A Thesis Submitted to Sylhet Engineering College for the Degree of **Bachelor of Science in Electrical and Electronic Engineering**

CatBoost Algorithm for Transmission Line Fault Detection and Classification

By

Himel Ahmed

Tanbir Rahman

&

Talab Hasan

Supervised by

Arif Ahammad

Assistant Professor

Department of Electrical and Electronic Engineering

Shahjalal University of Science & Technology, Sylhet

January, 2023

Sylhet Engineering College, Sylhet

Affiliated with

Shahjalal University of Science & Technology

The thesis titled **"CatBoost Algorithm for Transmission Line Fault Detection and Detection"** submitted by Tanbir Rahman, Student ID: 2017338532; Talab Hasan, Student ID: 2017338551 and Himel Ahmed, Student ID: 2017338514; Session 2016-2017, to the Department of Electrical and Electronic Engineering, Sylhet Engineering College, has been accepted as satisfactory in partial fulfilment of the requirement for the Degree of Bachelor of Science in Electrical and Electronic Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

____

Md. Shahid Iqbal
Assistant Professor and Head
Department of Electrical and Electronic Engineering
Sylhet Engineering College, Sylhet.

Chairman

____

Arif Ahammad
Assistant Professor
Department of Electrical and Electronic Engineering
Shahjalal University of Science & Technology.

Supervisor and Member

____

Md. Ashraful Alam
Lecturer
Department of Electrical and Electronic Engineering
Sylhet Engineering College, Sylhet.

Member

____

Salman Fazle Rabby
Lecturer
Department of Electrical and Electronic Engineering
Sylhet Engineering College, Sylhet.

Member

Acknowledgement

We begin by praising the Almighty, who is the founder and creator of the universe and is the supreme force. He has provided us with the passion and enthusiasm & showers of blessings to accomplish all of the qualifications successfully for the Bachelor of Science in Electrical and Electronic Engineering (EEE) degree, including this thesis work.

Next, we would like to express our sincere & heartfelt gratitude to our respected supervisor **Arif Ahammad**, Assistant Professor, Department of EEE, Shahjalal University of Science & Technology, for his productive academic advice, innovative instruction, endless support, valuable suggestions, and countless support throughout the study & made this work possible. His excellent monitoring & guidance has benefitted us tremendously.

It is also an honor to thank **Md. Shahid Iqbal**, Assistant Professor & Head and another faculty members **Salman Fazle Rabby**, Lecturer, Department of EEE and **MD. Ashraful Alam**, Lecturer, Department of EEE for their excellent assistance, insightful conversations, encouragement and collaboration throughout this journey.

We are grateful to all of those whom we have had the pleasure to work with, our batchmates from the EEE Department of SEC for their invaluable assistance in completing the Bachelor Degree successfully.

Nobody has been more important to us in the pursuit of this thesis more than our family. We are extremely grateful to our parents & siblings for their love, prayers, care and sacrifices for educating and preparing us for our future.

Abstract

Recently, Bangladesh experiences heavy power cuts and a system loss (power transmission & distribution loss) of 11.11% in FY 2020-21. And, fault in the transmission line is one of the key reasons behind this. This paper proposed the CatBoost machine learning method for detecting electric power transmission line faults and classification. The method is formed on the theory of the repeated combination of weaker classifiers based on the tree structure. The proposed method is adaptive towards balanced, imbalanced, and noisy data states. Also, the method achieves 100% accuracy in detection and 99% accuracy in classification. Besides, the designed method eliminates the need for explicit pre-processing of categorical data, has high time performance, and is resilient towards various operating conditions of transmission system parameters such as fault resistance, and distance.

Keywords: *Transmission line, faults, classification, Machine learning algorithm, CatBoost, Noise, Data imbalance.*

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Overview

Generally, a power system consists of three parts. They are mainly generation, distribution, and transmission parts. To efficiently transport bulk amounts of energy at high voltage without loss, power transmission lines are used to interconnect power plants and substations as well as for connections between substations and from generation to supply end. Therefore, it plays a significant part in the provision of electricity. Moreover, different types of faults in transmission lines may occur because of natural disasters i.e. heavy winds and lightning, as well as human interference or any abnormal condition of the power system. So, the protection of transmission lines is very important. And, a robust protection system should be used along with its continued development to overcome power system failures. Typically, protective relays are often used to identify the transmission line fault defects which needs a long time to identify the fault. From Bangladesh's standpoint, the subject is crucial. In FY 2020-21, Bangladesh has a system loss (power transmission & distribution loss) of 11.11% [1]. Another set of reports also mentioned that Bangladesh is struggling hard to ensure electricity supply to its 160 million citizens in 2021 due to frequent power cuts, which further affects badly to the business and dissuades foreign investment [2, 3]. And, fault in the transmission line is one of the main ones of frequent power cuts [4]. Therefore, detecting and classifying the fault types in the shortest possible time is the highest priority.

As technology has advanced, new methods and procedures have been developed that can handle problems quickly and effectively. Over the past 20 years, several sectors have made significant progress toward categorizing and detecting faults in power systems [5]. The traditional methods rely on events that have been recorded, but the new ways of approaches such as machine learning work with system data. There are different methodologies invented in recent times which overcome one another’s limitations with their continuous development process [6]. Mainly three types of techniques are used to diagnose transmission line faults [7].

i) Prominent methods, which include wavelet transformation-based analysis (WT), artificial neural network (ANN), and fuzzy logic approach.

(ii) Hybrid techniques that include various combinations of these fundamental techniques and finally,
(iii) Modern methods, such as support vector machine (SVM), different artificial intelligence (AI)-based methods, principal component analysis (PCA)-based approaches, deep learning based techniques and many others [8].

The purpose of this research work is to develop an adaptive machine learning model to detect and classify the faults which can works well on different data states as well as on the different operating conditions of the transmission line. To do so, i.e. detection and classification of the fault in the transmission line using data from data measurements unit, such as current and voltage signals are taken as input and the output shows the fault/no fault, if the fault then classifies this, a prototype of power transmission system is developed in MATLAB/SIMULINK and a model of machine learning using CatBoost algorithm is developed in Python.

1.2 Significance of CatBoost Algorithm for Fault detection and Classification

Voltage and current variation from regular values is referred to as an electrical fault. Power system components or lines operate with pre-determined voltages and currents under normal conditions, making the system safer to operate**.** But when a fault occurs, it causes excessively high currents to flow which causes damage to equipment and devices. There are mainly two types of faults in the electrical power system:
1. Symmetrical Faults.
2. Unsymmetrical Faults.

1. Symmetrical faults:
These are very severe faults and occur rarely in power systems. These are also called balanced faults and are of two types: line to line to ground (L-L-L-G) and line to line (L-L-L). Only 2-5 percent of system faults are symmetrical faults. If these faults occur, the system remains balanced but severely damages the electrical power system equipment.

2. Unsymmetrical Faults:

There are mainly three types. Mainly, single line-to-ground (L-G), line-to-line (L-L), and double line-to-ground (LL-G) fault.

Line-to-ground fault (L-G) is the most common fault and 65-70 percent of faults are of this type. It causes the conductor to contact the earth or ground. Double line-to-ground (LL-G) faults, which cause the two conductors to contact the ground, occur for 15 to 20 percent of all faults. 5-10 percent of faults are line-to-line (L-L) faults.

This happens when two conductors come into touch with one another, typically while wires hang in the wind. Since their occurrence takes the system out of balance, these errors are also known as imbalanced faults.

There are many severe effects on the power system when its fault occurs. The fault generated an enormously large and powerful fault current. The equipment utilized in the power system network may be damaged by the fault current. A huge amount of fault current generates heat which causes a high temperature and mechanical stress in the conductors. There is always a hazard of fire due to arcing caused by heavy currents. If the fault remains for a long time, then the fire may spread to other portions of the system. Rotating machines associated with the system may become heated by imbalanced current and voltage. Generators may stop synchronizing because of an imbalance in current and voltage which results in a complete shutdown of the system and, in the worst case, a blackout. So, the issue has an impact on system stability and reliability. And, in Bangladesh the traditional fault detection procedure involves relay based protective system which needs a long time to detect and classify the fault types. Thus, hampers reliability and creates instability and discontinuity in the system. Hence, detecting and classifying the fault types in the shortest possible time is the highest priority. Rapid fault classification assists notably in locating and clearing the faulty system quickly. Increases the reliability, and stability and provides continuity (uninterrupted power) of the system. Machine learning, an extremely potent, dynamic artificial technology offers outstanding performance to support classification within a short period [9]. And, in this work CatBoost is chosen because of the following reasons:

- It performs well on balanced, imbalanced as well as on noisy data condition.
- It maintains good performance metrics along with changes in different operating condition of the transmission line such as the resistance and the distance variation.

- It achieves good accuracy in both cases of detection and classification task.
- It doesn't require explicit pre-processing of categorical data. Hence, it has great time performance.
- It can be performed on small data as well as on big data also.

1.3 Thesis Organization

The thesis report is organized as follows:

Chapter 1: *Introduction,* Overview of the research work is presented in this section.

Chapter 2: *Literature Review,* related background and scope of contributions is presented in this section.

Chapter 3: *Proposed CatBoost Model,* Overall methodology proposed for this research work is presented in this section.

Chapter 4: *Results and Discussion***,** Outcomes obtained from this research work, performance evaluation, etc. is presented in this section.

Chapter 5: *Conclusion,* the thesis is concluded in this section with some opinion and recommendations.

Chapter 2: Background & Previous Work

This chapter focuses on introducing fundamental ideas about fault detection and classification and doing a literature review of prior research of machine learning uses for transmission line fault detection and classification.

2.1 Traditional approach

2.1.1 What is protective relay

A protective relay is a device that is utilized to identify the fault and initiate the circuit breaker's operation to isolate the faulty elements from the rest of the system.

2.1.2 Working principle of protective relay

The relays measure the electrical parameter (voltage, current, frequency, and phase angle) continuously in both normal and faulty conditions. According to this, detect the abnormal condition of the transmission line. Through changes in the parameter in abnormal conditions, the faults indicate their presence, type, and location to the relays. As the fault is detected, the relay initiates to close the trip circuit of the breaker. Therefore, the contacts of the breaker are opened and separation of the faulty section from the rest of the system.

2.1.3 Types of relays

Protective relays are available in several types and provide a wide range of functions in electrical power systems. Protective relays can be of various common sorts, such as:

- Overcurrent relay: This type of relay is used to detect excessive current in a circuit and trip the circuit breaker to protect against electrical overload. It is mainly used in the low-voltage circuit as it needs low cost.
- Directional relays: These relays can be utilized to prevent reverse current flow since they can determine the direction of current flow in a circuit. Mainly used in combination with some other relay.
- Differential relays: These relays are working on comparing the current flowing into a transformer with the current flowing out of it and trip the circuit breaker if there is a significant difference between the two.

- Distance relay: These relays determine the electrical distance between two points in a power system and can trip the circuit breaker if the distance exceeds a certain threshold. It is utilized together with the impedance relay.
- Pilot relays: A pilot relay is used to control the operation of other relays or electrical circuits. These relays work by using the transmission network that connects two distant substations. Wire, microwave pilot, and power line carriers are the 3 most used pilot relays in protecting functions.

2.2 Machine learning based approaches

2.2.1 What is machine learning

Since the beginning of time, humans have used a variety of instruments to complete various jobs more quickly. The inventiveness of the human mind resulted in the development of numerous machines. These devices made life easier for humans by allowing them to fulfill a variety of demands, such as travel, industry, and computing. And among these, there is machine learning (ML). Simply, machine learning is the method to learn by a machine from data through rigorously followed procedure of training, testing and optimizing. The application of ML teaches computers how to handle data more effectively. Sometimes, even after viewing the data, we are unable to evaluate or extrapolate the information. In that case, we utilize machine learning. The availability of a large number of datasets has increased the demand for machine learning which uses a variety of techniques to resolve data issues.

2.2.2 Evolution of machine learning:

The development of machine learning has a long history, which starts from the year 1950. Here the key milestones are described shortly:

- 1950s: the field of Artificial intelligence is established, with the ambition of making intelligent machines that can perform different tasks that would require human-like intelligence.
- 1960s: The first development of Neural networks, inspired by the function and structure of the human brain.

- 1980s: Decision trees, a machine learning algorithm are developed and become used widely for tasks of credit fraud detection and diagnosis of medical.
- 1990s: Support vector machine (SVM), is another kind of machine learning that can achieve advanced performance in many sectors.
- Late 1990s: Boosting algorithms, which combine multiple weak models to develop a stronger model, can show great performance.
- 2000s: Neural networks, which had previously fallen in disfavour, are improved leading to the development of deep learning algorithm
- 2010s: the area of machine learning continues to enlarge, with numerous developments of the new algorithm and application of ML to large range of industries.

2.2.3 Types of Machine Learning:

It can be broadly classified as four types and is shown in below in Figure 2.1.

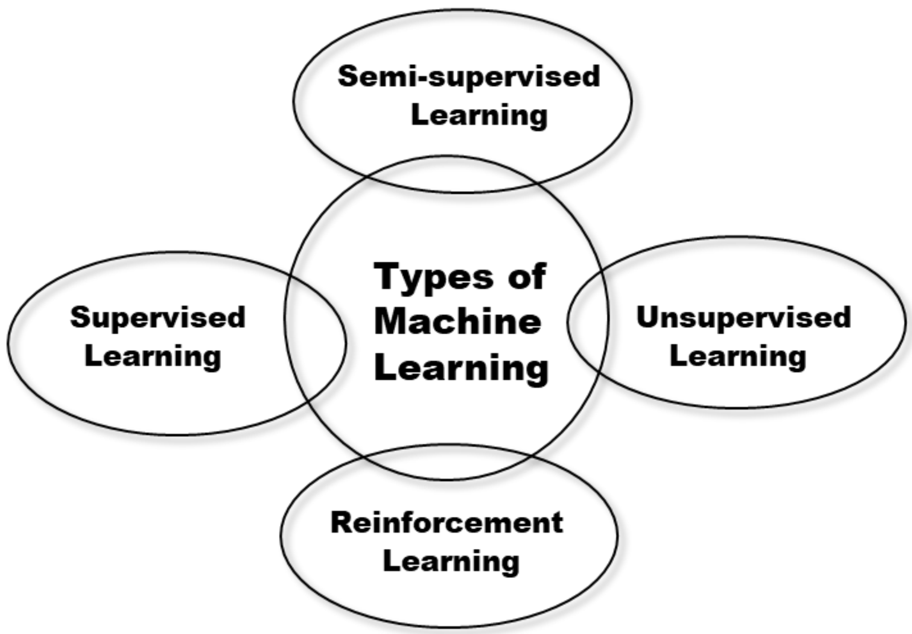


Figure 2.1: Types of Machine learning

2.2.3.1 Supervised Machine Learning Algorithms

As the name suggests, supervised learning involves a supervisor serving as an instructor. In its simplest form, supervised learning refers to the process of guiding or training a computer system utilizing labelled data. Which indicates that the right answer has already been assigned to certain data. For the supervised learning algorithm to analyse the training data (set of training examples) and provide an accurate result from labelled data, the machine is then given a fresh set of examples (data).

There are mainly two types of supervised learning: regression and classification.

2.2.3.2 Unsupervised Machine Learning Algorithms

Unsupervised learning is a technique in which models are not supervised using a training dataset. Instead, the model itself finds hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things. The goal of unsupervised learning is to find the underlying structure of the dataset, group that data according to similarities, and represent that dataset in a compressed format.

2.2.3.3 Reinforcement machine learning algorithms

It is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or a penalty.

Since there is no labelled data, the agent is bound to learn by its experience only. The agent interacts with the environment and explores it by itself. The primary goal of an agent in reinforcement learning is to improve performance by getting the maximum positive rewards.

2.2.3.4 Semi-supervised Machine Learning Algorithms

This technique is a combination of supervised and unsupervised learning algorithms. It utilizes a combination of labelled and unlabelled datasets during the training period. It can be effective in those fields of machine learning and data mining where the unlabelled data is already present and getting the labelled data is a monotonous process. With more common supervised machine learning methods, we train a machine learning algorithm on a “labelled” dataset in which each record includes the outcome information.

And, as mentioned earlier, mainly three types of techniques are used to diagnose transmission line faults.

i) Prominent methods, which include wavelet transformation-based analysis (WT), artificial neural network (ANN), and fuzzy logic approach.

(ii) Hybrid techniques that include various combinations of these fundamental techniques and finally,

(iii) Modern methods, such as support vector machine (SVM), different artificial intelligence (AI)-based methods, principal component analysis (PCA)-based approaches, deep learning based techniques and many others.

2.3 Review of the literature

Since the early stages of research, machine learning techniques was used to detect and classify the faults which involve various approaches such as the Support vector machine(SVM), Decision tree, artificial neural network, etc. [6], [7], [8]. Although the decision tree and SVM have greater accuracy among them, especially SVMs are well-suited for classification problems with a small number of cases and a high number of features. In these cases, other algorithms may struggle to find a decision boundary, but SVMs are capable to find a boundary that maximizes the distance between the different classes, which can lead to better generalization of new data, but ANN has drawn the attention of researchers because of its ability to parallel processing of data, learn quickly, and due to adjusting only a few parameters for training [9]. Such as, [10], [11], [12] come up with the backpropagation neural network, the autonomous neural network, which achieves significant speed, and accuracy but didn't examine the model against the noisy condition. Moreover, with the advancement of time scholars gives outstanding efforts to evolve the approaches which proposed many brand-new and hybrid techniques. [13] Proposed DWT-based ANN that works with the noisy condition but the accuracy of the model deteriorates in terms of different signal-to-noise ratios. In addition, recent literature [14], [15], [16] suggests deep learning-based, ultrafast and precise SAT-CNN, long short-term memory-CNN, and CNN architecture. However, studies in [17] present that CNN requires a high sampling rate to achieve good accuracy, which further increases the network size and computation time. Additionally, deep learning cannot learn much from a little amount of data and inhibits generalization due to overfitting [18]. Also in real-world appliances, each class of any classification problem doesn't maintain proportional equality. That means, in most cases data for each class are imbalanced in nature [19]. And, deep learning experiences less performance for classification tasks with imbalanced data [20].

Due to the drawback of these issues learned in literature, it is requisite to develop a method that overcomes the shortcoming of the abovementioned methods. In the next chapter, to detect and classify the transmission line faults, the CatBoost model is designed and implemented considering variations in various operating conditions of transmission system parameters such as fault resistance, distance, etc. as well as variations on different data states also.

Chapter 3: Proposed CatBoost Model

CatBoost is built on the principle of the sequential iterations of weaker classifiers. It is less prone to overfitting problems, has a default overfitting detector system, and performs well for a small amount of data and big data. No need for hot encoding and good in terms of accuracy, speed and time performance [21], [22]. However, assume $M = \{(Xp, Yp)\}p=1,\ldots,n$ is the sample of data where Xp is the set of input features and Yp is the response features. Then, the objective of the model training is to pick the Yp depending on the provided Xp for any given case (classification, regression). This section provides concise details of the simulation, data preparation, and model implementation.

3.1 Transmission line model development

The model of the power transmission system that is considered for generating the training data used for model implementation is shown in figure 3.1.

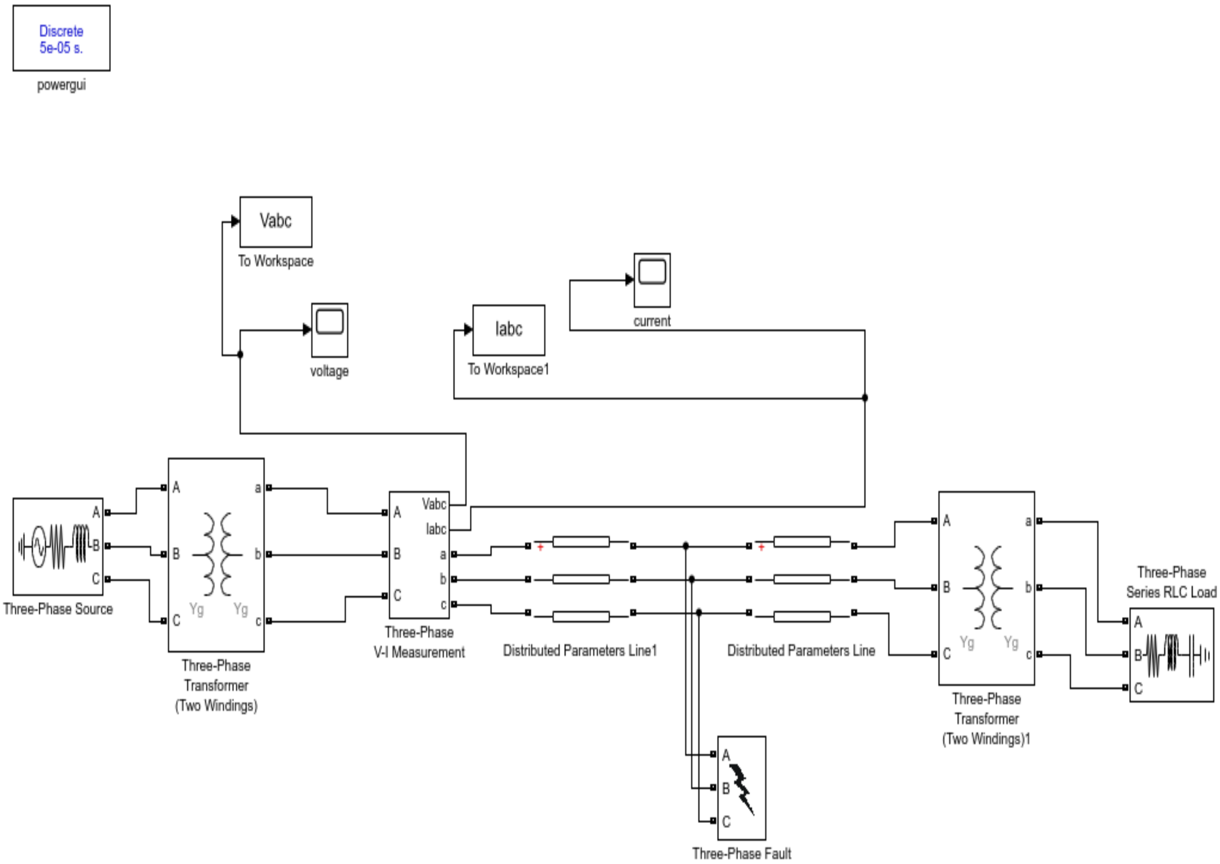


Figure 3.1 Transmission system Simulink diagram

The transmission line is 200Km long, has a frequency of 50Hz, and is operated at 132 kV with two sources (one for generation, the other is load) connected at each end. The whole system is developed using the Simulink environment. Also, the zero and positive sequence parameters of the transmission line as shown in table 3.1.

Table 3.1 Parameters of the transmission line

Parameters	**Symbols**	**Value**
Zero and positive sequence resistances	$R_{\mathbf{0}}, R_1$	0.11240, 0.044965 (Ω/km)
Zero and positive sequence inductances	$L_{\mathbf{0}}, L_1$	0.00202, 0.00101 (H/km)
Zero and positive sequence capacitances	$C_{\mathbf{0}}, C_1$	4.394, 7.471 (nF/km)

The sampling frequency of generated three-phase instantaneous voltages and currents signal is 12 kHz. Besides, various operating condition is also considered in the simulation. Such as signal is generated at fault resistance of 0.001, 50, and 100 Ω and fault distances of 50 and 100km. Further, in total 5 types of faults, phase A to G, A to B, A to B to G, A to B to C, and no fault signals are generated. Samples of the fault and no-fault waveforms are shown in figure 3.2.

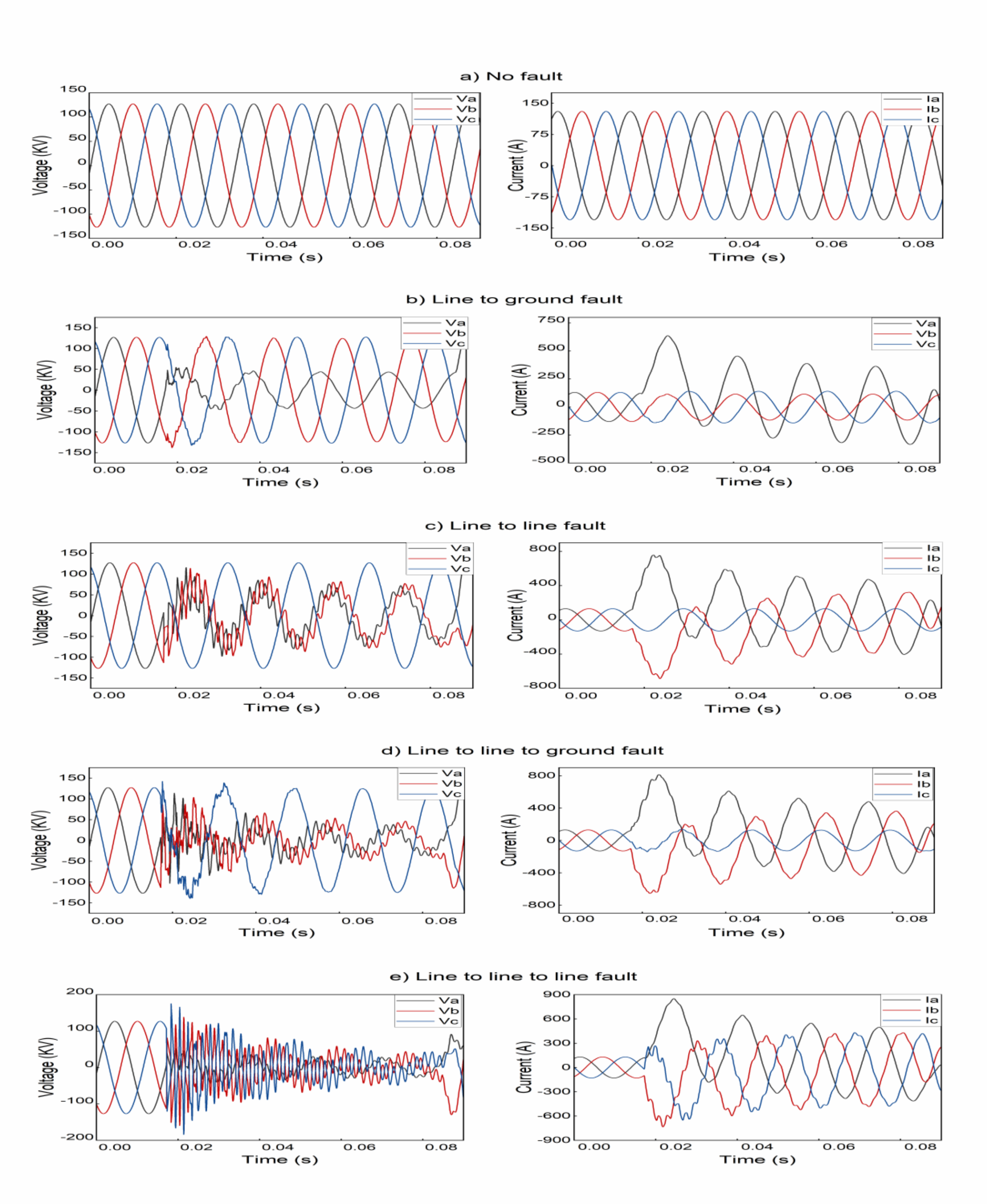


Figure 3.2 Sample waveforms for data collection.

From figure 3.2 (a) it is clear that without fault the voltage and the current waveform are sinusoidal. When fault figure 3.2 (b-e) arises, it is seen that faulty phase voltages decreased and the currents increased rapidly. Also, there is heavy distortion.

3.2 Data preparation

Data is the most fundamental need of the machine learning model. Data is collected or generated from various sources according to their function. The collection of data is called a dataset. It is very important to have a quality dataset. Because the datasets that have been utilized to train or test each machine learning model are ultimately what determines how well it performs. Any machine learning model's accuracy and error depend significantly on the data that were used. Hence, the model will be more effective if the data is more precise and organized.

After obtaining signals, it is required to achieve labeled datasets for model implementation. Thus, data is extracted from waveform using the Simulink model. For the detection task, the objective is to detect faults or not faults. For the classification task, the objective is to classify line to ground (LG), line to line (LL), line to line to ground (LLG), line to line to line (LLL), and no fault. Therefore, for detection, a dataset of 1000 data containing three-phase instantaneous voltages and currents labeled as the faulty and no-faulty state is prepared. And, for classification, a dataset of 2500 data containing three-phase voltages and currents labeled as above specified types is prepared. Diversification of data (as discussed in section 3.1) is made in both circumstances (detection and classification) to ensure almost like real-world problems. Since we prepared datasets from raw data, it is obvious to properly check and thus eliminate the null value or any values that has no relation with function for proper results. Although classification data involves categorical features, there is no need to explicitly preprocess the data because CatBoost allows using non-numeric values.

3.3 Model implementation

The model is designed and implemented using python programming language. So, at first it is necessary to-

- Import libraries
- Import datasets

A. Importing libraries

For the processing of data, python has some predefined library functions, which are used vastly in machine learning. The used libraries for model implementation are introduced below:

NumPy:

The core Python library for performing scientific calculations is called NumPy. As a result, it is utilized to add any kind of mathematical procedure to the code. Large multidimensional arrays and matrices can be inserted into the code using NumPy.

Pandas:

One of the most well-known Python libraries is utilized for organizing and importing datasets. It is an open-source library for data analysis and manipulation.

Matplotlib:

In python, matplotlib is used to plot 2D plotting of different types of charts. We must import the pyplot sub-library. For the code, this library is used to plot any kind of chart.

CatBoost:

In this work, prebuild CatBoost library is used by optimizing the hypermeter for better performance.

B. Import dataset

Now, have to import the dataset.

Then, data is reorganized as input and output combination. Input is three-phase instantaneous voltage and currents. And, for detection output is binary (fault or no-fault). For classification, the output is multiclass (LG, LL, LLG, LL, and no-fault). Then, the entire data is separated into two parts taking a large integer (80%) for training and twenty percent for testing purposes. Finally, the model is implemented considering three categories of data. Figure 3.3 shows the first and second categories of data, the balanced and imbalanced conditions.

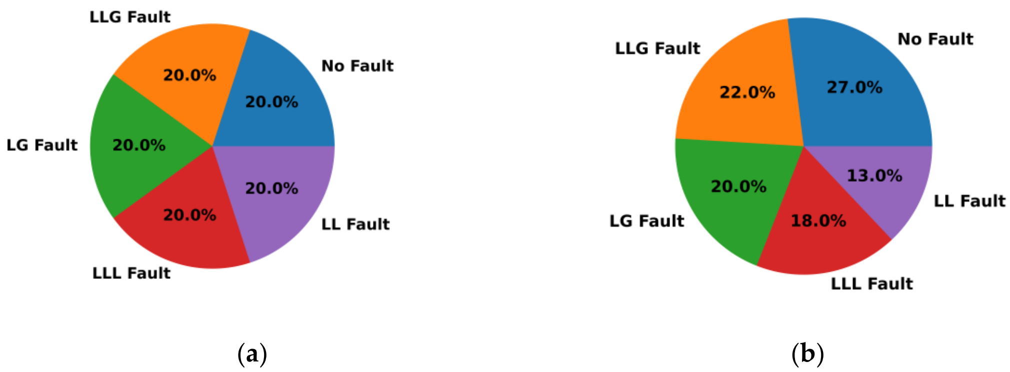


Figure 3.3 Data category : (**a**) Balanced condition (**b**) Imbalanced condition

In addition, the model is also examined for the third category, the noisy condition. To do so, Additive White Gaussian Noise is applied to three-phase voltages and currents signal. The added signal-to-noise ratio of Gaussian Noise is 20, 37dB. Figure 3.4 shows some samples of the noisy waveform.

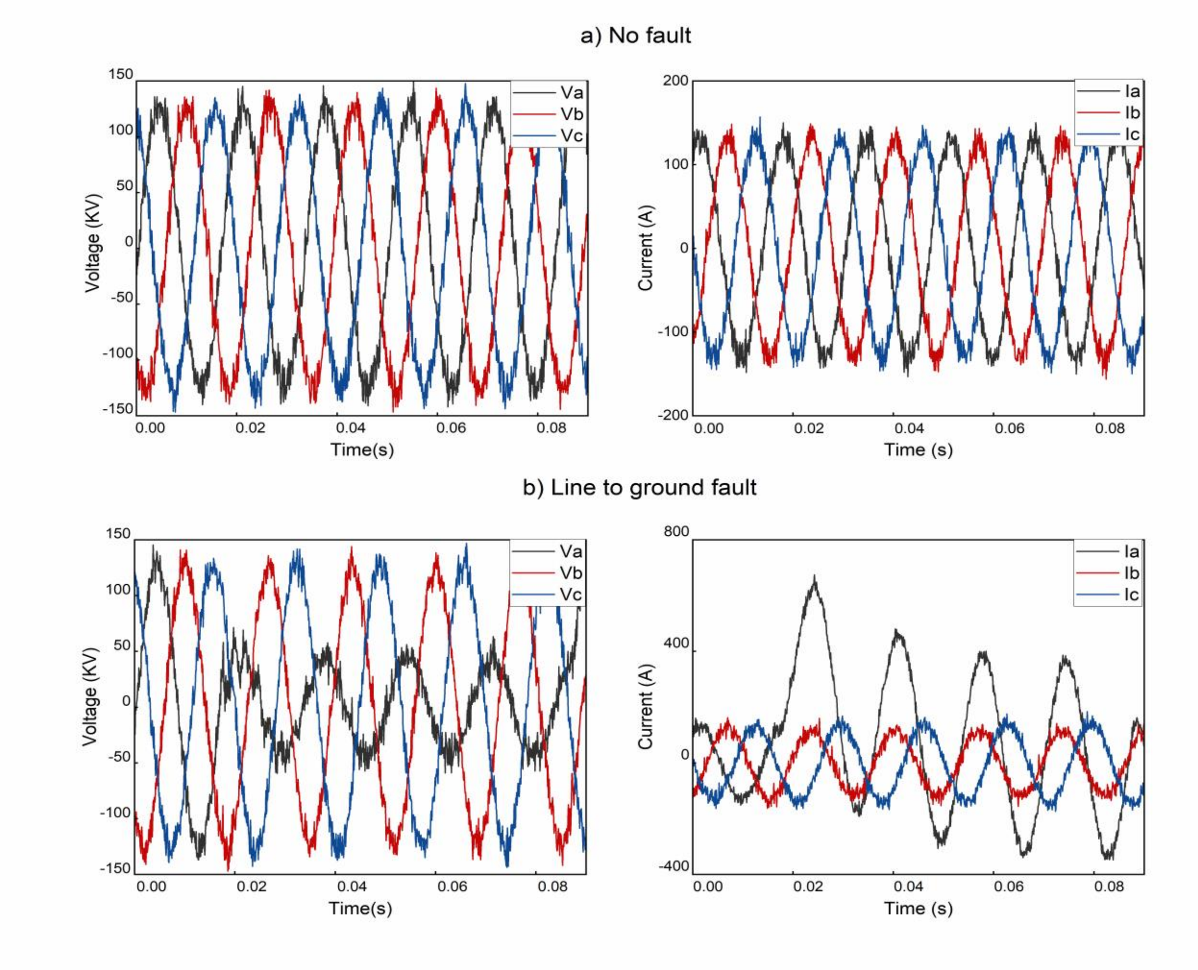


Figure 3.4 Sample waveforms for noisy condition data collection.

Also, for achieving the best performance the model hyper-parameter is optimized as presented in table 3.2.

Table 3.2 Optimized parameters.

Hyper-parameter	Classification	Detection
Number of iterations	50	50
Loss function	Multiclass	logloss
Leaf estimation method	Newton	Newton
Learning rate	0.5	0.5
Number of trees	50	50

Chapter 4: Results & Discussion

The outcome of the model for tasks involving the detection and classification of power transmission line faults is discussed in this chapter. In the discussion section, the model is also analyzed using a variety of performance metrics.

4.1 Results

4.1.1 Detection of fault

After implementing the CatBoost model, it shows outstanding performance for all three categories (balanced, imbalanced, and noisy data). It achieves 100% accuracy for detection in all three categories. The confusion matrix is shown in figure 6.

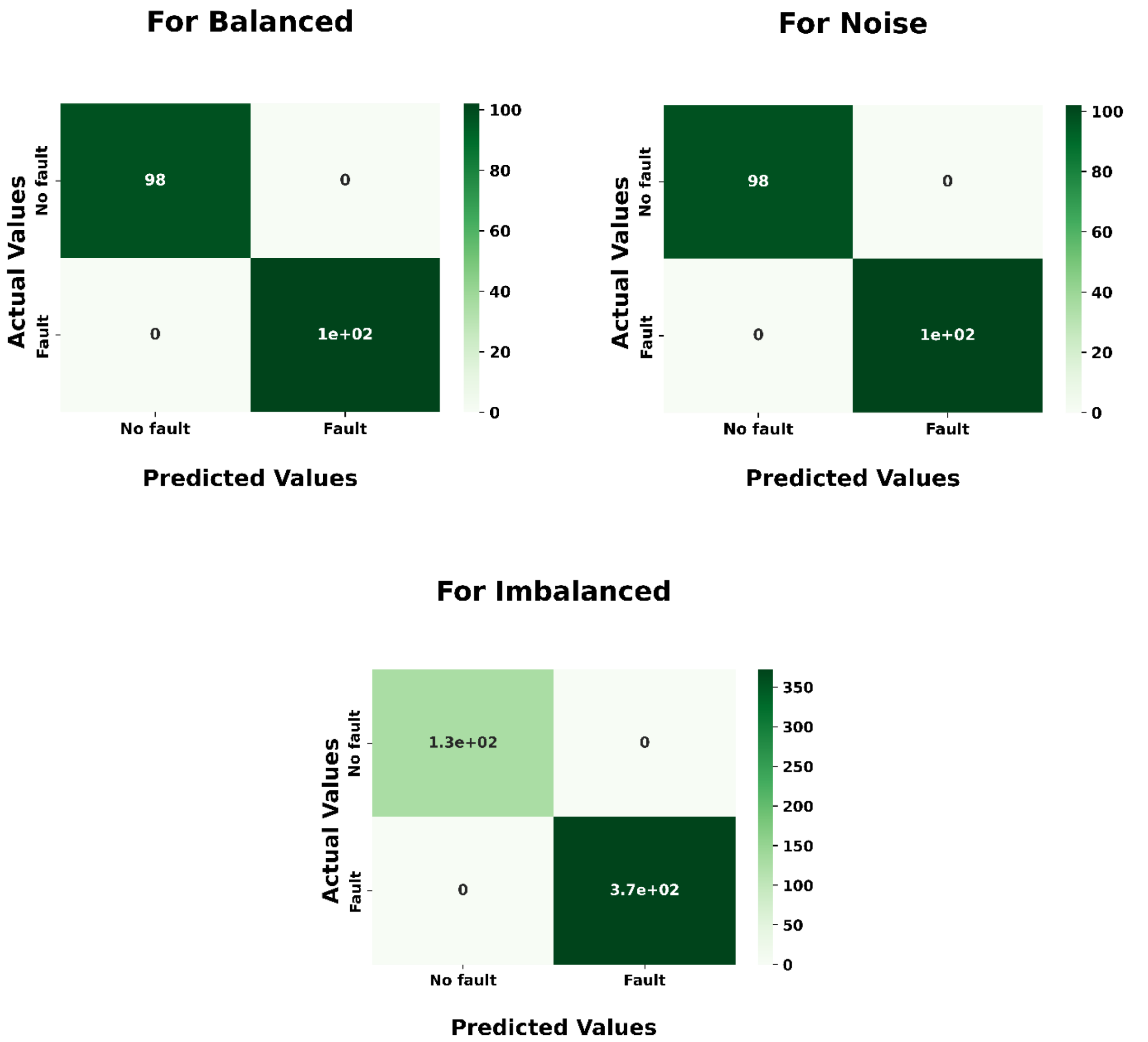


Figure 4.1: Detection confusion matrix a) for balanced condition (**b**) for noisy condition (c) for imbalanced condition.

4.1.2 Classification of fault

Similarly, the model is also a good fit for the classification of faults. Details of the results as described in the below sub-section:

4.1.2.1 Balanced condition test

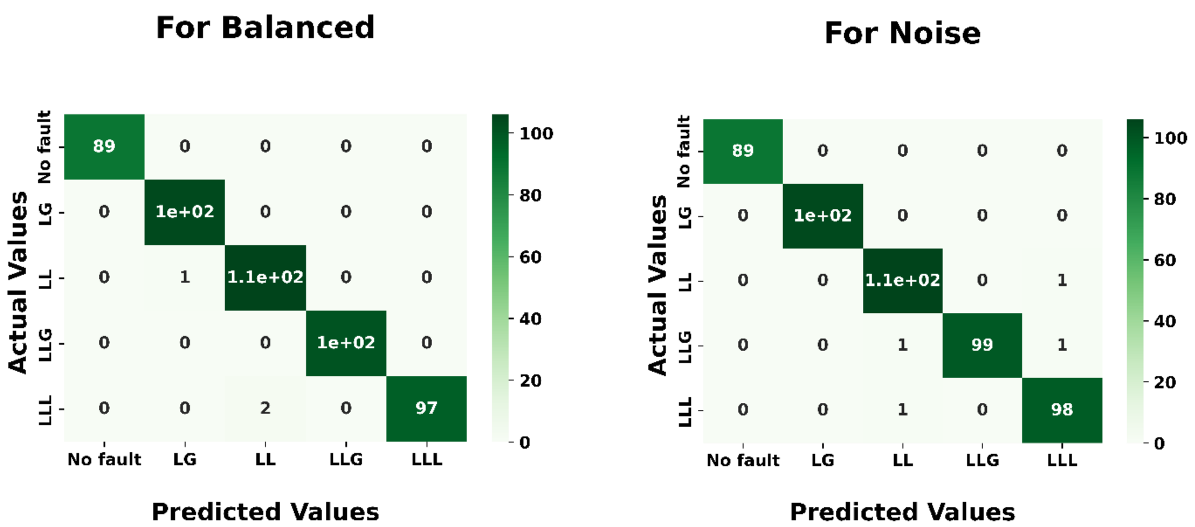
In balanced data conditions, the model acquires 99% accuracy to classify the predefined types of faults. Figures 8 and 9 show the graphical representation of the accuracy and loss. Also, the confusion matrix is shown in figure 7.

4.1.2.2 Noisy condition test

In noisy data conditions, the model also performs well. No deterioration occurs in accuracy. It acquires 99% accuracy to classify the faults. Figure 7 shows the confusion matrix. Also, the accuracy and loss graph is shown in figures 8 and 9.

4.1.2.3 Imbalanced condition test

In imbalanced data conditions, the model also acquires the same accuracy. It achieves 99% accuracy to classify the fault types. Figure 8 and 9 show the graphical representation of the accuracy and loss. Additionally, the confusion matrix is shown in figure 7.



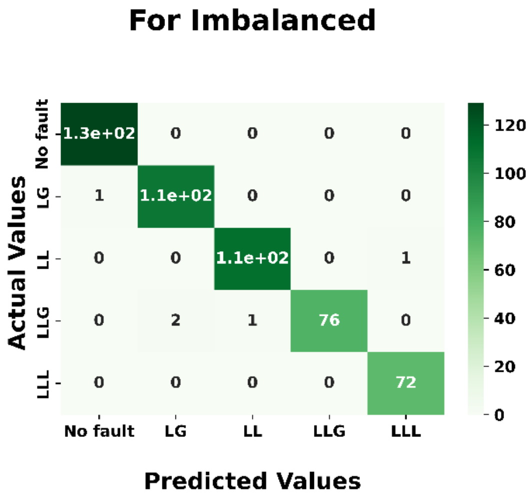
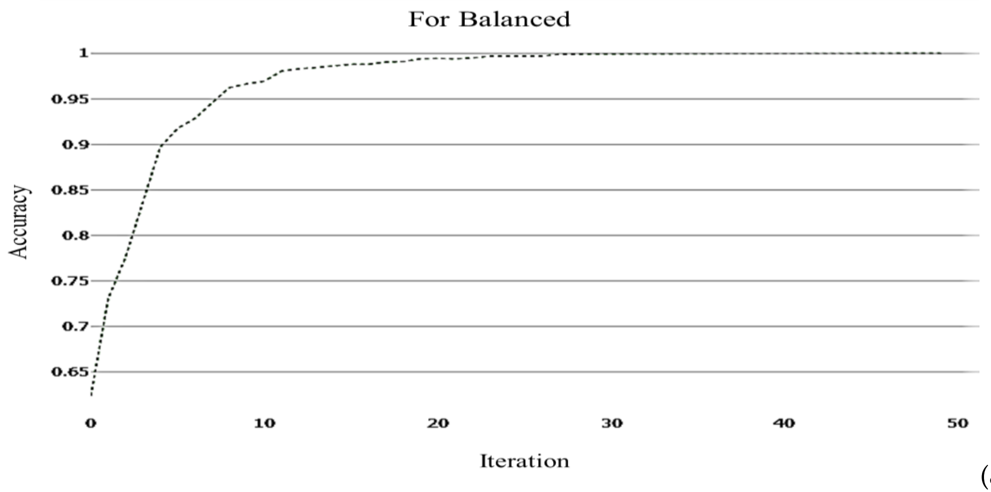
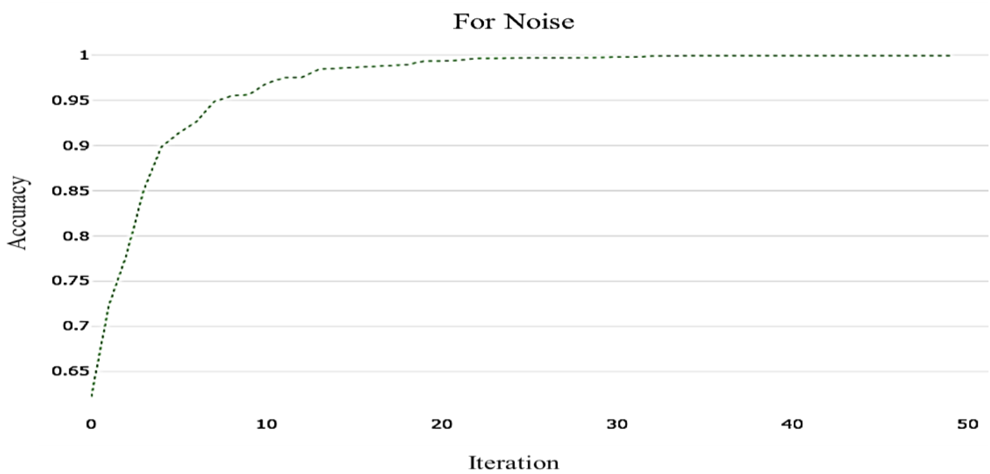


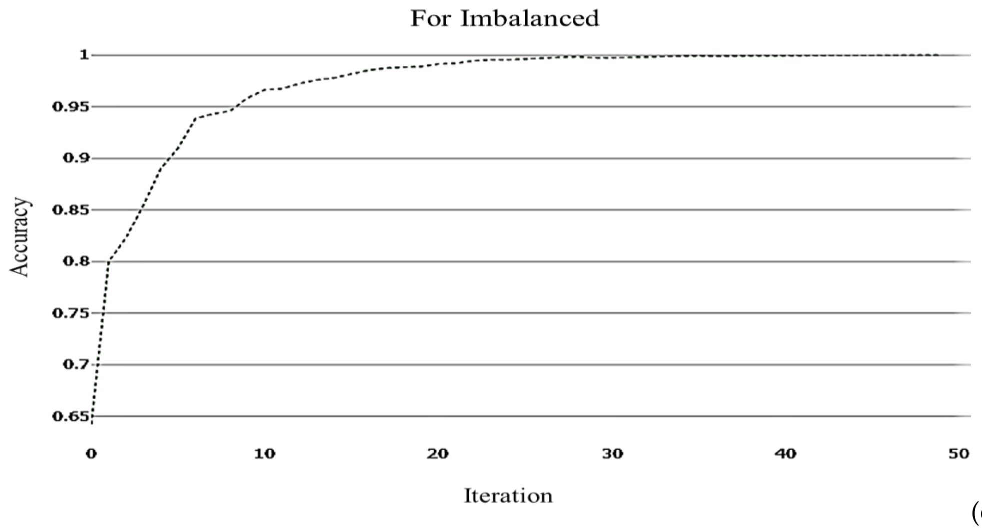
Figure 4.2: Clasification confusion matrix a) for balanced condition (**b**) for noisy condition (c) for imbalanced condition.



(**a**)

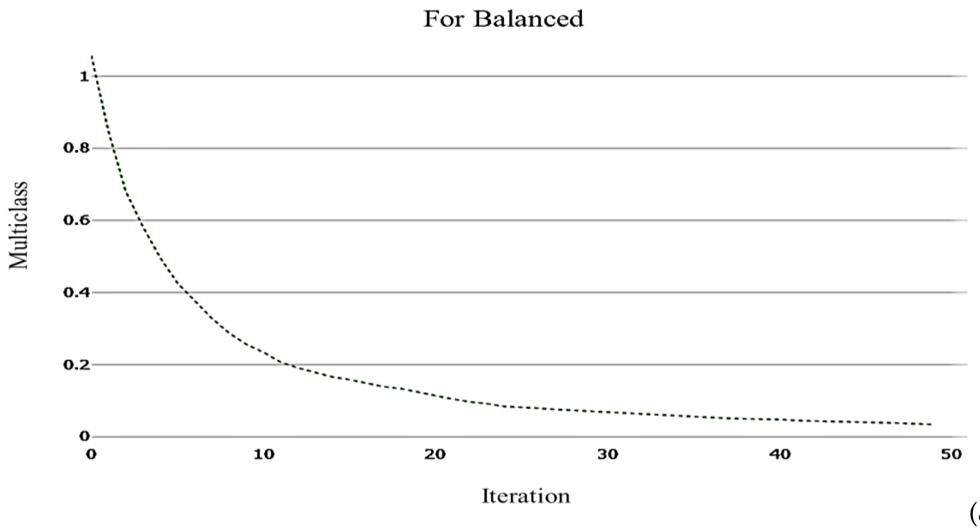


(**b**)

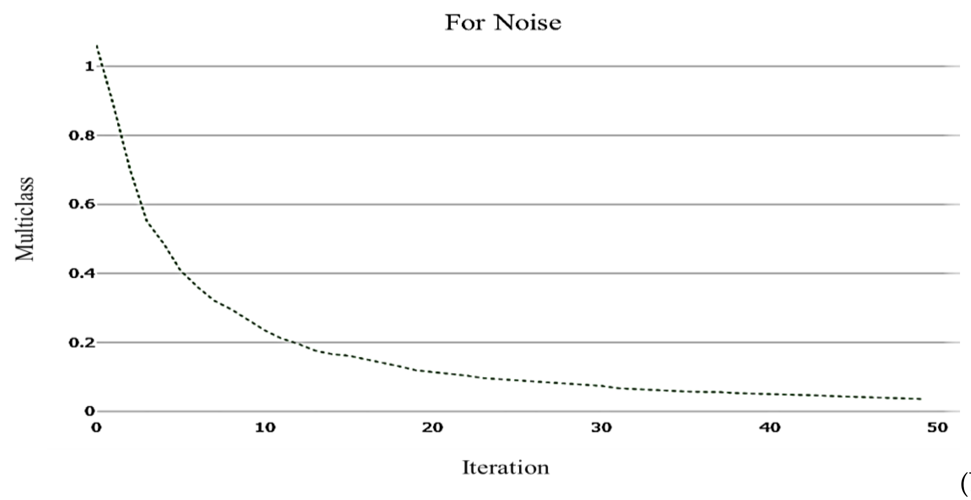


(**c**)

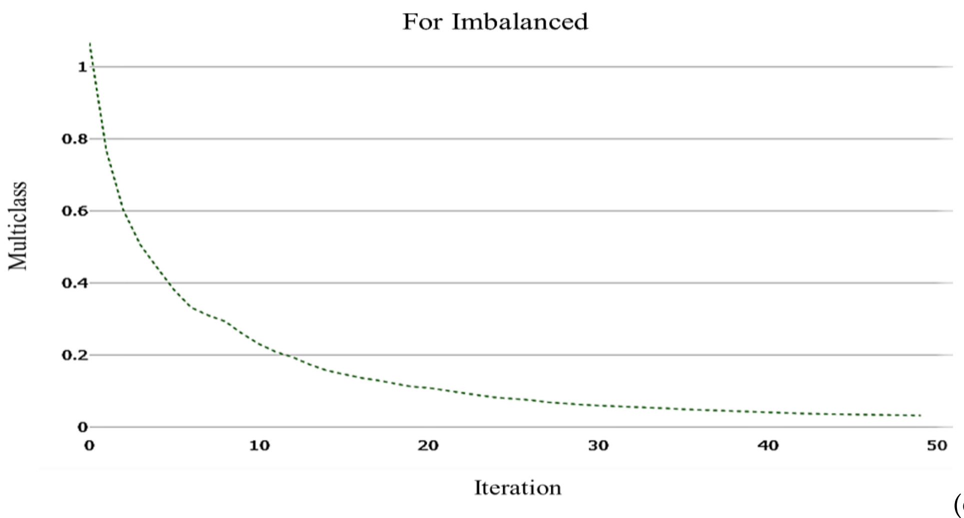
Figure 4.3 Classification accuracy: (**a**) In balanced condition; (**b**) In noisy condition; (**b**) In imbalanced condition.



(**a**)



(**b**)



(**c**)

Figure 4.4 Classification loss: (**a**) In balanced condition; (**b**) In noisy condition; (**b**) In imbalanced condition.

4.2 Discussion

It is important to evaluate the method’s performance briefly. Table 3 discusses the method's precision, recall, F1, and accuracy score for detection and classification tasks. It shows the method is well-suited. Simply, precision is the fraction of correct positive examples among all positive examples and is expressed by,

$$\text{Precision} = \frac{T_p}{F_p + T_p}, \tag{1}$$

Further, recall measures the ability to identify positive examples and is expressed by,

$$\text{Recall} = \frac{T_p}{T_p + F_n}, \tag{2}$$

And the F1 score is the harmonic mean of recall and precision and is expressed by,

$$\text{F1} = 2 * \frac{\text{Recall x Precision}}{\text{Recall + Precision}}, \tag{3}$$

Finally, accuracy measures the effectiveness of a model as a whole and is expressed by,

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + Fp + T_n + F_n}, \tag{4}$$

where, Tp= True positive, Tn= true negative Fp= false positive, and Fn= false negative [22].

Table 4.1. Performance measurement metrics.

Task	Data type	Precision	Recall	F1	Accuracy
Detection	Balanced	1	1	1	1
	Noise	1	1	1	1
	Imbalanced	1	1	1	1
Classification	Balanced	0.99	0.99	0.99	0.99
	Noise	0.99	0.99	0.99	0.99
	Imbalanced	0.99	0.99	0.99	0.99

Chapter 5: Conclusion

Quality power is that of an uninterrupted supply of power with accurate parameters. The transmission line is an essential element that ensures power supply between the generating stations and consumers. If any fault occurs, the quality of power is hampered. This paper presents an adaptive CatBoost learning method built on the gradient-boosted decision trees for fault detection and classification of the transmission line. The suggested method shows extraordinary performance towards noisy, balanced, and imbalanced data conditions. The method also simplifies the data pre-processing tasks by eradicating the need for conversion of non-numeric features to numeric ones. It is also resilient to various operating conditions of the transmission line. Finally, the method can be improved and optimized by testing on real-world applications.

References

1. Bangladesh Power Development Board, Annual Report 2020-21, p.09. Available online: http://bd.bpdb.gov.bd/bpdb/new_annual_reports.

2. Hossain, E. Power Cuts Make Households, Industries Suffer in Bangladesh. *New age*. March 11, 2020. Available online: https://www.newagebd.net/article/101874/power-cuts-make-households-industries-suffer-in-bangladesh.

3. Fairley, P. Bangladesh scrambles to deliver electricity to its 160 million residents in 2021. *IEEE Spectrum*. Jan 29, 2020. Available online: https://spectrum.ieee.org/bangladesh-scrambles-to-deliver-electricity-to-its-160-million-residents-in-2021.

4. Bangladesh Power Development Board. Annual Report 2019-20, 2020-21, p.17, p.18. Available online: http://bd.bpdb.gov.bd/bpdb/new_annual_reports.

5. S.A. Aleem, N. Shahid, I.H. Naqvi, Methodologies in power systems fault detection and diagnosis. Energy Syst. 6(1), 85–108(2014).

6. A. Prasad, J.B. Edward, K. Ravi, A review on fault classification methodologies in power transmission systems: Part—I. J. Elect. Syst. Inf. Technol. 5(1), 48–60 (2018).

7. A. Prasad, J.B. Edward, K. Ravi, A review on fault classification methodologies in power transmission systems: Part-II. J. Elect .Syst. Inf. Technol. 5(1), 61–67 (2018).

8. Mir Mohammad Taheri, Heresh Seyedi, Behnam Mohammadi-ivatloo, "DT-based relaying scheme for fault classification in transmission lines using MODP", Generation Transmission & Distribution IET, vol. 11, no.11, pp. 2796-2804, 2017.

9. Hossain, E.; Khan, I.; Un-Noor, F.; Sikander, S.S.; Sunny, Md.S.H. Application of Big Data and Machine Learning in Smart Grid, and Associated Security Concerns: A Review. *IEEE Access* **2019**, *7*, 13960–13988.

10. Jamehbozorg, A.; Shahrtash, S.M. A Decision-Tree-Based Method for Fault Classification in Single-Circuit Transmission Lines. *IEEE Trans. Power Delivery* **2010**, *25*, 2190–2196.

11. Hassan, J.U.; Nizami, I.F. Machine Learning Algorithm Analysis for Detecting and Classification Faults in Power Transmission System. In Proceedings of the 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2); IEEE: Rawalpindi, Pakistan, May 24, 2022; pp. 1–5.

12. Zhengyou, H.; Shibin, G.; Xiaoqin, C.; Jun, Z.; Zhiqian, B.; Qingquan, Q. Study of a New Method for Power System Transients Classification Based on Wavelet Entropy and Neural Network. *International Journal of Electrical Power & Energy Systems* **2011**, 33, 402–410.

13. Mukherjee, A.; Kundu, P.K.; Das, A. Transmission Line Faults in Power System and the Different Algorithms for Identification, Classification and Localization: A Brief Review of Methods. *Journal of The Institution of Engineers (India): Series B* **2021**, 102, 855–877.

14. Jamil, M.; Sharma, S.K.; Singh, R. Fault Detection and Classification in Electrical Power Transmission System Using Artificial Neural Network. *SpringerPlus* **2015**, *4*, 334.

15. Leh, N.A.M.; Zain, F.M.; Muhammad, Z.; Hamid, S.A.; Rosli, A.D. Fault Detection Method Using ANN for Power Transmission Line. In Proceedings of the 2020 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE); IEEE: Penang, Malaysia, August 2020; pp. 79–84.

16. Ferreira, V.H.; Zanghi, R.; Fortes, M.Z.; Gomes, S.; Alves da Silva, A.P. Probabilistic Transmission Line Fault Diagnosis Using Autonomous Neural Models. *Electric Power Systems Research* **2020**, 185, 106360.

17. Abdullah, A. Ultrafast Transmission Line Fault Detection Using a DWT-Based ANN. *IEEE Trans. on Ind. Applicat.* **2018**, *54*, pp. 1182–1193.

18. Fahim, S.R.; Sarker, Md.R.I.; Arifuzzaman, Md.; Hosen, Md.S.; Sarker, S.K.; Das, S.K. A Novel Approach to Fault Diagnosis of High Voltage Transmission Line - A Self Attentive Convolutional Neural Network Model. In Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP); IEEE: Dhaka, Bangladesh, 2020; pp. 1329–1332.

19. Fan, D.; Yi, Y.; Chen, S.; Song, L.; Li, J. Transmission Line Fault Type Identification Based on Transfer Learning and Deep Learning. In Proceedings of the 2021 6th International Conference on Power and Renewable Energy (ICPRE); IEEE: Shanghai, China, September 17 2021; pp. 202–207.

20. Zhu, Y.; Liu, Y.; Wu, K. Fault Phase Identification Method Based on Convolutional Neural Network for Double Circuit Transmission Lines. In Proceedings of the 2019 IEEE 8th International Conference on Advanced Power System Automation and Protection (APAP); IEEE: Xi'an, China, October 21 2019; pp. 668–672.

21. Shiddieqy, H.A.; Hariadi, F.I.; Adiono, T. Effect of Sampling Variation in Accuracy for Fault Transmission Line Classification Application Based On Convolutional Neural Network. In Proceedings of the 2018 International Symposium on Electronics and Smart Devices (ISESD); IEEE: Bandung, October 2018; pp. 1–3.

22. Cremer, C.Z. Deep Limitations? Examining Expert Disagreement over Deep Learning. *Progress in Artificial Intelligence* **2021**, *10*, 449–464.

23. Tanha, J.T.; Abdi, Y.A.; Samadi, N.S.; Razzaghi, N.R.; Asadpour, M.A. Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data* **2020**, 7.

24. Gormez, Y.; Aydin, Z.; Karademir, R.; Gungor, V.C. A Deep Learning Approach with Bayesian Optimization and Ensemble Classifiers for Detecting Denial of Service Attacks. *International Journal of Communication Systems* **2020**, *33*, e4401.

25. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient Boosting with Categorical Features Support. *arXiv preprint arXiv:1810.11363* **2018**.

26. Gulin, A.G CatBoost – open-source gradient boosting library. Available online: https://catboost.ai (accessed on 20 September 2022).

27. Sokolova, M.; Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management* **2009**, *45*, 427–437.

Appendix

Code for Transmission line fault detection

Import necessary library

```
import numpy as np
import pandas as pd
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, classification_report
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
import seaborn as sns
from catboost import CatBoostClassifier
```

Read dataset

For balanced

```
df = pd.read_csv('Detect-Data-Without-Noise-Edited.csv')
```

For imbalanced

```
df = pd.read_csv('Detect-Unbalance-Without-Noise.csv')
```

For noise

```
df = pd.read_csv('Detect-Data-With-Noise-Edited.csv')
```

Select input and output data

Input

```
X = df.drop(['OUTPUT'],axis=1)
```

Output

```
y = df['OUTPUT']
```

Features declaration

```
cat_features = list(range(0, X.shape[1]))
print(cat_features)
```

Split data into train and test

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20, random_state = 0)
```

CatBoost implementation

```
from catboost import CatBoostClassifier

clf = CatBoostClassifier(
    iterations=5,
    learning_rate=0.1,
    #loss_function='CrossEntropy'
)


clf.fit(X_train, y_train,

)

print('CatBoost model is fitted: ' + str(clf.is_fitted()))
print('CatBoost model parameters:')
print(clf.get_params())
```

Stdout of the training

```
from catboost import CatBoostClassifier

clf = CatBoostClassifier(
    iterations=50,
    random_seed=42,
    learning_rate=0.5,
    custom_loss=['AUC', 'Accuracy']
)

clf.fit(
    X_train, y_train,


)
```

Model predictions

```
y_pred=print(clf.predict(data=X_test))
```

Metrics calculation and graph plotting

```
clf.fit(X_train,y_train,

        verbose=False ,

          plot=True)
```

Confusion matrix

```
cf_matrix = confusion_matrix(y_test, y_pred)

ax = sns.heatmap(cf_matrix, annot=True, cmap='Greens')

ax.set_title(' Confusion Matrix \n\n');
ax.set_xlabel('\nPredicted Values')
ax.set_ylabel('Actual Values ');
```

Code for Transmission line fault classification

Import necessary library

```
import numpy as np
import pandas as pd
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, classification_report
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
import seaborn as sns
from catboost import CatBoostClassifier
```

Read dataset

For balanced

```
multi_data= pd.read_csv('test 1.csv')
```

For imbalanced

```
multi_data = pd.read_csv('Classification-Unbalance-Without-Noise.csv')
```

For noise

```
multi_data= pd.read_csv('Train-Data-With-Noise.csv')
```

Select input and output data

Input

```
X = multi_data.drop(['A','B','C','G','faultType'], axis=1)
```

Output

```
y = multi_data['faultType']
```

Features declaration

```
cat_features = list(range(0, X.shape[1]))
print(cat_features)
```

Split data into train and test

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20, random_state = 0)
```

Catboost implementation

```
from catboost import CatBoostClassifier

clf = CatBoostClassifier(
    iterations=5,
    learning_rate=0.1,
    #loss_function='CrossEntropy'
)


clf.fit(X_train, y_train,

)

print('CatBoost model is fitted: ' + str(clf.is_fitted()))
print('CatBoost model parameters:')
print(clf.get_params())
```

Stdout of the training

```
from catboost import CatBoostClassifier

clf = CatBoostClassifier(
    iterations=50,
    random_seed=42,
    learning_rate=0.5,
    custom_loss=['AUC', 'Accuracy']
)

clf.fit(
    X_train, y_train,


)
```

Model predictions

```
y_pred=print(clf.predict(data=X_test))
```

Metrics calculation and graph plotting

```
clf.fit(X_train,y_train,

        verbose=False ,

          plot=True)
```

Confusion matrix

```
cf_matrix = confusion_matrix(y_test, y_pred)

ax = sns.heatmap(cf_matrix, annot=True, cmap='Greens')

ax.set_title(' Confusion Matrix \n\n');
ax.set_xlabel('\nPredicted Values')
ax.set_ylabel('Actual Values ');
```