Fall 2022

# Predicting Electrical Faults in Power Distribution Network

Ahmad Adel Bin Sulaiman
aab9033@rit.edu

# Predicting Electrical Faults in Power Distribution Network

by

**Ahmad Adel Bin Sulaiman**

**A Capstone Submitted in Partial Fulfilment of the Requirements for the**

**Degree of Master of Science in Professional Studies:**

**Data Analytics**

**Department of Graduate Programs & Research**

**Rochester Institute of Technology**

**RIT Dubai**

**Fall 2022**

# RIT

**Master of Science in Professional Studies:**

**Data Analytics**

**Graduate Capstone Approval**

Student Name**: Ahmad Adel Bin Sulaiman**

Graduate Capstone Title**: Predicting Electrical Faults in Power Distribution Network**

**Graduate Capstone Committee:**

**Name: Dr. Sanjay Modak** **Date:**

**Chair of committee**

---

**Name: Dr. Ehsan Warriach** **Date:**

**Member of committee/Mentor**

---

# Acknowledgments

First and foremost, I would like to praise Allah, all praises to Allah, the Most Compassionate, the Most Merciful, Glory to Allah.

My mentor and advisor, Dr. Ehsan Warriach, deserves special recognition and gratitude for his unwavering assistance, collaboration, and guidance during the entire capstone project. Additionally, I would like to express my gratitude to Dr. Sanjay Modak, Chair of Graduate Programs and Research at RIT Dubai, for his suggestions, direction, and criticism during the entire program.

In addition, I want to give a special gratitude to my parents, who have always supported me, especially during the master's program. I would also like to thank my family, friends, and coworkers for their unwavering support during the research.

# Abstract

Electricity is becoming increasingly important in modern civilization, and as a result, the emphasis on and use of power infrastructure is gradually expanding. Simultaneously, investment and distribution modes are shifting from the large-scale centralized generation of electricity and sheer consumption to decentralized generators and extremely sophisticated clients. This transformation puts further strain on old infrastructure, necessitating significant expenditures in future years to ensure a consistent supply. Subsequent technical and prediction technologies can help to maximize the use of the current grid while lowering the probability of faults. This study discusses some of the local grid difficulties as well as a prospective maintenance and failure probabilistic model. To provide an effective and convenient power source to consumers, a high Volta protects and maintains under fault conditions. Most of the fault identification and localization approaches rely on real and reactive power converter observations of electronic values. This can be seen in metrics and ground evaluations derived via internet traffic. This paper provides a thorough examination of the mechanisms for error detection, diagnosis, and localization in overhead lines. The proposal is then able to make suggestions about the ways that can be incorporated to predict foreseen faults in the electrical network. The three classifiers, Random Forest, XGBoost and Decision tree are producing high accuracies, while Logistic Regression and SVM are producing realistic accuracy results.

**Keywords:** Distribution Network, Electrical Fault, Predictive Maintenance, Power, Machine Learning

# Table of Contents

# List of Figures



# List of Tables

# Chapter 1 Introduction

## 1.1 Background Information

The transmission line is the backbone of the electrical grid. Power and its reliability have become increasingly important in the modern era, and a transmission line's primary function is to move electricity from its generation facility to its distribution system. As a result of its dynamic and interconnected nature, the electrical power system is constantly vulnerable to disruption or malfunction. The interconnected nature of the electrical power system results in a vulnerability to disruption or malfunction. The electrical power system is vulnerable to disruption or malfunction.

Identifying where a problem exists in the electrical grid is an important responsibility for those who work in the industry. Because of the potential for damage to network devices, disruptions in service, and overall network instability, the fault has the potential to decrease network reliability and affect the ability of network devices to function properly. Therefore, identifying and locating faults within the power grid is essential to maintaining the reliable operation of the electrical grid. The utility industry relies on information from a variety of sources, including field workers and power plant operators, to identify system-wide problems such as faults in transmission lines, substations, or other network devices within an electric utility's grid (Eskandarpour & Khodaei, 2017).

Customers and power corporations both suffer monetary losses as a result. When a distribution network spans a large geographical area, standard fault location methods in the distributed network are ineffective. There is a high cost in time and resources needed to cover a large area with redundant distribution networks. The location can be determined by the node with the best quality of service, by the most resources, or a combination thereof. One method of improving reliability is to enlarge and duplicate the size of power stations to have more capacity (Viegas et al., 2016). This

increases costs, but also increases reliability due to redundancy in terms of power generation and distribution networks must have the ability to automatically detect and locate faults. Time, labor, system readiness for power maintenance, future scheduling flexibility, and economic variables are just a few of the benefits of automatic fault prediction and localization. Customer satisfaction and system dependability both rise because of these changes. The operation of a distribution system is based on the premise that a fault will always be present and can be found through routine monitoring. If the system does not function properly, it could lead to both safety and financial problems. For example, if the power goes out in one location for a long time, it could lead to fire or electrocution hazards.

This paper will discuss the different methods and algorithms that can be used to predict electrical faults in distribution networks. Distribution networks are the backbone of our society. These networks transport the electricity from power plants to substations, and then through high voltage transmission lines to households and businesses. The distribution networks are composed of many cables that run underground or over the ground to multiple nodes where electricity is distributed. Electrical faults in these networks can lead to blackouts and even fire hazards. As these networks become ever more complex, the chance of a fault occurring becomes greater and greater. The consequences of a network fault can be disastrous, with power outages leading to injuries or even loss of life.

The location of faults is the key to preventing power outages. This paper will discuss the different methods and algorithms that can be used to predict electrical faults in distribution networks. Distribution power networks often experience electrical faults that cause power outages. These faults can be caused by equipment malfunctions, lightning strikes, and other events. Methods used to prevent these failures include using advanced warning systems such as voltage sensors and telemetry

in the distribution network. Manufacturing fluctuations also contribute to large-scale failures on power networks. A power distribution network is a chain of interconnected electrical substations and transmission lines that carry electricity from large energy sources to end users. A power network may be operated by an electric utility, a local or regional government agency cooperative.

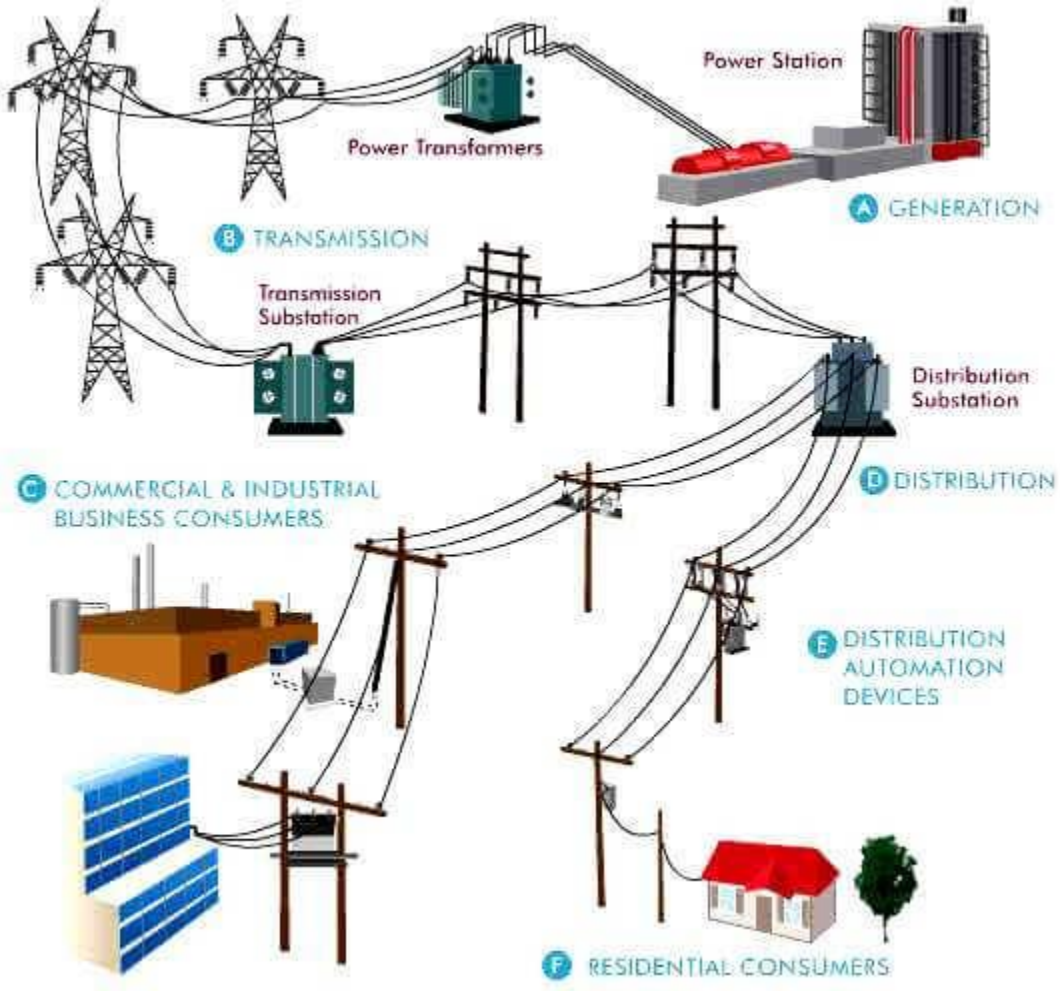


*Figure 1 Electrical Network*

Faults in energy systems lead to potentially dangerous transients, equipment failure, and power outages, all of which lower system reliability and leave customers frustrated. Fault prediction is crucial in the shift from reactive and wasteful maintenance procedures to a more proactive maintenance plan. To avoid defects and failures in energy systems, it is important to identify developing problems before they become obvious. For instance, we can replace or fix the faulty parts

that cause symptoms i.e., cable with a spark in it, a transformer with partial discharge, etc. These components are defective to the point that the entire system can be shut down by the slightest change. Predictive maintenance is the process of keeping systems and components in good working order by anticipating potential problems. (Balouji et al., 2018) Electrical distribution network fault prediction is possible with the help of machine learning technologies and model training developed using real and/or simulated data. The system's reliability is improved because of this strategy, which reduces the likelihood that the defect would occur in the first place. This requires classifying the system's actual data. The weather conditions recorded by weather stations, the locations of any breakdowns, or even the voltage currently recorded at regular intervals by the energy system's components, are all examples of the types of information that could be collected. (Skydt et al., 2021) Therefore, two general cases may exist based on the data type:

- Predicting fault in an electrical network using weather conditions and/or characteristic data of systems components.
- Predicting fault in the equipment of the network such as transformer, cable, etc. using periodically recorded voltage and current. The model assumes that when measured values are low or zero and model values are high, then the fault is most likely to occur. In this case, it is assumed that data labels were applied by the process designer before creating a model.

Previous generating systems that used massive revolving masses, like as enormous power plants, are rapidly getting supplanted by battery technology which precisely manages grid frequency ratios because of these trends. Such expansion reduces the total electricity state's resilience to errors and shocks. Sliding failures on a platform base would become more important in the future even without stabilizing the aspect of spinning forces. It is critical to address such issues as part of that process toward a more responsive Power System (Andresen et al., 2018).

## 1.2 Introduction

A steady supply of power is becoming increasingly important to the functioning of modern society. The repercussions of a power outage can range from being a little annoyance to causing large financial losses or even posing serious risks to the health and safety of local residents (BUTLER, 2001).

The complexity of the systems that distribute power makes it more likely that there will be frequent faults (OFGEM, 2017). The leading equipment in the network, in particular, is always susceptible to multiple failures, which can take place in any of the leading equipment's primary components or subcomponents. These failures can take place at any time. If there is a problem with one of the network's components, the power will go out not just in the region that is being served by those components, but also in the areas that are adjacent to those components. If there is a problem anywhere in the system that distributes energy, it will produce a severe disturbance throughout the entire grid. When something goes wrong with the system, there is potentially a large risk of incurring significant financial consequences. To avoid incurring the monetary penalties that the authorities threaten them with, it is reasonable for electrical businesses to prevent any delays in the provision of electricity and to quickly regain their customers' confidence in the event of a breakdown. When a fault occurs, determining where in the distribution systems the issue can be found is of the utmost importance. It is very vital to be able to predict problems in distribution networks particularly along with their locations.

### 1.2.1 Statement of Problem

Due to increased load on the network, electrical faults are a common issue that causes electrical interruption to consumers, usually up to two hours of interruption. It is hard to check the path of the

faults manually to judge whether which part has the defect and which connection must manage properly. Predicting electrical faults in distribution networks using online data helps in avoiding long interruptions for the consumers which could increase satisfaction.

### 1.2.2 Problem Question

Failures in the energy distribution network should be thoroughly understood by both the operators and the customers of the network. This is of the utmost significance. On the other hand, due to the unpredictability of the problems, it can be difficult to provide accurate failure predictions for a certain time.

This study attempts to answer the following primary research question: “Use of machine learning techniques can provide the accurate prediction and forecasting of defects in distributing networks"

## 1.3 Project Definition and Goals

To predict electrical issues in the distribution network to reduce electrical interruption. The key goals of the project are:

- To use preceding research to explore the edge.
- To determine the types of faults and their causes.
- To lessen usage of manpower.
- To suggest areas of conflict having electric faults.
- To build a classification model that predicts an electrical fault in the network before it occurs

## 1.4 Research Methodology

The suggested methodology is supported by a schema. The very first stage is unquestionably the most crucial since it makes it possible to identify the distribution system and any potential defects using information from linked devices and appliances.

- **Step 01:** Source dataset is collected from Kaggle which contains total of 12001 instances with 7 attribute value pairs.
- **Step 02:** For these analysis and design processes, appropriate dataset will be explored and visualized using R programming language working in RStudio IDE.
- **Step 03:** For evaluation, dataset will be split into standard 70/30 split ratio as train-test subset datasets.
- **Step 04:** Dataset will be trained and tested using 5 different Machine Learning Models.
    - Logistic Regression
    - SVM Classifier
    - Decision Tree
    - XG Boosting Classifier
    - Random Forest
- **Step 05:** Obtained results will be evaluated using the Accuracy as evaluation measure.

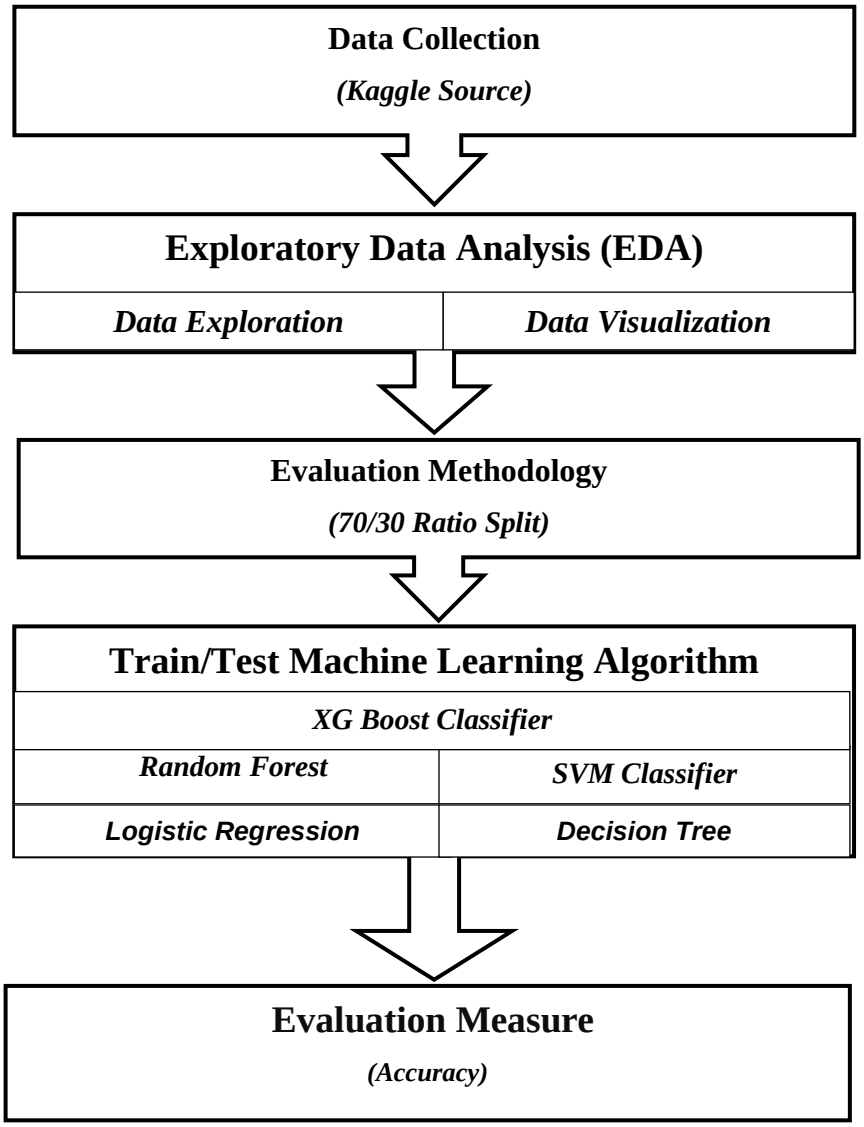


*Figure 2 Methodology*

## 1.5 Limitations of the Study

This study was limited based on the availability of the data set, service providers usually have confidentiality that prevent them from live information from smart electric meters; voltage and current. Therefore, a MATLAB-simulated data set was collected from Kaggle. The study will be limited on a set of data that is simulated by a computer software that may not reflect in real life.

# Chapter 2 Literature Review

## 2.1 History

Over the years, approaches for detecting, classifying, and locating faults in transmission lines and distribution systems have been extensively investigated. With power system notions causing academics to become increasingly concerned, the need for training a deep fault tracking and definite form capable of identifying and finding various sorts of problems cannot be stressed. In this paper, we first provide an overview of the most used fault detection and location mechanisms in practice, then we evaluate their accuracy with a case study. Summary of basic fault detection and locating methods, Basic approaches for fault detection and locating include various means to detect failures in the system.

The diagnosis, categorization, and placement of defects in power sources have advanced rapidly over the last Twenty years in a variety of disciplines. Recent advancements in methods, cognitive computing, location-based services, and communication channels have empowered a growing number of scientists to support research with a wide scope and depth, allowing traditional fault defense methodologies to be pushed to their limits. In addition, two main limitations of web defect diagnostics are indeed being addressed. The first limitation is the delay in receiving data. Recently designed sophisticated electrical gadgets are really being implemented (Wang et al., 2011) to acquire data at various nodes in the systems, in complement to classic test rigs like voltage transformers, current transformers, and peripherals. The second limitation is the cost of web services. For example, when a device is not directly connected to an external network and data are being acquired via radio waves, the accuracy is lessened by the positioning errors in geographical space. From a learning

perspective, I would say that this field has evolved rapidly over time. It's important to understand where your strengths and weaknesses are and to work on improving those areas.

Self-powered non-intrusive sensors are also being developed with the potential to form sensors Self-powered, non-intrusive sensors are now being developed, with the possibility of forming sensor nodes for power system online control (Han et al., 2015)**.** Researchers can construct intelligent fault diagnostic systems by mining knowledge from data corresponding to various conditions as more data becomes accessible. When current and voltage signals are collected by many interspersed sensors, the effect of complicated and diverse network designs can be minimized. The lack of adequate terms of computational capabilities is the next constraint. In a self-powered sensor, the power for its operation is typically generated by a renewable energy source such as solar cells. Renewable sources can produce more power on demand than non-renewable sources like fossil fuels. This ability can be utilized to construct simple, robust sensors that are not dependent on external power supplies. However, more conventional sensors are more complicated and require more power. In the case of a robotic system, the processing power is often used to control its movement. For example, mobile robots use GPS receivers to keep track of their position in space and need powerful processors to carry out complex calculations in real-time. Sensor systems may also depend on bulky batteries or fuel cells, which require powerful processors. Power outages in the smart grid must be managed and assessed quickly enough to protect framework components and maintain normal operations. Many of the new technologies that are being developed to enhance the grid will require a more efficient and resilient power supply. New technologies in this area include demand response and distributed energy.

The model reported in this study (Tyvold et al., 2020) has been proven to have high accuracy and an impressive level of reliability when used on a wide range of data samples that includes some faulty

examples. No faulty data will comprise most of the information that only an operating model will examine. For example, if the normal fault incidence had one or two each week, but the model adjusts the fault risk once each second, this is a common circumstance. As a result, if the model's number of false isn't particularly low, the number of false alarms can rapidly overwhelm the genuine ones. .- Incorrect data will comprise the minority of the information that only an operating model will examine. Likewise, if the normal fault incidence had limited numbers each week, but the model adjusted for it once every second, this is a common circumstance. As a result, if the model's number of false isn't particularly low, it can be hard to tell where there are genuine faults from false alarms. False alarms also cause some of the model's data to be noisy. This can be partially mitigated by increasing the number of model parameters that are being used and/or increasing the period in which the model is being compared to its prior data.

The report (Mahmoud et al., 2021) focused on flaws in the smart grid system from several perspectives.

The following are some of the flaws that were identified:

- Smart grid system is not able to produce electricity during power outages.
- Power consumption is higher than what was expected and predicted in terms of frequency and duration of power outages.
- The smart grid system has environmental side effects with an increased amount of CO2 emissions and nitrogen oxide.

In this study, power problems in the smart grid are discussed, including the causes of an event, fault management, types of failures, and protection. Locating, identifying, and isolating defective lines are all part of fault management research. In the transition to a smarter network grid, advancements in this study area are projected to enable a more robust and effective grid utilization. Finally, predictive

maintenance algorithms can largely profit from the expanded use of data frameworks created using big data technologies and algorithms.

The suggested technique (Picard et al., 2021) improves the traditional distribution network planning process by incorporating DER from a network perspective and utilizing the data made accessible by smart grid deployment. Furthermore, the proposed technique was successfully evaluated in a real distribution network with 450,000 delivery points over a 5000 km2 geographical area, identifying zones where new DER will be more effective. It was determined which areas in the distribution system will benefit most from the new DER and which zones are potential bottlenecks. Through this, a roadmap was created to guide the further deployment of DER. The key outcome was the establishment of a method for improving present network knowledge by breaking it down into subsystems and planning with specific base cases for each region and their related needs operation, type of consumption, and generation. This novel approach to planning was proven in terms of economics in a regulated distribution framework, taking into account new network infrastructure. , including the creation of a series of requirements for the new network, such as security and component redundancy to ensure reliability. The study's contribution is shown through its ability to identify key areas for improvement to position itself as the global reference for smart energy networks.

This study (Andresen et al., 2018) discusses the anticipated obstacles in the utility grid, as well as various metering and activities that could aid the power grid. The increasing availability of long-term monitoring and control data will help reduce data latency, along with developments in big data analysis and machine learning, allowing for efficient failure classification and prevention. In the shift to a smarter grid, advancements in this study area are projected to enable a more resilient and economical grid utilization.

The climate machine-learning techniques created in this article for predicting voltage instability have been proven to outperform games of chance or random speculation. Given those models' restricted power, the gain proved significant across several cases: the median reliability for the highest performing designs for each period and activity genre ranged between +0:06 to +0:26 greater than that of the benchmark value. "This study finds that grouping and density segmentation techniques can be used to anticipate electric grid failure incidents. The techniques have low predictability, but they may be used in conjunction with classifiers, for example." The research also shows how addressing imbalanced instead of equitable statistics reduces accurate predictions, which is essential because consecutive actual data is imbalanced.

In this paper (Høiem et al., 2020) the author did a comparison of defect prediction models based on several supervised machine learning algorithms was undertaken in this paper.

Norway is a beautiful country, known for its amazing landscapes, long coastlines, and rich cultural and natural history. These majestic qualities attracted power quality experts from IBV. With the help of automated data acquisition equipment provided by our close partner Rockwell Automation™*, IBV was able to employ high-resolution power quality measurements in Norway to train the machine learning models used in its PQD.IBV's PQD suppliers internally use a model based on machine learning technology to predict and identify power quality problems in the grid.

The energy grid is vulnerable to various errors, each with its own characteristics and dispersion pattern. One of these types of errors is predictable, so identifying it can help mitigate the issue. This hallmark must be present before the crucial tragedies. That some of the most common issue incidents in the main supply are thought to grow over time but have a defect generation hallmark before the disruption, as seen (Russell et al., 2009)**.** An isolator defect that leads to an earth fault or capacitance, for instance, might also cause releases which could be discovered with measurement tools, but the

releases could be so small that they escape detection. It is important to note that the error process can be detected with measurement tools, but the releases might be so small that they escape detection. In a comprehensive schematic of this kind, it is imperative to have engineers analyze propagation patterns because each different type of defect and its various frequencies correspond to different causes of the damage. Another aspect of root cause analysis is determining whether an environmental factor or a defect contributed to a particular failure. If a part cracked but no defects were found on the part, then it could be attributed to temperature cycling.

## 2.2 Machine Learning Based Solution

## 2.3 Supervised Learning

Learning via example is at the heart of supervised machine learning (Gorunescu 2011), which is accomplished using data that has been labelled and previously categorized. The term "supervised learning" refers to a sort of task in which an algorithm is trained based on past knowledge. Supervised learning is also known as "supervised machine learning." This means that the labelling of the outcome, which is also known as the response variable, dependent variable, or target variable, must be predicted in advance for this type of learning. The assumption that learning is directed by the learning results of the training data is where the supervision originates from. The learning algorithm is given a set of inputs that correspond to preset correct outputs; from these, it learns the pattern and adapts itself appropriately. In this study, the methods of classification and regression, both of which fall under the umbrella of supervised learning, will be employed to make predictions (G. Lei wang,s 2011).

## 2.4 Decision Tree

A decision tree is a type of supervised classification algorithm that draws inferences from previous data to correctly categorize new data. The algorithm learns from previous data and then applies those learnings to classify new data (Hongquan 2019).

Random Forest which is an extension of Decision Trees that works by building an ensemble of Decision Trees on various subsets of the complete dataset and then aggregating the outputs of these Decision Trees to obtain a more accurate prediction model. The Decision Tree is a technique for solving classification issues that is used in machine learning. This technique works by inferring rules for partitioning the dataset into numerous subsets based on the characteristics of the data.

## 2.5 Support Vector Machine

Support Vector Machines, or SVM for short, are a form of machine learning that have seen extensive use in the field of classification problems. It accomplishes this by locating, within a space consisting of many dimensions, the hyperplane that best differentiates the specified categories by accounting for the greatest distance between itself and the nearest sample (Andrew, 2011).

## 2.6 Models Comparison

A comparison of defect prediction models that are based on several supervised machine learning approaches has been carried out in this piece of research. To train the machine learning models, high-resolution power quality measurements were taken from the Norwegian power system. The predictive models were trained to forecast four distinct event categories, which are referred to as voltage dips, ground faults, rapid voltage changes (RVC), and interruptions, respectively. The outcomes of the comparison between the Random Forest model, the Support Vector Machine model,

the Feed Forward Neural Network model, and the Recurrent Neural Network model indicate that the Random Forest model is performing marginally better than the other models, with an accuracy of 0.602. However, the performance is not yet at a level where it would be possible to implement the technology in an operating setting. The rate of false positives as well as the rate of false negatives are too high for such a step to be taken. It has also been noticed that the models' performances vary depending on the different kinds of events that they are trained on. This is another thing that may be examined. According to the findings, the PQ events that are simplest to forecast are rapid voltage shifts, which have an accuracy of 0.710, and voltage dips, which have an accuracy of 0.601.

## 2.7 Summary of Literature Review

Failure detection and prevention strategies in the utility grid and various metering activities could help adequately control and manage the grid system. Long-term monitoring and data control strategies help reduce data latency in the systems. The energy grid system is vulnerable to many errors that have different characteristics. The measurement tools could discover and prevent significant defects in the systems. The only challenge is that some of the errors in the grid system might be smaller to the point of escaping detection. This means that it is essential to have engineers who would analyze the propagation patterns and detect the releases that escape detection. The aspect of cause in defect detection and prevention varies, and defects might correspond to different causes and damages. This means it is important to analyze a defect to establish if it causes system failures.

- Machine learning solutions in defect detention rely on massive amounts of data which could be problematic for other predictive methods
- Data mining methods become an efficient method that is supported by complex algorithms that could be used in the detection of faults through the processing of enormous amounts of data

- Data gathering devices could also help in error detection in the transmission line networks
- Data gathering method employs the use of complex machine-learning algorithms
- Supervised learning is a system of learning by example.
- Supervised learning works through predictive variables or target variables that must have been labelled based on past incidents
- A decision tree is a method in a smart grid system that utilizes previous data in the generation of new data
- Support vector machine works by locating within a vast space a hyperplane that best suits a specific category and differentiates it from the nearest sample.
- The predictive models in smart grid system focuses on four distinct events; voltage dips, ground faults, rapid voltage change and interruptions
- A comparison between various models shows that random forest model is performing better than other models, with accuracy of 0.602.
- The performance of the random forest model has however not reached the projected levels.

# Chapter 3 Project Description

## 3.1 Project Overview

To ultimately attain the fundamental goal of creating numerous machine learning classification algorithms that can categorize electrical faults, the project will go through several processes. The project will begin with exploratory data analysis (EDA) and data visualization for the chosen dataset after it has been chosen. Before moving on to the data modelling stage, additional manual data pre-processing activities will be carried out to clean the data, cope with empty cells, and attempt to enhance the accuracy of the dataset if required. The algorithms will be tested using the appropriate metrics to judge their performance after being built using the clean data set. Additionally, all work involving data will be carried out using the RStudio software.

## 3.2 Dataset Description

The data set was gathered from the Kaggle website, a well-known platform and online community for practitioners of data science and machine learning (ML), which offers a wide range of data science and ML issues. The problem confronting is “Detecting and Classifying Electrical Faults using Machine Learning Algorithms”, and in this project an open-source dataset is used for analysis and implementation.

The downloaded dataset folder contains two csv files. Our concerned file is named “detect_dataset.csv” containing total of 12001 rows and 7 attribute value pairs with unique values.

*Table 1 Dataset Description*

| Feature | Type | Details |
|---|---|---|
| Output (S) | Boolean (0 or 1) | 0: there is no fault<br>1: there is a fault |
| Ia | integer | Electric Current of phase ‘A’ |
| Ib | integer | Electric Current of phase ‘B’ |
| Ic | integer | Electric Current of phase ‘C’ |
| Va | integer | Voltage of phase ‘A’ |
| Vb | integer | Voltage of phase ‘B’ |
| Vc | integer | Voltage of phase ‘C’ |

**Source of Dataset:**

Dataset is available in this link below

https://www.kaggle.com/datasets/esathyaprakash/electrical-fault-detection-and-classification

# Chapter 4 Data Analysis

## 4.1 Data Understanding

The working data set must first be understood and studied; this is a crucial initial step in trying to draw attention to some of the properties and characteristics of our data collection. The creation of a word cloud, evaluating attitudes, and visualizing attributes to uncover more information about them, such as keyword and location properties, are additional aspects of exploratory data analysis (EDA). For the EDA process, we used R programming language and to support R language used RStudio software.

## 4.2 Dataset Exploration

In RStudio, initially, we installed the necessary packages for our required EDA procedure sequence. The dataset was explored into different phases and steps for better understanding and clarity of each phase. Each package contains specific metadata about visualizations, modelling, preprocessing, and other tasks needed to be done.

```{r}
install.packages('dplyr')
install.packages('tidyr')
install.packages('tidyverse')
install.packages('hrbrthemes')
install.packages('viridis')
install.packages('ggplot2')
install.packages('forecast')
install.packages('caret')
install.packages('rpart.plot')
install.packages('e1071')
install.packages('caTools')
install.packages('class')
install.packages('ISLR')
install.packages('rpart')
install.packages('DAAG')
install.packages('mlbench')
install.packages('tree')
install.packages('pROC')
install.packages('xgboost')
```

*Figure 3 Installing packages*

Then responding libraries were imported into RStudio.

```
## Importing Libraries
```{r}
library(dplyr)
library(tree)
library(DAAG)
library(mlbench)
library(e1071)
library(caTools)
library(class)
library(tidyr)
library(tidyverse)
library(hrbrthemes)
library(viridis)
library(ggplot2)
library(forcats)
library(caret)
library(rpart.plot)
library(party)
library(ISLR)
library(rpart)
library(caTools)

```
```

*Figure 4 Importing Libraries*

Then, the dataset named “detect_dataset.csv” was imported.

```
## Load Data set
```{r}
df<-read.csv("C:/Users/Administrator/Desktop/Capstone/Data/Electrical Fault detection and classification/detect_dataset.csv",header=TRUE)


```
```

*Figure 5 Loading dataset in a data frame*

To display the initial rows and columns of the dataset, we used the head() function.

```
## Load head of data set
```{r}
head(df)
```
```

Description: df [6 x 7]

| | Output..S. <int> | Ia <dbl> | Ib <dbl> | Ic <dbl> | Va <dbl> |
|---|---|---|---|---|---|
| 1 | 0 | -170.47220 | 9.2196135 | 161.25258 | 0.0544900 |
| 2 | 0 | -122.23575 | 6.1686674 | 116.06709 | 0.1020000 |
| 3 | 0 | -90.16147 | 3.8136322 | 86.34784 | 0.1410255 |
| 4 | 0 | -79.90492 | 2.3988035 | 77.50611 | 0.1562725 |
| 5 | 0 | -63.88525 | 0.5906674 | 63.29459 | 0.1804515 |
| 6 | 0 | -55.95468 | -1.0018817 | 56.95656 | 0.1934141 |

6 rows | 1-6 of 7 columns

*Figure 6 Head of the dataset*

Then, the dataset summary is estimated as the Means, Median and Mode of the complete dataset.

```
## All Summary of Data set
```{r}
summary(df)
```
```

```
  Output..S.          Ia                 Ib
Min.   :0.000   Min.   :-883.542   Min.   :-900.527
1st Qu.:0.000   1st Qu.: -64.349   1st Qu.: -51.422
Median :0.000   Median :  -3.240   Median :   4.711
Mean   :0.458   Mean   :   6.709   Mean   : -26.558
3rd Qu.:1.000   3rd Qu.:  53.824   3rd Qu.:  69.638
Max.   :1.000   Max.   : 885.739   Max.   : 889.869
      Ic                 Va                  Vb
Min.   :-883.3578   Min.   :-0.620748   Min.   :-0.659921
1st Qu.: -54.5623   1st Qu.:-0.237610   1st Qu.:-0.313721
Median :  -0.3994   Median : 0.002465   Median :-0.007192
Mean   :  22.3530   Mean   : 0.010517   Mean   :-0.015498
3rd Qu.:  45.2745   3rd Qu.: 0.285078   3rd Qu.: 0.248681
Max.   : 901.2743   Max.   : 0.609864   Max.   : 0.627875
      Vc
Min.   :-0.612709
1st Qu.:-0.278951
Median : 0.008381
Mean   : 0.004980
3rd Qu.: 0.289681
Max.   : 0.608243
```

*Figure 7 Summary of the dataset*

We retrieved the dimension of the dataset, which notifies that dataset has a total of 12001 no. of rows and a total of 7 columns.

```
## Dimension of data set
```{r}
dim(df)
```
```

```
[1] 12001     7
```

*Figure 8 Dataset dimension*

Then displayed the complete dataset in a single plot to represent the different dataset column values as an individual value set.

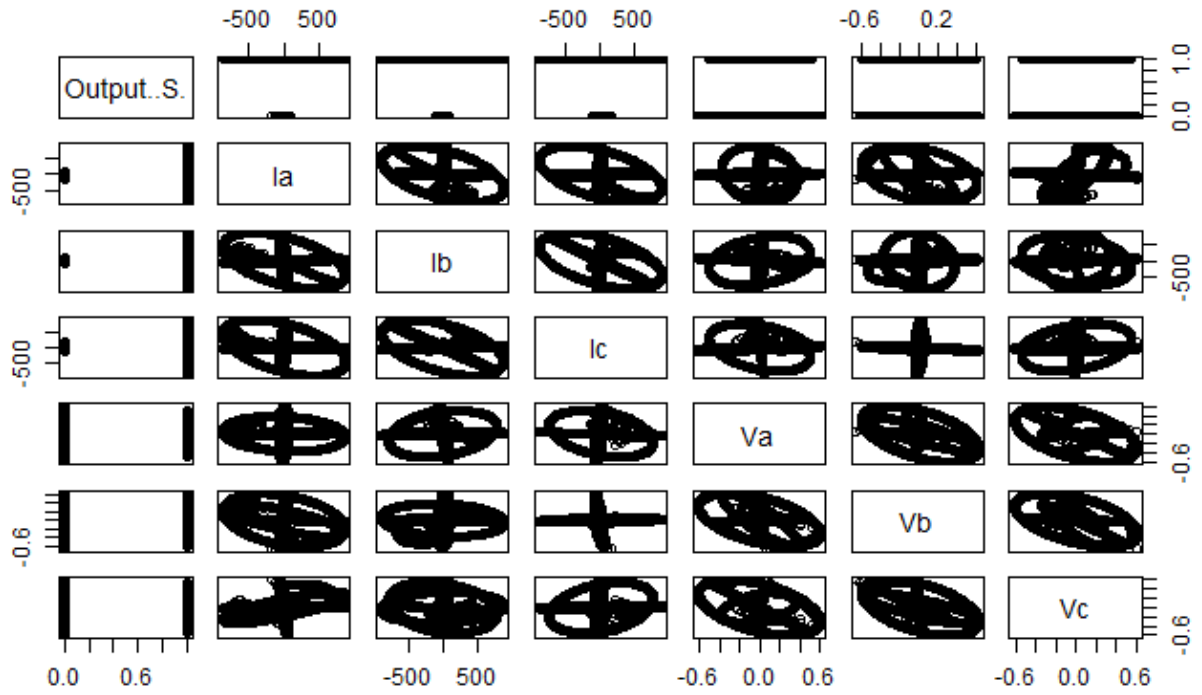


*Figure 9 Dataset values in a graph*

Plotted a graph using different columns means the y-axis shows actual values present in the dataset and the x-axis shows times ranges.

```r
plot(df$Output..S.,col='yellow')
plot(df$Ia,col='light green')
plot(df$Ib, col='lightgray')
plot(df$Ic,col='pink')
plot(df$Va,col='purple')
plot(df$Vb,col='red')
plot(df$Vc,col='light pink')
```

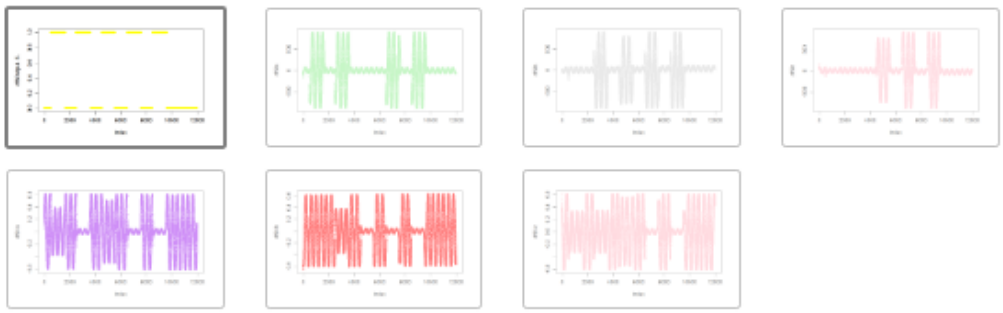

*Figure 10 Dataset value on X-axis & Y-axis*

Plotted different graphs to visualize the attributes' index values.

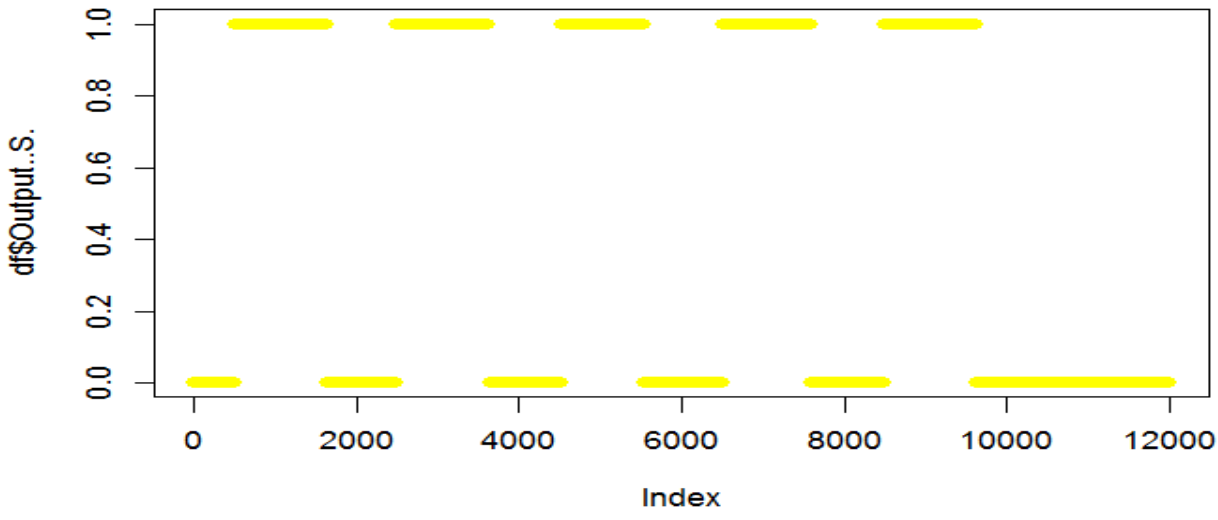


*Figure 11 Plotting the output in the data frame*

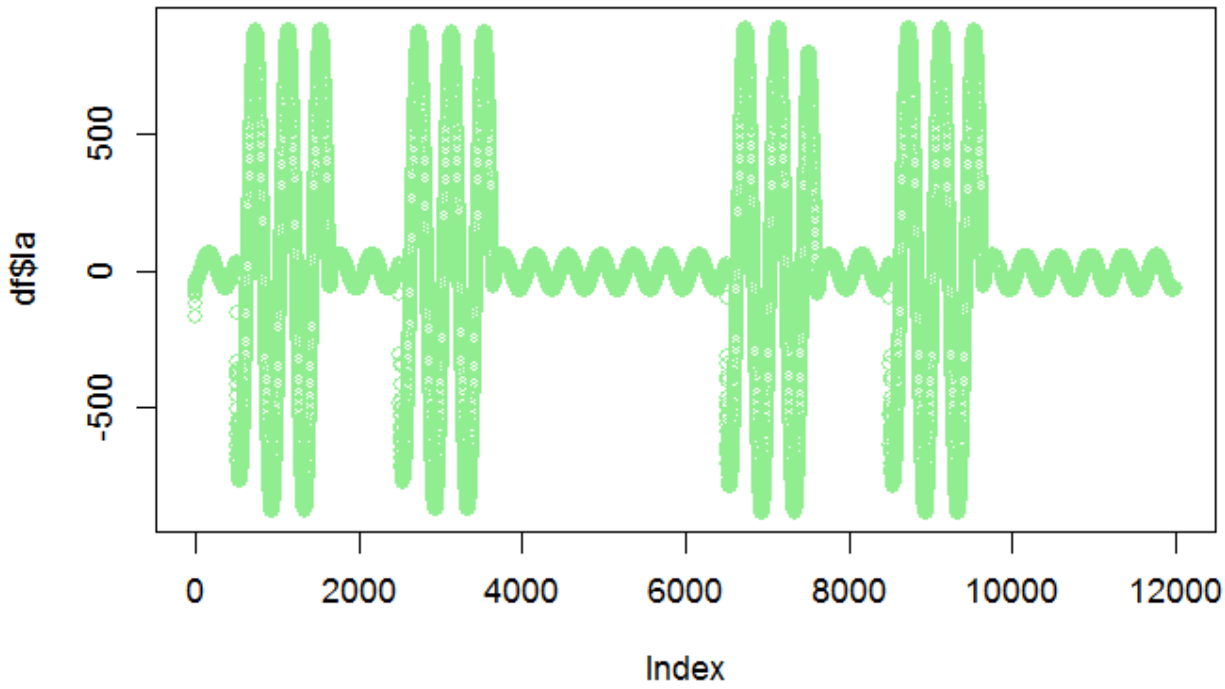


*Figure 12 Plotting Ia in the data frame*

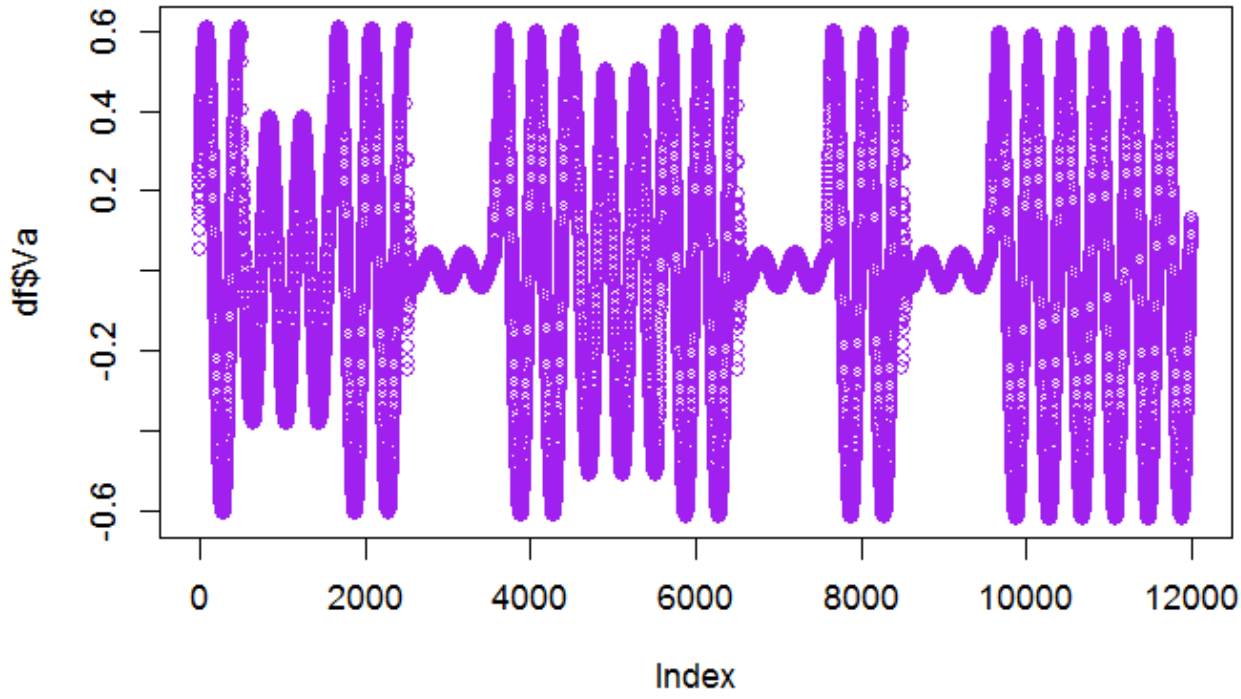


*Figure 13 Plotting Va in the data frame*

The histogram graph here shows that the y-axis is the frequency, and the x-axis shows the actual value or column present in the dataset. So, this graph also shows frequency value ups and downs compared to the value of the dataset.

```r
```{r}
hist(df$Output..S.,
main="Show class Data set for colomn Output (S)",
xlab="G colomn for class data set",
col="green",
)

hist(df$Ia,
main="Show class Data for coloumn Ia",
xlab="Ia for class data",
col="Brown",
)
hist(df$Ib,
main="Show class Data for coloumn Ib",
xlab="Ib for class data",
col="pink",
)
hist(df$Ic,
main="Show class Data for coloumn Ic",
xlab="Ic for class data",
col="purple",
)
hist(df$Va,
main="Show class Data for coloumn Va",
xlab="Va for class data",
col="darkblue",
)
```

*Figure 14 Displaying Histograms*
```

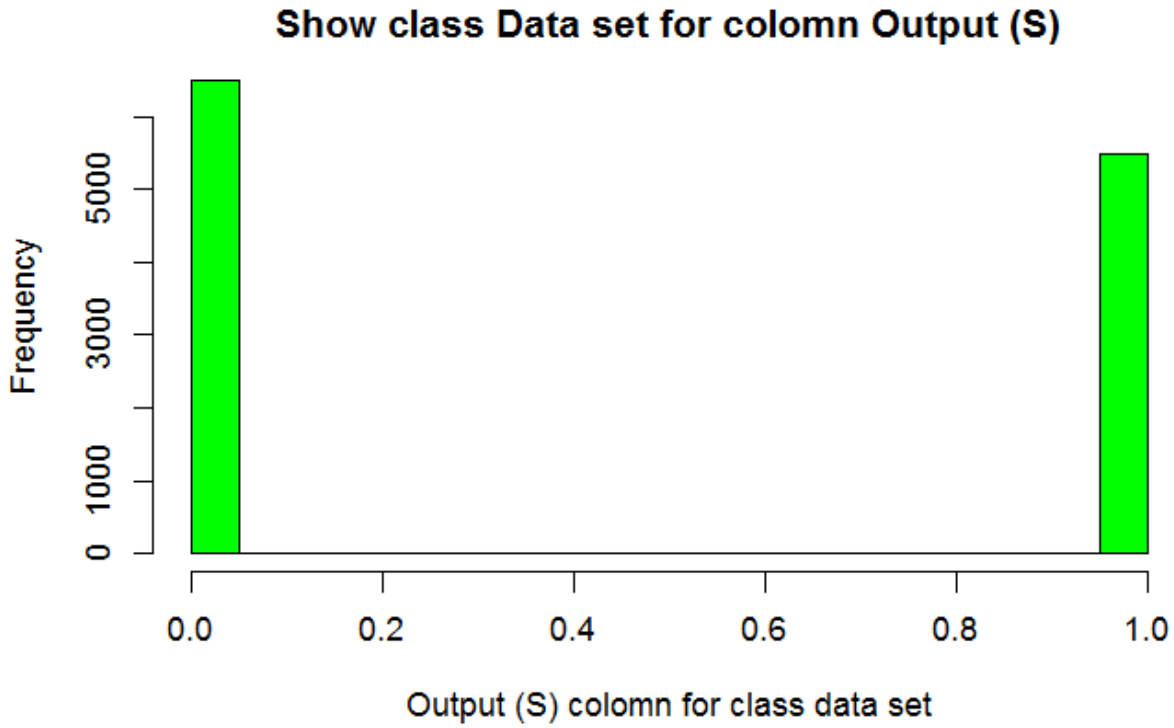


*Figure 15 Frequency of Output (S)*

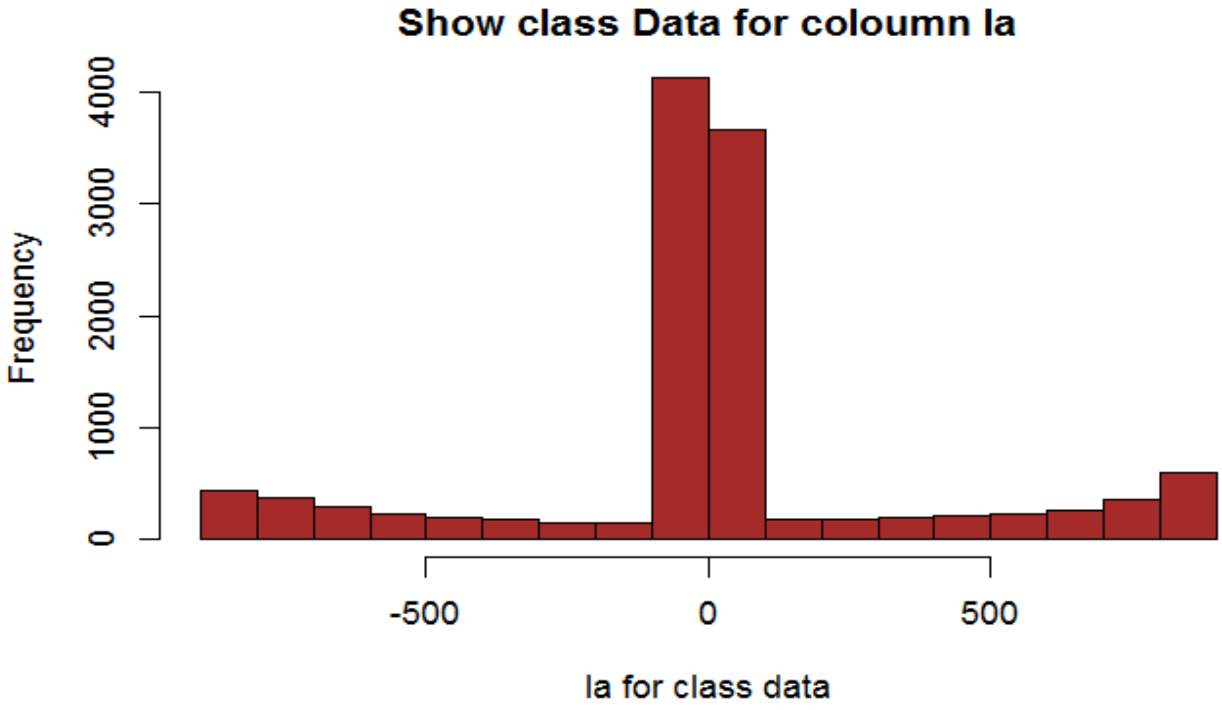


*Figure 16 Frequency of Ia*

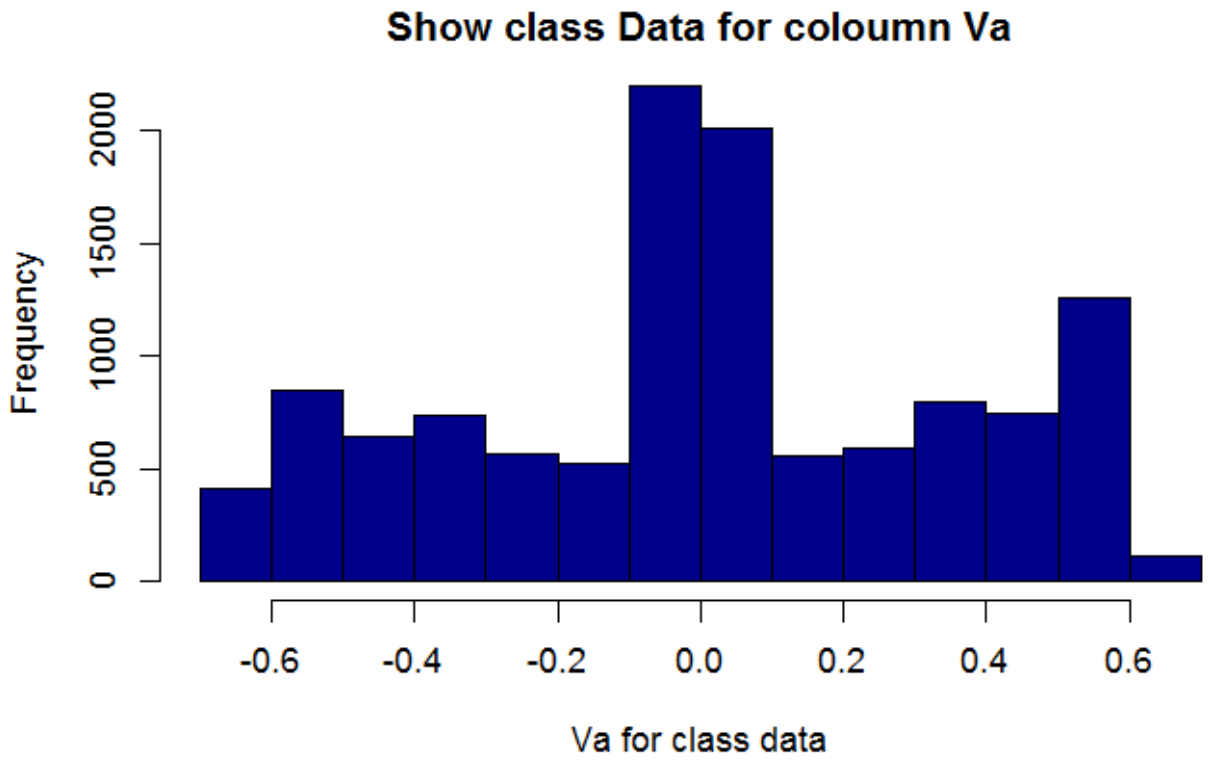


*Figure 17 Frequency of Va*

## 4.3 Train-Test the Dataset

To train and test the dataset, we followed the standard approach which is a 70-30 ratio. From the total dataset, 70% of the dataset is used to train and 30% of the dataset is used to test the trained models. Dimensions of train and test dataset is also presented.

```
## Train & test your data set
```{r}
set.seed(1)

#Use 70% of dataset as training set and remaining 30% as testing set
sample <- sample(c(TRUE, FALSE), nrow(df), replace=TRUE, prob=c(0.7,0.3))
train  <- df[sample, ]
test   <- df[!sample, ]

#view dimensions of training set
dim(train)

#view dimensions of test set
dim(test)
```
```

```
[1] 8371    7
[1] 3630    7
```

*Figure 18 Splitting dataset into training & testing subsets*

## 4.4 Data Pre-processing

In the data exploration phase, we observed our desired dataset is a nominal attribute-paired dataset containing only numeric data for all attributes. So, we did not require any preprocessing steps to clean our dataset. We also found there is no null value associated with any column because every value for the individual column is calculated from the generators and automatically stored.

## 4.5 Data Modeling

We processed our data using a variety of machine-learning algorithms. We selected the most popular and appropriate classifiers for the classification task that involves electric fault identification. The machine learning model will be trained on the training subset in the modelling part, and its performance will be evaluated against the unknown testing subset by creating a confusion matrix.

### 4.5.1 SVM Classifier

SVM, or support vector machines, are effective models for sorting data with various attributes. SVM builds a dimensional space in high-dimensional space by using the linear function hypothesis. Using a hyperplane in p-dimensional space, classes can be divided. Additionally, SVM uses the margin, a method to choose the best hyperplane that divides classes maximizing. SVM is among the most effective and widely used among all classification techniques, text classification methods are widely used. SVMs possess the being capable of learning independently of the dimension of the feature set is a special quality. This argues that our data can be distinguished from other data with high confidence using variables from the hypothesis. Despite the presence of many features in the space, we can generalize. The selection of parameters that results in the hypothesis also with minimum VC-Dimension is the ideal one. This makes expensive cross-validation unnecessary and enables fully automated parameter tuning.

In RStudio, we used the SVM function and confusion matrix function as seen in the following chunk:

```
#SVM Classifier
```{r}
M_SVM = svm(Output ~.,data = train,kernel='linear',cost=0.1)
P_SVM = predict(M_SVM,test)
head(P_SVM)
accuracy_svm <- mean(P_SVM == test$Output)
sprintf("Accuracy of SVM: % s ",accuracy_svm)
confusionMatrix(P_SVM,test$Output)
```
```

*Figure 19 SVM Classifier*

```
[1] "Accuracy of SVM: 0.7254792998055022 "
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1951  988
         1    0  660

               Accuracy : 0.7255
                 95% CI : (0.7106, 0.74)
    No Information Rate : 0.5421
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.42

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 1.0000
            Specificity : 0.4005
         Pos Pred Value : 0.6638
         Neg Pred Value : 1.0000
             Prevalence : 0.5421
         Detection Rate : 0.5421
   Detection Prevalence : 0.8166
      Balanced Accuracy : 0.7002

       'Positive' Class : 0
```

*Figure 20 SVM Confusion Matrix*

We can see that the accuracy estimated through the confusion matrix is 72.55%.

### 4.5.2 Decision Tree

A supervised non-parametric classifier for the classification method is the decision tree. To forecast the output value, DT uses training data to learn simple rules. Although DT is simple to comprehend and depict, it does not accept missing data. As a result, it set itself up to work without normalizing the data. DT is applicable to category and numeric data formats. Although DT uses white boxing, statistical analysis allows us to estimate its performance

In RStudio, we used the decision tree function and confusion matrix function as seen in the following chunk:

```
#Decision Tree
```{r}
tree <- rpart(Output~. , data = train)
result <- predict(tree,test,type = 'class')
accuracy_tree <- mean(result == test$Output)
sprintf("The accuracy of Decision Tree: %s",accuracy_tree)
confusionMatrix(result,test$Output)
```
```

*Figure 21 Decision Tree Classifier*

```
[1] "The accuracy of Decision Tree: 0.987774381772715"
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1945   38
         1    6 1610

               Accuracy : 0.9878
                 95% CI : (0.9836, 0.9911)
    No Information Rate : 0.5421
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9753

 Mcnemar's Test P-Value : 2.962e-06

            Sensitivity : 0.9969
            Specificity : 0.9769
         Pos Pred Value : 0.9808
         Neg Pred Value : 0.9963
             Prevalence : 0.5421
         Detection Rate : 0.5404
   Detection Prevalence : 0.5510
      Balanced Accuracy : 0.9869

       'Positive' Class : 0
```

*Figure 22 DT Confusion Matrix*

We can see that the accuracy estimated through the confusion matrix is 98.78% which is very high and considered to be over fitted.

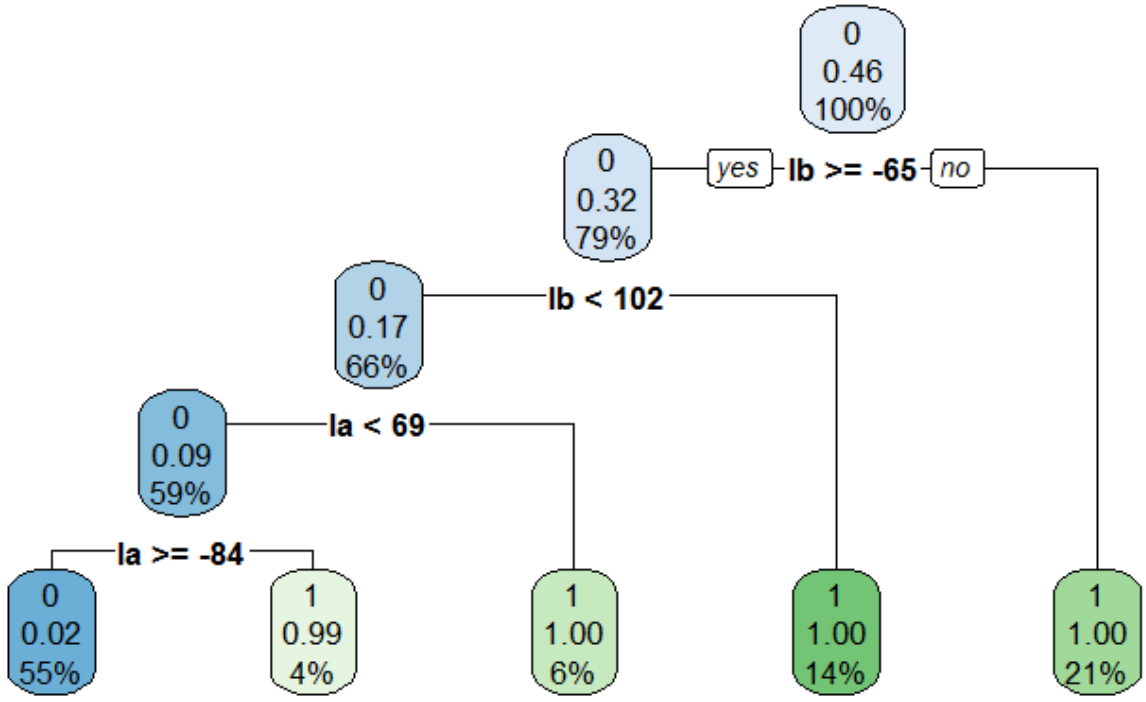


*Figure 23 Decision Tree*

As per the decision tree, if the current of B-phase is less than -65A, there's a fault in the network. If the current is greater than or equal to -65A, the output depends on other values as shown in the figure.

### 4.5.3 XG Boost Classifier

XG is a classification and regression problem-based classifier, the boosting classifier. It is a component of ensemble machine-learning techniques. Boosting is mentioned in this context since a cost function is being utilized to optimize the model. Its operation relies on the cooperation of several inaccurate weak learners and accurate strong learners. Each case receives full attention, and it picks up additional rules from the distribution of past misclassifications. It uses a decision tree to spread predecessors in a network. It instantly over fits the corpus data using the greedy method. It utilizes regularization techniques to penalize the modelling of the various sections. As a result, overfitting

was decreased, and model performance was enhanced.

In RStudio, we used the XGBoost and confusion matrix functions as seen in the following chunk:

```
#XGBoost
```{r}

train_control <- trainControl(method = 'cv',
                              number = 3,
                              verboseIter = TRUE,
                              allowParallel = TRUE)

boosting1 <- train(Output~., data = train,
                   method ='xgbTree',
                   trControl = train_control,
                   tuneGrid = expand.grid(nrounds = 500,
                                          max_depth = 3,
                                          eta = 0.2,
                                          gamma = 2.1,
                                          colsample_bytree=1,
                                          min_child_weight = 1,
                                          subsample = 1))
boost_pred <- predict(boosting1, newdata = test)
cm_boost <-table(test$Output,boost_pred)
confusionMatrix(cm_boost)

```
```

*Figure 24 XGBoost Classifier*

```
Confusion Matrix and Statistics

   boost_pred
       0    1
  0 1945    6
  1    9 1639

               Accuracy : 0.9958
                 95% CI : (0.9931, 0.9977)
    No Information Rate : 0.5429
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9916

 Mcnemar's Test P-Value : 0.6056

            Sensitivity : 0.9954
            Specificity : 0.9964
         Pos Pred Value : 0.9969
         Neg Pred Value : 0.9945
             Prevalence : 0.5429
         Detection Rate : 0.5404
   Detection Prevalence : 0.5421
      Balanced Accuracy : 0.9959

       'Positive' Class : 0
```

*Figure 25 XGBoost Confusion Matrix*

We can see that the accuracy estimated through the confusion matrix is 99.58% which is very high and considered to be over fitted.

### 4.5.4 Random Forest

Random Forest is a component of the ensemble learning approach, which is particularly effective in solving the classification problem. With accurate prediction, it performs better than a sizable amount of data. To obtain the forecasts, a variety of configuration packages are provided. The parameter settings were made defaults. Anybody can adjust the parameter settings to suit their needs. To generate trees, it uses bagging and randomization features. Instead of looking for the most crucial feature, it looks for the best features among all the grown nodes. There are various parameters and properties for this classifier. It is a high-dimensional data classifier that is off the shell. For each tree distribution, it generates a mixture of tree predictors. The tree distribution fully extends and may reach a very huge tree with the default configuration value. Parameter tweaking is essential in accordance with the criteria to control this problem. Additionally, it resolves the memory usage problem.

In RStudio, we used the Random Forest and confusion matrix functions as seen in the following chunk:

```
#RandomForest
```{r}
RFM = randomForest(Output ~.,data = train)
P_RFM= predict(RFM,test)
accuracy_RFM <- mean(P_RFM == test$Output)
sprintf("Accuracy of Random Forest: % s ",accuracy_RFM)
confusionMatrix(P_RFM,test$Output)

```
```

*Figure 26 Random Forest Classifier*

```
[1] "Accuracy of Random Forest: 0.997777160322312 "
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1944    1
         1    7 1647

               Accuracy : 0.9978
                 95% CI : (0.9956, 0.999)
    No Information Rate : 0.5421
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9955

 Mcnemar's Test P-Value : 0.0771

            Sensitivity : 0.9964
            Specificity : 0.9994
         Pos Pred Value : 0.9995
         Neg Pred Value : 0.9958
             Prevalence : 0.5421
         Detection Rate : 0.5402
   Detection Prevalence : 0.5404
      Balanced Accuracy : 0.9979

       'Positive' Class : 0
```

*Figure 27 Random Forest Confusion Matrix*

We can see that the accuracy estimated through the confusion matrix is 99.78% which is also very high and considered to be over fitted.

### 4.5.5 Logistic Regression

Predictive analytics and categorization frequently make use of this kind of statistical model, also referred to as a logit model. Based on a particular dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent variable's range is 0 to 1. In logistic regression, the odd that is, the chance of success shared by the probability of failure is transformed using the logit formula.

In RStudio, we used the logistic regression and confusion matrix function as seen in the following chunk:

```r
#Logistic Regression
```{r}
M_LOG <-glm(Output ~.,data = train, family = "binomial")
summary(M_LOG)
P_LOG <- predict(M_LOG, test, type = "response")
head(P_LOG)
predict_L <- ifelse(P_LOG >= .5 , 1 , 0)
accuracy <- mean(predict_L == test$Output)
sprintf("Accuracy of Logistic Regression: % s ",accuracy)
confusionMatrix(factor(predict_L),test$Output)

```
```

*Figure 28 Logistic Regression*

```
[1] "Accuracy of Logistic Regression: 0.732147818838566
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1950  963
         1    1  685

               Accuracy : 0.7321
                 95% CI : (0.7174, 0.7466)
    No Information Rate : 0.5421
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4349

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9995
            Specificity : 0.4157
         Pos Pred Value : 0.6694
         Neg Pred Value : 0.9985
             Prevalence : 0.5421
         Detection Rate : 0.5418
   Detection Prevalence : 0.8094
      Balanced Accuracy : 0.7076
```

*Figure 29 Logistic Regression Confusion Matrix*

## 4.6 Results

In the modeling part, we used five supervised machine learning models, SVM, Decision Tree, XGBoost, random forest and logistic regression. The table below summarizes the comparison of them all. Data was split 70/30, and the values are of the test data.

*Table 2 Models Statistical Performance*

| Classifier | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| **SVM** | 0.7255 | 1 | 0.4005 |
| **Decision Tree** | 0.9878 | 0.9969 | 0.9769 |
| **XGBoost** | 0.9958 | 0.9954 | 0.9964 |
| **Random Forest** | 0.9978 | 0.9964 | 0.9994 |
| **Logistic Regression** | 0.7321 | 0.9995 | 0.4157 |

Random Forest scored the highest in accuracy (99.78%) followed by XGBoost (99.58%), Decision tree (98.78%), logistic regression (73.21%) and SVM with the lowest accuracy (72.55%). High accuracy usually indicated over fitting.

Sensitivity is the percentage of true Positive (TP, model predicted positive and the actual is positive) divided by TP and False Negative (FN, model predicted negative and the actual is positive). Although SVM scored low on accuracy, it has the highest sensitivity, 100%. The next model with the highest sensitivity is logistic regression, followed by decision tree, random forest and finally XGBoost. They all scored over 99%.

Specificity is the percentage of true negative (TN, model predicted negative and the actual is negative), divided by TN and false positive (FP, model predicted positive and the actual is negative). Random forest scored highest with 99.94%, followed by XGBoost, decision tree, logistic regression and finally SVM with 40.05%.

To sum up, we produced three models that could be over fitted (Decision tree, random forrest, and XGBoost). SVM and logistic regression had reasonable and realistic accuracies, around 73%.

# Chapter 5 Conclusion & Recommendations

## 5.1 Conclusion

The processes for error detection, detection, and identification in overhead wires are thoroughly examined in this study. The solution is then able to offer advice on potential strategies for incorporating to forecast anticipated issues in the electric system. The three classifiers, Random Forest, XGBoost and Decision tree are producing high accuracies, while Logistic Regression and SVM are producing realistic accuracy results. As a result of its ability to predict events in the most efficient manner, random forest is currently ideally outperforming other models in terms of fault detection, just as what was learned in the literature review.

## 5.2 Recommendations

For future work, we recommend few suggestions to work with advance machine learning models and deep learning models. Then identifying which model is best fit for electrical faults in electric system using same dataset. Another recommendation is to create a new dataset with different no. of parameters for the same problem and them deploying the previous and new suggested methodology. These results will be verified using WEKA tool and then further exploring using R language in RStudio or any other R language supportive IDE. Moreover, the used models, such can undergo parameters tuning which might yield a higher/realistic accuracy, one challenge was the time it takes to build the model.

# References


1. Andresen, C. A., Torsaeter, B. N., Haugdal, H., & Uhlen, K. (2018, October 16). Fault Detection and Prediction in Smart Grids. 9th IEEE International Workshop on Applied Measurements for Power Systems, AMPS 2018 - Proceedings. https://doi.org/10.1109/AMPS.2018.8494849
2. A. Liaw and M. Wiener, "Classification and Regression with Random Forest," R News, 2002.
3. A. M. Andrew, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods," 2001.
4. Balouji, E., Gu, I. Y. H., Bollen, M. H. J., Bagheri, A., & Nazari, M. (2018). A LSTM-based deep learning method with application to voltage dip classification. Proceedings of International Conference on Harmonics and Quality of Power, ICHQP, 2018-May, 1–5. https://doi.org/10.1109/ICHQP.2018.8378893
5. BUTLER, S. (2001). The nature of UK electricity transmission and distribution networks in an intermittent renewable and embedded electricity generation future, Centre for Environmental Technology.
6. Eskandarpour, R., & Khodaei, A. (2017). Machine Learning Based Power Grid Outage Prediction in Response to Extreme Events. IEEE Transactions on Power Systems, 32(4), 3315–3316. https://doi.org/10.1109/TPWRS.2016.2631895
7. Gorunescu, F. (2011). Data Mining: Concepts, models and techniques (Vol. 12). Springer Science & Business Media.

8. Guo, H., Nguyen, H., Vu, D.-A., & Bui, X.-N. (2021). Forecasting mining capital cost for open-pit mining projects based on artificial neural network approach. Resources Policy, 74, 101474.
9. Han, J., Kamber, M. and Pei, J. (2012). Data Mining Concepts and Techniques (3rd ed.).
10. Han, J., Hu, J., Yang, Y., Wang, Z., Wang, S. X., & He, J. (2015). A Nonintrusive Power Supply Design for Self-Powered Sensor Networks in the Smart Grid by Scavenging Energy from AC Power Line. IEEE Transactions on Industrial Electronics, 62(7), 4398–4407. https://doi.org/10.1109/TIE.2014.2383992
11. Høiem, K. W., Santi, V., Torsæter, B. N., Langseth, H., Andresen, C. A., & Rosenlund, G. H. (2020). Comparative study of event prediction in power grids using supervised machine learning methods. 2020 International Conference on Smart Energy Systems and Technologies (SEST), 1–6.
12. Lei, G., Dai, M., Tan, Z., & Wang, Y. (2011). The Research of CMMB Wireless Network Analysis Based on Data Mining Association Rule. 2011 7th International Conference on Wireless Communications, Networking and Mobile Computing, 1–4.
13. Mahmoud, M. A., Md Nasir, N. R., Gurunathan, M., Raj, P., & Mostafa, S. A. (2021). The current state of the art in research on predictive maintenance in smart grid distribution network: Fault's types, causes, and prediction methods—a systematic review. Energies, 14(16), 5078.
14. Picard, J. L., Aguado, I., Cobos, N. G., Fuster-Roig, V., & Quijano-López, A. (2021). Electric distribution system planning methodology considering distributed energy resources: A contribution towards real smart grid deployment. Energies, 14(7).

https://doi.org/10.3390/en14071924

15. Russell, B. D., Benner, C. L., Cheney, R. M., Wallis, C. F., Anthony, T. L., & Muston, W. E. (2009). Reliability improvement of distribution feeders through real-time, intelligent monitoring. 2009 IEEE Power & Energy Society General Meeting, 1–8.
16. Shuvro, R. A., Das, P., Hayat, M. M., & Talukder, M. (2019). Predicting cascading failures in power grids using machine learning algorithms. 2019 North American Power Symposium (NAPS), 1–6.
17. Skydt, M. R., Bang, M., & Shaker, H. R. (2021). A probabilistic sequence classification approach for early fault prediction in distribution grids using long short-term memory neural networks. Measurement: Journal of the International Measurement Confederation, 170. https://doi.org/10.1016/j.measurement.2020.108691
18. Sowah, R. A., Dzabeng, N. A., Ofoli, A. R., Acakpovi, A., Koumadi, K. M., Ocrah, J., & Martin, D. (2018). Design of power distribution network fault data collector for fault detection, location and classification using machine learning. IEEE International Conference on Adaptive Science and Technology, ICAST, 2018-August. https://doi.org/10.1109/ICASTECH.2018.8506774
19. Tyvold, T. S., Nybakk Torsater, B., Andresen, C. A., & Hoffmann, V. (2020, September 1). Impact of the temporal distribution of faults on prediction of voltage anomalies in the power grid. SEST 2020 - 3rd International Conference on Smart Energy Systems and Technologies. https://doi.org/10.1109/SEST48500.2020.9203569
20. Viegas, J. L., Vieira, S. M., Melício, R., Matos, H. A., & Sousa, J. M. C. (2016). Prediction of events in the smart grid: Interruptions in distribution transformers. 2016 IEEE International Power Electronics and Motion Control Conference (PEMC), 436–441.

21. Wang, Y., Pordanjani, I. R., & Xu, W. (2011). An Event-Driven Demand Response Scheme for Power System Security Enhancement. IEEE Transactions on Smart Grid, 2(1), 23–29. https://doi.org/10.1109/TSG.2011.2105287
22. OFGEM. (2017). RIIO electricity distribution annual report 2016-17 (Rep.).
23. Han, J., Kamber, M. and Pei, J. (2012). Data mining. 3rd ed. Haryana, India: Elsevier
24. Gorunescu, F. (2011). Data Mining: Concepts, models, and techniques (Vol. 12), 250-251. Springer Science & Business Media.
25. G. Lei, M. Dai, Z. Tan and Y. Wang, 2011. "The Research of CMMB Wireless Network Analysis Based on Data Mining Association Rule," 2011 7th International Conference on Wireless Communications, Networking and Mobile Computing, Wuhan, pp. 1-4.
26. Guo, Hongquan & Nguyen, Hoang & Vu, Diep-Anh & Bui, Xuan-Nam. (2019). Forecasting mining capital cost for open-pit mining projects based on an artificial neural network approach. Resources Policy. 10.1016/j.resourpol.2019.101474
27. Joaquim L. Viegas, Susana M. Vieira, Rui Mel´ıcio, "Prediction of events in the smart grid: interruptionsin distribution transformers," 2016 2016 IEEE International Power Electronics and Motion Control Conference (PEMC) doi: 10.1109/EPEPEMC.2016.77520
28. C. A. Andresen, B. N. Torsæter, H. Haugdal and K. Uhlen, "Fault Detection and Prediction in Smart Grids," 2018 IEEE 9th International Workshop on Applied ~ 202 ~ Measurements for Power Systems (AMPS), Bologna, 2018, pp. 1-6, doi: 10.1109/AMPS.2018.8494849.
29. R. A. Shuvro, P. Das, M. M. Hayat, and M. Talukder, “Predicting Cascading Failures in Power Grids using Machine Learning Algorithms,” 51st North American Power Symposium, NAPS 2019, no. 1541148, pp. 0–5, 2019