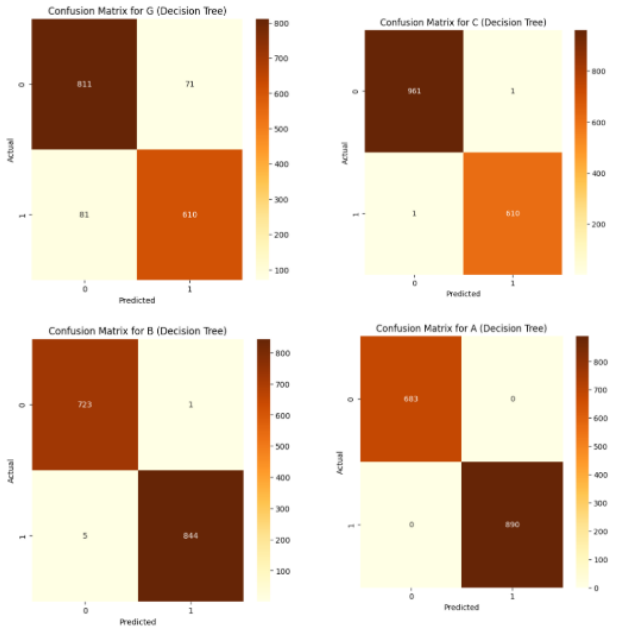

Fig. 21. Decision Tree Confusion Matrix

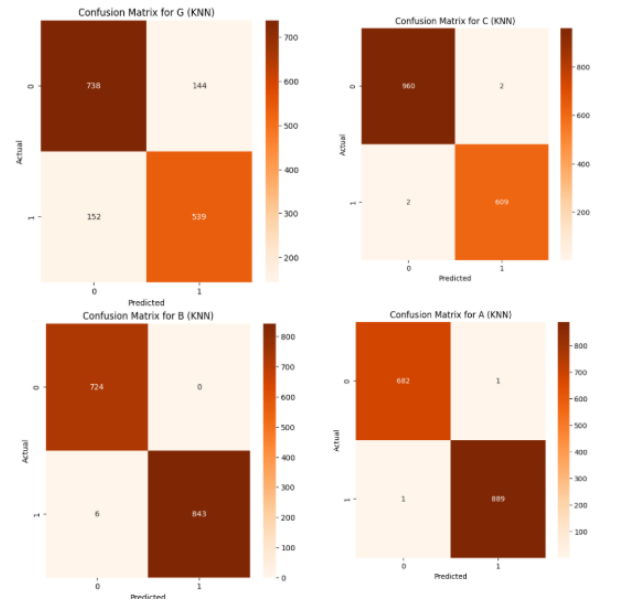

Fig. 22. KNN Confusion Matrix

V. Conclusion

In this exploratory project, we explored the field of faults in transmission lines starting from classifying them and then used Machine Learning Models on two different datasets for detection and classification of faults. The results thus obtained indicated the usefullness of each model, specially SVM for fault detection and Random Forest for fault classification. We achieved an accuracy of **99.98%** by SVM for detection and **89.95%** by Random Forest for classification of faults.

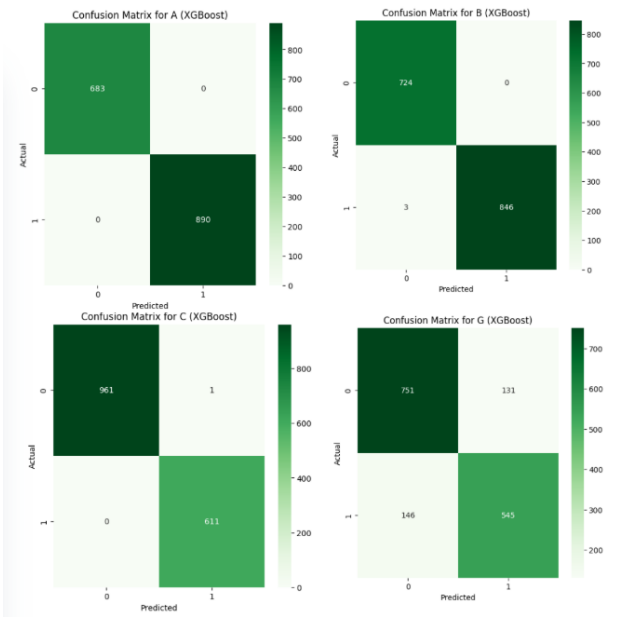

Fig. 23. XGBoost Confusion Matrix

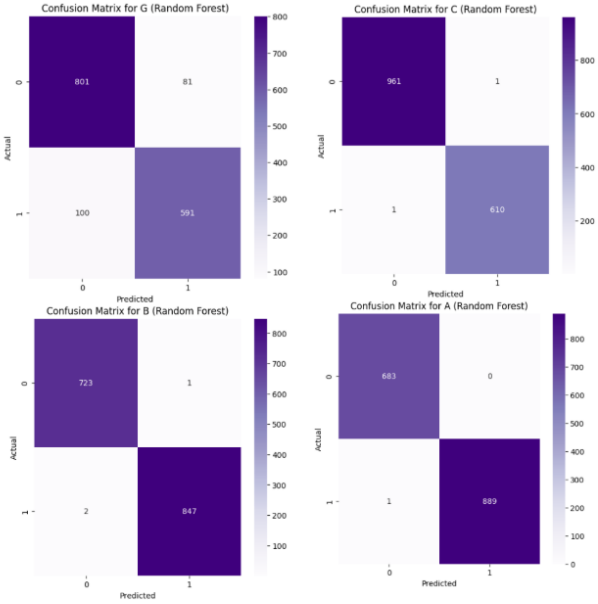

Fig. 24. Random Forest Confusion Matrix

VI. Acknowledgment

The authors would like to thank their faculty advisors and research coordinators at IIT(BHU) for their continuous guidance and encouragement throughout this project.

References

[1] Fault Occurrence Detection and Classification Fault Type in Electrical Power Transmission Line with Machine Learning Algorithms
[2] Transmission line fault detection and classification based on SA-MobileNetV3
[3] Fault Detection and Classification for Transmission Line Protection System Using Artificial Neural Network

[4] Classification of Faults in Power System Transmission Lines Using Deep Learning Methods with Real, Synthetic, and Public Datasets
[5] Fast and Accurate Fault Detection and Classification in Transmission Lines using Extreme Learning Machine
[6] Google Colab Notebook for Fault Detection and Classification