Comparative Data Analysis of a PV Module system considering weather parameters.

By

Nafiz Ahmed
16121123
Khandoker Samiul Hoque
16321099
Sabbir Ahmad
17321022
Didar Alam Siddiki
16321057

A thesis submitted to the Department of Electrical and Electronic Engineering in partial fulfillment of the requirements for the degree of
Bachelor of Science in Electrical Engineering.

Department of Electrical and Electronic Engineering
Brac University
June 2021

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I/We have acknowledged all main sources of help.

Student's Full Name & Signature:

Nafiz Ahmed 16121123	**Sabbir Ahmad** 17321022
Didar Alam Siddiki 16321057	**Khandoker Samiul Hoque** 16321099

Approval

The thesis “Comparative Data analysis of a PV system considering weather parameters” submitted by

1. Nafiz Ahmed (16121123)
2. Khandoker Samiul Hoque (16321099)
3. Sabbir Ahmad (17321022)
4. Didar Alam Siddiki (16321057)

of Summer, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science in Electrical and Electronic Engineering on 9th June 2021.

Examining Committee:

Supervisor:
(Member)

_____
Dr. Md. Mosaddequr Rahman
Professor and Chairperson Dept. of EEE
BRAC University

Program Coordinator:
(Member)

_____
Abu S.M. Mohsin, PhD
Professor, Dept. of EEE
BRAC University

Departmental Head:
(Chair)

_____
Dr. Md. Mosaddequr Rahman
Professor and Chairperson, Dept. of EEE
BRAC University

Abstract/ Executive Summary

Fastest growing economy of Bangladesh increase the great demand of power generation using renewable energy sources. However, uncertainty in the output power of the photovoltaic (PV) power generation station due to variation in meteorological parameters is of serious concern. As a solution to this issue this work aims to predict the accurate power of a PV system. The performance results of this study are presented in terms of Random forest, Artificial Neural Network (ANN) and Multiple Linear Regression model. Additionally, the performance results obtained with Random forest, Artificial Neural Network (ANN) and Linear Regression are compared to show that which model has better prediction accuracy and less error. This paper aims to employ and perform a comparison study of PV systems considering weather parameter using different algorithms of above-mentioned data forecasting methods. The data which will be taken from the mentor of our thesis will be used in this paper. The present study will also be very helpful to provide technical guidance to the prediction of the PV power System.

Keywords: photovoltaic (PV); Random forest; Artificial Neural Network (ANN); Multiple Linear Regression model; Short Circuit Current; PV Module.

Dedication

This thesis is dedicated to our beloved parents who raised us to be the persons we are today, who were always there for us whenever we needed them, and who constantly support us with their love and kindness. Also, to our supervisor who continuously guided us to do our work properly. Lastly to Siddiqure Rahaman, father of Didar Alam Siddiki who passed away during the process of our thesis work.

Acknowledgement

All praise to Almighty Allah, who enabled us to complete our thesis work on schedule and without major setbacks. We would like to express our heartfelt appreciation to our distinguished supervisor, Dr. Mosaddequr Rahman Chairperson of BRAC University, for his guidance and support and excellent directions, as well as his ongoing support and enthusiasm throughout the thesis work We are really grateful for his dedicated participation at every step, professionalism, and invaluable guidelines that opened the road for the completion of this thesis work.

Table of Contents

List of Tables

List of Figures

List of Acronyms

RF = Random Forest Regression

ANN= Artificial Neural Network

MLR= Multiple Linear Regression

PV= Photo Voltaic

ML= Machine Learning

MPP=Maximum Power Point

MPPT= Maximum Power Point Tracking

Chapter 1

Introduction

1.1 Introduction

Sunlight is one of the energy sources what we get from nature which is utilizing for the solar system. From trees to our electric bulb all feed on this energy. Energy can be produced in various ways. Most of them cause great damage to our environment. Producing energy in an environmentally friendly way is difficult unless solar cells are used. Using solar cells to produce energy does not harm our environment and it is cost-efficient. As sunlight, a source of renewable energy, using solar cells can decrease the scarcity of fuels and other natural resources used to produce electrical energy. The whole world is running towards renewable energy so using solar cells is a popular choice. Here in our country almost 14% of the households get electricity directly from solar cells and the government have some huge plans for the future related to this. So, here the importance of using solar cells to produce energy in the present time and for the future. Solar cells are also known as PV (Photovoltaic) panels. However, PV panel generation is sensitive towards some weather parameters such as Wind Speed, Humidity, Temperature, Rain Drop, Solar Irradiance, etc. Here in this study is considering Wind Speed, Humidity, Air Pressure, and Temperature parameters and using ANN, Random Forest and Linear Regression machine learning algorithms to forecast data that were collected from two Mono-Silicon PV modules to find out more accurate day long forecasting which can be essential for commercial and non-commercial users as the generation of PV panel is always uncertain. So, in this paper we tried day long forecasting data testing a whole month of day long data by machine learning algorithms considering some weather parameters.

1.2 Literature Review

Natural resources are limited in our world and by producing energy using those resources made the reservation even less. To save those reserved resources people need methods of producing energy where these resources are not used. So, researchers started studying on how can energy be produced without using natural resources and also what can be friendly towards our nature. Producing energy using direct sunlight is ecofriendly and also it is the biggest source of renewable energy. To produce energy using sunlight we need to use Photovoltaic Panel or Solar Panel. Many scientists and researchers started working on increasing the efficiency on PV module and this study is still going on. Forecasting data of PV module is a part of this study. Many papers have been published related to this study. Therefore, some studies related to this study is mentioned below. In the paper, the author, discussed about the solar power generation forecasting using RF model [1]. Then daily PV power generation using RF model in China is discussed in this work [2]. RF based approach had been used to track MPP of PV system operating under actual environment been discussed in another work [3]. Multiple Liner Regression model works better with clean sky hours had been proven by two university researchers in a research work [4]. Finally, In a work in this paper tried to find out, by using multiple linear regression, which factors affect a footballer's market value. Finally, they concluded that young talents are valuable and players with 180 cm and 184 cm height have proven their success [5]. All these works inspired us to do this work.

1.3 Aim and Objective

Aim of this study is to compare analysis using machine learning algorithms. Here we used PV module data considering some weather parameters which are Wind Speed, Air Pressure, Humidity and Temperature. We used 5 months day long of PV module data and tried to predict

last day of each month data and compared with the original data that we had. The main objective of this thesis book is to compare the results that we got from our machine learning algorithms and thus choose the model that gives the highest accuracy.

1.4 Organization of the Thesis

In our thesis book the very first section that we wrote is the Introduction Section. Later on, we wrote about the theory related to this study which is Theoretical Background. On the third section of this thesis book three machine learning algorithms were explained which are Artificial Neural Network (ANN), Multiple Linear Regression, Random Forest. After that on the fourth section Result Analysis was explained that we got from our machine learning models. Firth section of the thesis book is the Conclusion of the book and lastly References were given.

Chapter 2

Theoretical Background

2.1 Introduction

In this section of theoretical background, we will be explaining about the PV panels, data collection, effects of weather parameters on PV panel generation and will try to give a glimpse of machine learning algorithms that we used in our thesis. For the weather parameters we are considering Wind Speed, Temperature, Humidity and Air Pressure. Although we did not consider Solar Irradiance in our forecasting method, we will try to give a little explanation

about how Solar Irradiance effects the generation of PV panels and the performance of PV panels.

2.1.1 PV Panel

Before diving in to the PV panel that is used to collect data for this work we need to know some things about PV panel. What is PV panel, how does this work, mechanism of it and some characteristics of PV panel are thigs that are essential to know. Firstly, we will be talking about what is PV panel and later on we will be explain how does the PV panel works, mechanism of PV panel and the characteristics of PV panel in this chapter. In the characteristics of PV panel portion short and brief explanation of Open Circuit Voltage (Voc), Short Circuit Current (Isc), Maximum Power Point (MPP), Fill Factor (FF), Power Efficiency (%eff) had been given. Also, on the Effects of Weather Parameters on PV Panel section weather parameters like Wind Speed, Humidity, Air Pressure and temperature had been also been explained briefly. Knowing about all these things will enrich the understanding of PV panel and will further help us understand this paper even more.

2.1.2 What is PV panel

Before knowing about the PV panel, we have to know about how the PV panel is made. PV panel is the short form of Photovoltaic panel. PV panel is also known as Solar panel. PV panel or Solar panel are the arrangement of a number of PV cells or solar cell together. A solar cell or PV cell is an electronic device that works directly to convert light energy into electrical energy through the photoelectric effect, which is a physical and chemical mechanism altogether. We all know that Sun is the power house of our world. This solar cell or PV cell converts this sunline directly into electrical energy and how this cell does this will be explained into how the PV panel or Solar panel works. Solar cells are nothing but p-n junctions, while solar panels are nothing more than semiconductors constructed of silicon. The current and voltage of a solar panel are increased by connecting many cells in a series-parallel configuration. Solar cells, like batteries, contain 2 layers. One positive layer and one negative layer. These positive and negative layers of Solar cells altogether create an electric field when it is exposed to direct sunlight on the day time. There several types of solar panel. Among them, three of the Solar panels are Monocrystalline, Polycrystalline and the Thin-film. All these three panels have their own advantages and disadvantages upon one another. A table of advantages and disadvantages of different type of solar panel or PV panel is mentioned below

Table-2.1 – Advantages and Disadvantages of PV Panels.

Type of Solar Panel	Advantages	Disadvantages
Monocrystalline	High efficiency	Expensive
Polycrystalline	Cost efficient	Low efficiency
Thin-film	Portable and flexible	Low efficiency

Although both monocrystalline and polycrystalline solar panels include silicon cells, the silicon content in monocrystalline and polycrystalline panels changes. A single, pure silicon crystal is used to make monocrystalline solar cells. Polycrystalline solar cells, on the other hand, are constructed out of bits of silicon crystal that are fused together in a mold before being sliced into wafers. Solar panels with a thin layer are known as thin-film panels [6]. Photos of Monocrystalline, Polycrystalline and Thin-film are given below

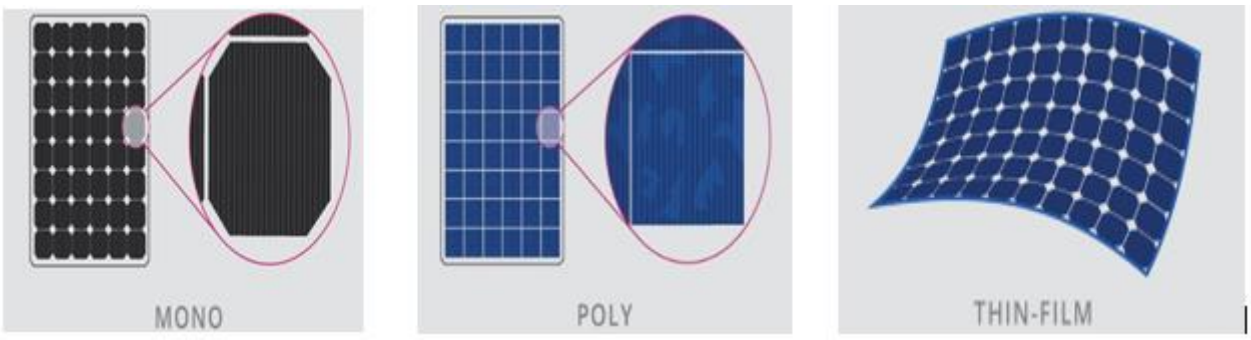


Figure 2.1: Mono, Poly and Thin Film Crystalline PV Panel.

2.1.3 How PV panels work

Photovoltaic cells need to operate on an electrical field. When opposite poles are separated, a magnetic field is formed while an electric field is formed when opposite charges are separated. Manufacturers of silicone "dope" with other materials, to produce this field, which gives a positive or negative electric charge for each piece of the sandwich. They sew phosphorus into the top of the silicon layer, which provides extra negative electrons to the layer. In the meantime, boron is added to the underlying layer, which results in fewer electrons and a positive charge. All this adds up to a silicone layer connecting electric field. The electric field drives the electron from the silicone connection if a sunlight photon knocks a loosening electron. These electrons are transformed by a few additional cell components into useable electricity. The electrons are gathered and transported on the side of the cell by metal conductive plates. The electrons can then flow freely, just as any other electricity source.

2.1.4 Mechanism of Solar Panel

The photovoltaic effect is at the heart of the science of generating electricity with solar panels. The photovoltaic effect, discovered by Edmond Becquerel in 1839, is a property of certain materials (known as semiconductors) that allows them to generate an electric current when exposed to sunlight. The photovoltaic process is broken down into the following simple steps:

- Solar radiation is absorbed by the silicon photovoltaic solar cell.
- When the sun's rays collide with the silicon cell, electrons begin to move, resulting in an electric current flow.
- Wires collect and transmit direct current (DC) electricity to a solar inverter, which converts it to alternating current (AC).

2.1.5 Characteristics of PV Panel

To understand the characteristics of PV panel we have to discuss about the single diode model, I- V characteristics and the electrical characteristics of PV array. A short explanation of all these things are given below –

Firstly, the one-diode model, known as the five-parameter model, is one of the most common solar cell models. The Model includes a combination of a controlled current source photogenerated IPH, a diode described as an exponential single-source Shockley equation, a shunt resistance Rsh, and a series of power loss modelling RS. This model is given the equivalent circuit is given below

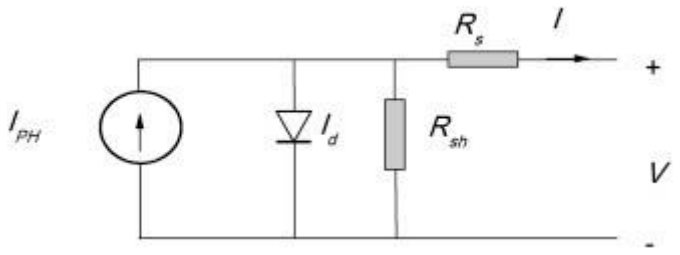


Figure 2.2: An equivalent circuit representing the five-parameter model of a solar cell.

Secondly, I-V Characteristics of a Solar Cell Curves are primarily a graphical representation of the operation of a photovoltaic cell or module, summarizing the connection between current and voltage under the current irradiance and temperature conditions. I-V curves provide the information required to assemble a system so that it operates as close to its best peak electrical outlet (MPP) as possible. The graphical representation of I-V Characteristics Curve is given below:

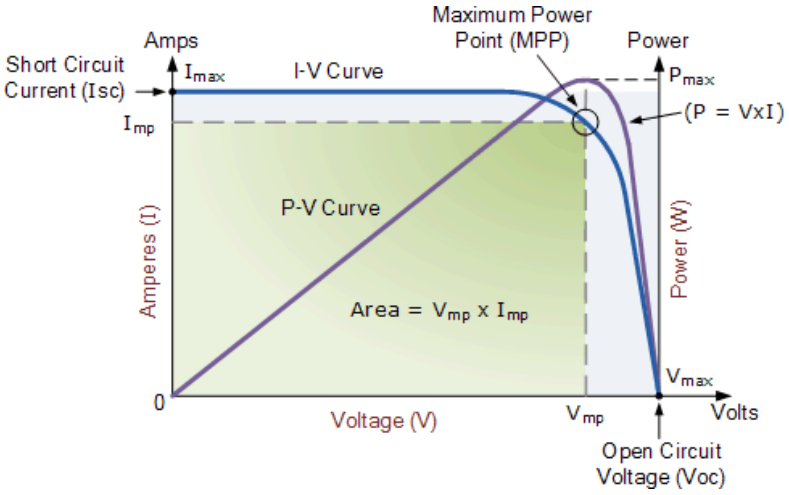


Figure 2.3: I-V Characteristics Curve [7].

The current-voltage (I-V) characteristics of a typical atomic number 14 PV cell operating under traditional conditions are shown in the graph above. The product of an electric cell's output current and voltage is the service provided by one electric cell or panel (I x V). The power curve above is achieved for a particular radiation level if the multiplication is performed point by point for all voltages from short-circuit to open-circuit circumstances. Finally, the link between output current and voltage summarizes a photovoltaic system's electrical attributes. The solar radiation quantity and intensity (solar radiation) regulates the output current (I) and the photovoltaic system cell operating temperature. The Solar cell's IV characteristics curves, which describe the current-voltage relation, are normally provided by the panel manufacturer and indicate the following -

Open-Circuit Voltage –Voc

This is the utmost voltage that the array provides once the terminals don't seem to be connected to any load. This value depends upon the amount of PV panels connected along in series [7].

Short-Circuit Current – I_{SC}

The maximum current provided by the PV array once the output connectors are shorted along (a tangency condition) [7].

Maximum Power Point –MPP

This relates to the purpose wherever the ability equipped by the array that's connected to the load (batteries, inverters) is at its maximum value, where MPP = Imp x Vmp.

the utmost wall plugs of electrical photovoltaic array is measured in Watts (W) or peak Watts (Wp) [7].

Fill Factor –FF

The fill factor is that the relationship between the most power that the array will truly offer below traditional operational conditions and also the product of the open-circuit voltage increased by the short-circuit current. This fill factor price provides a thought of the standard of the array and the nearer the fill factor is to one (unity), the additional power the array can provide [7].

Percent Efficiency –%eff

The efficiency is the ratio of the maximum electrical power that the array can generate compared to the amount of solar radiation that reaches the array. The efficiency of a typical solar system is usually 10-12%, depending on the type of cell used [7].

2.2 Effects of Weather Parameters on PV Panel

In this section we are going to talk about the effects of weather parameters on PV panel and power generation of PV panel. We already know that PV panel are sensitive towards many weather parameters such as Wind Speed, Humidity, Temperature, Air Pressure, Solar Irradiance etc. In our forecasting process we considered Wind Speed, Humidity, Temperature and Air Pressure parameters and we will try to discuss briefly about how these parameters effect the PV module and its generation on the following sub-sections of this section.

2.2.1 Wind Speed

Wind speed has less of an impact on the performance of solar modules. Typically, the solar module is positioned at an angle that corresponds to the area's latitude. Numerous studies have shown that when wind blows over a tilted solar panel, it exerts an uneven pressure on the panel. It has two sides. As a result, the solar module surface experiences drag in the direction of the wind flow and lift in the direction perpendicular to the wind flow on the other side. As a result, torque is generated. Although it is dishonest to claim that wind velocity has a direct impact on the efficiency of solar modules, it does play a significant influence in PV generation. When the wind does not add any more oomph to the sunlight rays while powering panels, the wind's action acts as a boost in solar efficiency. Because of the science behind the generation of power, if the surface of a solar panel gets too hot, the efficiency drops. However, the efficacy of the solar panel improves with a cooler panel. The efficiency of a colder solar panel, on the other hand, increases. The effect of wind 26 on solar cell performance can be summarized as follows: cooler panels enable more energy to pass through as an electric current than heated panels. The temperature of the solar cell decreases while the wind blows. The wind cools the solar modules, resulting in less electron oscillation, allowing the electrons to move more freely. As you get to

the higher state, you'll be able to transmit more energy. Solar modules that have been cooled by one degree Celsius become 0.05 percent more productive over time. Over time, this rate accumulates.

2.2.2 Temperature

Solar cells are sensitive to temperature fluctuations and are made of semiconductors. When the temperature of a semiconductor is increased, the bandgap shrinks, affecting the majority of the semiconductor material properties. The energy of the electrons in the material rises by a factor of two. As the temperature rises, the bandgap of a semiconductor shrinks. As a result, a reduced amount of energy is required to break the bond. A decrease in bond energy reduces the bandgap in the bond model of a semiconductor bandgap. As a result, increasing the temperature reduces a semiconductor's bandgap. The short circuit current and the open-circuit voltage are the two characteristics in a solar module, and the open-circuit voltage is the one that is most impacted by temperature.

The open-circuit voltage is the parameter in a solar cell that is most impacted by temperature changes. The effect of rising temperatures is depicted in the diagram below.

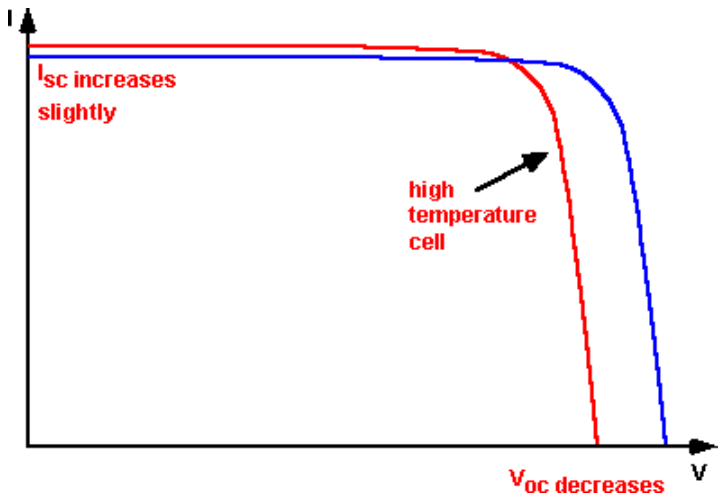


Figure 2.4: The effect of temperature on the IV characteristics of a solar cell [8].

2.2.3 Humidity

The amount of moisture or water vapor in the air is referred to as humidity. The gaseous condition of water, known as water vapor, is often invisible to the naked eye. The presence of precipitation, dew, or fog is indicated by the humidity level. When there is a lot of water vapor in the air, the humidity level is high, which indicates that rain is on the way. Summer humidity averages 76 percent, while winter humidity averages 74 percent. Absolute humidity, relative humidity, and specific humidity are the three different types of humidity measurements. The relative humidity method was used to measure the humidity in this experiment. Humidity has an effect on the solar module as well. As the relative humidity decreases, the output 25 short circuit current increases while the open-circuit voltage decreases. If the relative humidity is low, which indicates that there is little water vapor in the air, the solar flux is larger, which raises the short circuit current. As a result, it is possible to conclude that low relative humidity is preferable for solar module efficiency.

2.2.4 Air Pressure

A huge volume of air in the atmosphere that is relatively uniform in temperature and moisture is referred to as an air mass. Air masses can span thousands of kilometers in either direction and can reach up to 16 kilometers into the atmosphere from ground level. For the sake of simplicity, air mass is a measurement of how much atmosphere the sun's rays must pass through on their route to the earth's surface. Because light beams are absorbed and scattered by particles in the atmosphere, the solar energy retrieved is lower than intended.

In this chapter we tried to shortly but briefly tried to explain the theory related to this study. Here we explained about PV Panel, its working mechanism, characteristics and how weather effects the PV Panel, its generation and how it also effects the efficiency of PV Panel.

Chapter 3

Data Analysis Method

3.1 Random Forest

RF is a machine learning algorithm is an ensemble method. RF improves learning performance with a voting system given a set number of decision trees. RF exhibits the functionality of random selection random selection of features, bootstrap sampling, and decision tree making. The algorithm blends two main approaches: the bagging method and the random sub-space methodology. RF is typically used for classification, regression and clustering. These attributes are great for random forests for predicting PV power generation. PV power generation is easily affected by weather parameters. Usually the data series contain a lot of noise. These noises may reduce the generalization ability of the model. After inputting data samples, RF model will first extract some of the samples by bootstrap sampling, and then randomly select the features of these samples. These two random sampling phases make RF tolerant to outliers and noise and decrease the probability of over fitting. In this thesis, out of two algorithms RF regression algorithm also used to establish the forecast model [9].

3.1.1 Bootstrap approach

Bootstrapping is a statistical re-sampling approach in which a dataset is randomly sampled and replaced. It's frequently used to estimate the level of uncertainty in a machine learning model. The goal is to sample data from the original training set many times with replacement to create numerous independent training sets [10]. These are then utilized to lower the variance of "ensemble" methods predictions, considerably boosting their predictive performance. For estimating statistical quantities from samples, the bootstrap approach is used. Bootstrapping is also useful for small data sets that are prone to over-fitting. The bootstrap approach can be used

to test a solution's stability. It can improve robustness by using several sample data sets and then evaluating multiple models [18].

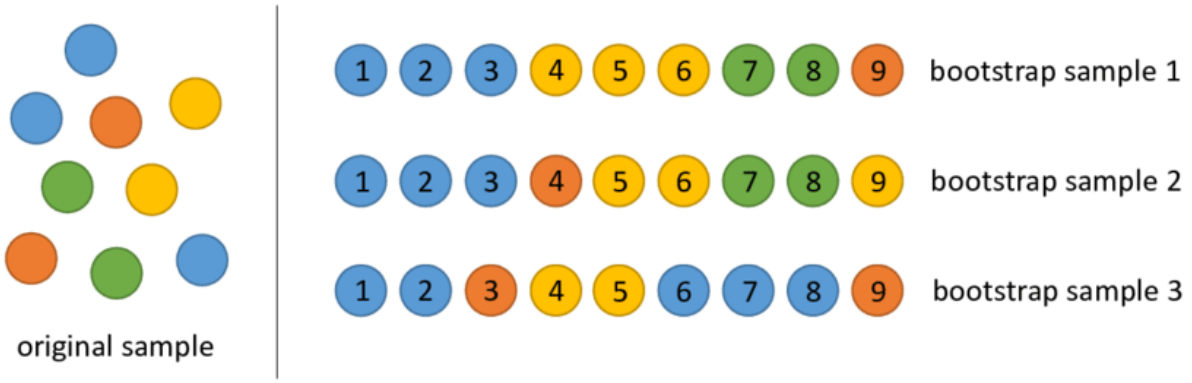


Figure 3.1: Bootstrap sampling approach [18].

3.1.2 Random Subspace

Random Subspace Ensemble is a ML algorithm that combines predictions of many decision trees that have been trained in the training dataset on distinct column subsets. Randomly differentiating the columns that are utilized to train every contributing member of the set has the effect of generating variety into the ensemble. Moreover, it can lift performance over using a single decision tree. It relates to other sets of decision-making trees such as 'bagging,' which builds trees with varied samples of rows from the training data set and a random forest which blends bagging principles with the random subspace ensemble. Random subsets of input characteristics can be selected to define random subspaces. This may be used as the foundation for an ensemble learning approach in which a model can fit into every random subspace [13].

3.1.3 Bagging

It is a form of ensemble learning algorithm called bootstrap aggregation or bagging. An ensemble method is a methodology which integrates predictions from many machine

algorithms to generate predictions more precise than any model. Bootstrap Aggregation is a generic approach which can be used to minimize the variance for high-variance algorithms. Decision trees are an algorithm of high variance, such as classification and regression trees [16], [17].

3.1.4 Decision tree

A decision tree is a supervised machine learning technique that can be used to solve problems in classification and regression. A decision tree is nothing more than a set of consecutive decisions that lead to a certain outcome. The aim is to construct a model that can predict a target variable value given a collection of input variables. Each leaf of the tree is labeled with a class or a probability distribution over the classes, indicating that the data set has been classified by the tree into one of the classes or a probability distribution over the classes [11]. Here Each element of the categorization domain is referred to as a class. Moreover, Trees respond to consecutive questions by leading us down a certain branch of the tree based on the response. It divides data into branches until a threshold unit is reached. Decision trees used in data mining are of two main types: a) Classification tree analysis (the class to which the data belongs is the projected outcome), b) Regression tree analysis (the expected result can be regarded as a real number) [15], [20].

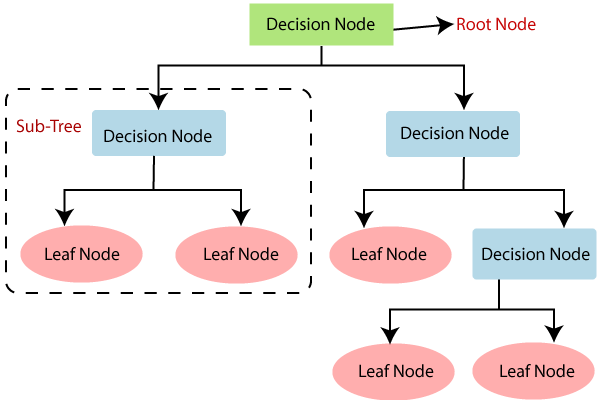


Figure 3.2: Decision Tree [20].

3.1.5 Steps of RF regression algorithm

1. Choose samples using the Bootstrap approach and use them as a training set.

2. In the set, start by growing a tree.

3. Calculate the optimum node division of the original tree by its characteristics.

4. Split the nodes until all of the samples are of the same type.

5. Assemble all of the trees into a forest, and then use the mean value of each tree's data to determine the forest's final prediction.

6. Model parameters must be determined before the model can be built [9].

3.1.6 RF regression parameters

1. Number of estimators or the number of trees in the forest.

2. Criterion index.

3. Max features. When looking for the best node, a function is chosen to find the best number of features.

There are three option available: a) Original value, which corresponds to the auto function. b) Square root of original value, corresponding to sqrt [9].

c) The logarithm of original value, corresponding to log2.

3.1.7 Model parameters

3.1.7.1 Criterion

The use of a criterion is frequently used to organize data. The criterion must meet two basic requirements.

1. Performance: When the natural boundaries of the data are well established, the resulting partition must fall along them.

2. Efficiency: There must be a fast algorithm for determining the best partition.

Here in RF this function is used to determine the quality of the split [12].

3.1.8 Performance Evaluation Indicators

Error analysis indicators are needed to understand the model's performance. In this method we used MSE. MSE represents as the mean of the square error.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - yi)^2 \quad (3.1)$$

Here $\hat{y}i$ is the prediction and yi is the real value. The predicted values range between (-10,000 to 10,000). The MSE loss (Y-axis) reaches its minimum value at prediction (X-axis) = 100. The range is 0 to ∞ [14].

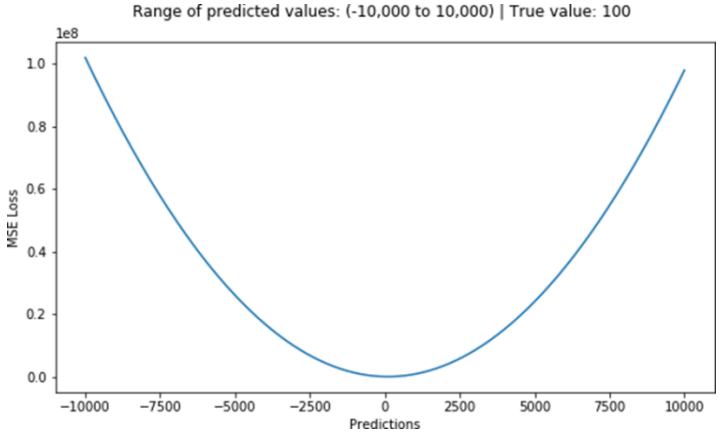


Figure 3.3: MSE Loss [14].

3.2 Data Analysis Method (ANN)

3.2.1 Artificial Neural Network

Artificial neural networks use in various research sectors, including education, engineering, medicine, and social applications. In the training phase, learning approaches entail adjusting the weights associated with each neuron. The network can generalize the result to execute the unseen data after the weights have been modified. An artificial neural network (ANN) is a unique process that may change its structural properties in response to the information it processes. It conducts by altering the connections weight. Each chain carries a certain weight, and the transmission between two neurons controlled by a weight, which is a number. Weights tweak for increasing the effectiveness of the results [21].

Artificial Neural Networks (ANN) are the foremost complex and delicate deep learning algorithms currently available within the Artificial Intelligence world. The basis of deep learning algorithms, artificial neural networks, and the cutting edge of machine intelligence are among the most revolutionary innovations of the recent decade. The concept of artificial neural networks has been around for a long time. Still, it is only recently that its use and implementation has become a reality in the artificial intelligence industry. Artificial neural networks primarily heavily influenced from biological neural systems.

Biological neural networks, on the other hand, are well-known for modifying the brain to great quantity of information in complex ways. The biological neural network of the brain, the fundamental process unit of the brain, is made up of about a hundred billion neurons.

Synapses are the vast connections between neurons that allow them to carry out their activities. There are 100 trillion synapses in the human brain, with 1,000 synapses per neuron. Electrical currents and chemical reactions run through a monumental number of

neurons in each brain function [22]. In ANN, their crucial thing is their neurons, and it plays an essential role in making decisions in this procedure. Neuron is the main component of the ANN method. Each neuron receives input from some of the other neurons, multiplies it by weights assigned to it, keeps adding it and afterwards transmits the portion to one or even more neurons. Prior to actually passing the output towards the next variable. If we put many thousands of neurons in numerous layers and stack them up to date on any alternative, we will acquire an artificial neural network that carries out complex jobs, including categorizing images or detecting speech and weather forecasts etc. A simple ANN model is presenting below:

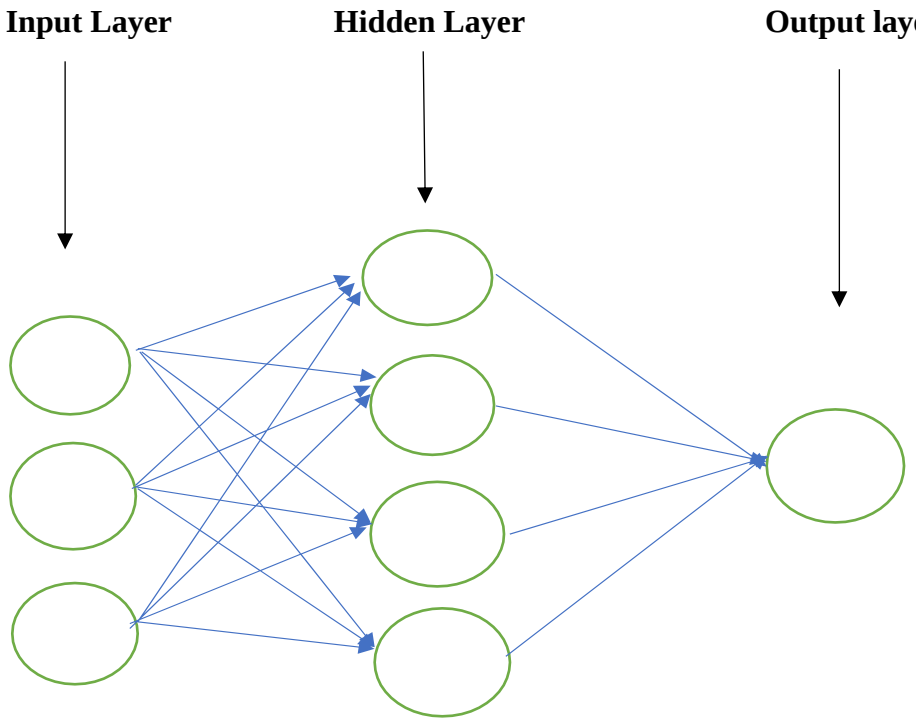


Figure 3.4: Basic Artificial Neural Network Model.

Basically, Artificial Neural Network has three different layers in their model. one is the input layer in the middle. There is the hidden layer which is the essential layer for this method. Lastly, there is an output layer. In the Input Layer here, it consumes or receives Data from

any given source, for examples, images, data files, hardware sensors, microphone, weather parameters reading data etc. In the hidden layer, there can be one or can be multiple layers in them. In this layer, data process according to the models and give output or numerous outputs depend on the functions of the network in the Output Layer.

In the Artificial Neural Network two types of things happen in the different layers. Firstly, when we give the data to the input layer Afterwards it trains the data. Usually, the weight of the neuronal networks in artificial neural networks initially allocate random values. Tweaking these weights to the appropriate numbers is pivotal for the ANN to accomplish its statutory duties correctly and efficiently. However, determining the suitable weights is difficult, mainly when dealing with numerous layers and thousands of neural connections. Furthermore, with these Machine Learning Algorithms, we essentially supply some data to train the model and then supply 10% to 20% testing data to evaluate the accuracy of our model predictions. During training, the forecasting model inherently extracts specific patterns from the input. Each layer identifies a unique class of attributes when the model has trained with elevated samples. We use a well-trained neural network to input certain data, the modified weights of neurons can extract the right properties and decide exactly which class the data is in. In this process basically it trains the data and predict the outcomes [12].

3.2.2 Advantages and Disadvantages

One of the subdivisions of artificial intelligence is that machine learning that develops behavioral patterns based on experience utilizes artificial neural networks as multiple methods. Many distinct machine learning models may be used to diagnose patterns in data and perform tasks such as categorization and prediction. On the other hand, Artificial Neural networks exceeded other algorithms while dealing with diverse and structured inputs, such

visuals, photos, sounds, data, etc. ANN and algorithms for deep learning would not need any training which is under strict observation. instead, they automatically retrieve features from input data when adequately trained. Information is processed stored on the entire network, not really in a database, as it does in traditional programming. The network does not cease to operate despite losing a few bits of information at one location. Even with limited information, after ANN training, the data may yield output. The significance of the missing data determines the performance loss here. The ANN will continue to generate output although if one or more cells are erroneous. This feature allows networks to be fault-tolerant. Through remarking on similar events, artificial neural networks learn events and make predictions. Artificial neural networks have the computational power to accomplish several tasks at once. Apart all this advantages this model has some disadvantages also. It is required to identify the examples and to train the network according to the desired outcome by delivering these samples to the network in order for it to learn. The networking success is proportional to the examples picked, and the network might provide false information if it cannot be shown to the network in all of its dimensions. Due to its structure, artificial neural networks demand parallel processing capabilities for hardware and software. As a result, the appearance of the equipment is dependent. ANN does not explain why or how to produce a test solution for this reason the confidence of the network is damaged. The structure of artificial neural networks is not determined by any precise rule. It is possible to build a correct network structure by doing a trial and error process [23].

3.2.3 Reason for selecting ANN

Our mission is to predict the weather by the info we get from the photovoltaic Modules. We decide Artificial Neural Network Model. The rationale for selecting an Artificial neural network is that it will simply handle untidy and unstructured information. By constructing a

model, if we train some data, it may also predict a decent output by process on their hidden layers. If there are any Corrupted information or missing information, ANN will still ignore those losses and predict better output and also, the model slows over time, suggesting that it does not simply run slow from the beginning of the process. Overall, these are the rationale we decide Artificial Neural Networks as our data analysis methodology.

3.3 Data Analysis Method (Multiple Linear Regression Model)

3.3.1 Linear Regression

Linear regression is a statistical method to find out the relationship between two variables. This relationship is established by fitting a linear equation to observed data. One variable is considered to be an independent variable, and the another one is considered as a dependent variable. Before using linear regression model, user should first assess whether or not there is a relationship between the dependent and independent variable. This does not necessarily mean that one variable causes the other, but rather that the two variables have a significant relationship. If the proposed independent and dependent variables appear to have no relationship, fitting a linear regression model to the data is unlikely to yield a useful model. Correlation coefficient is a valuable numerical measure of association between two variables. The strength of the linear relation between two variables is measured by the correlation coefficient [31]. Formula for computing the correlation coefficient for a given set of observations (x_1, y_1), (x_2, y_2), … (x_n, y_n) is

$$r = \frac{1}{n-1} \Sigma \left(\frac{x-\bar{x}}{S_x}\right)\left(\frac{y-\bar{y}}{S_y}\right) \tag{3.2}$$

Here,

r = correlation coefficient

x = Observed abscissa.

$\overline{x}$ = Average of all the observed abscissas.

y = Observed ordinate.

$\overline{y}$ = Average of all the observed ordinates.

s_x = Summation of all the abscissas.

s_y = Summation of all the ordinates.

The limit of correlation coefficient always remains between -1 and 1, where 1 or -1 indicting perfect correlation. A positive correlation means that when one variable increases in values the other corresponding variable will also increase, while a negative correlation means increasing values in one variable decreases the other corresponding variable in values. Correlation value close to 0 implies that there is no relation between variables [31]. Formula for calculating the correlation coefficient standardizes the variables that is why changes in scale or units of measurement have no effect on its value. As a result, when determining the strength of a relation between two variables, the correlation coefficient is usually more useful than a graphical representation

$$y = \beta x + \beta_0 \tag{3.3}$$

where *x* is the independent variable and y is the dependent variable. β, is the slope of the line and β_0 is the intercept (the value of y when x = 0).

3.3.2 Multiple Linear Regression

Multiple linear regression is an extension of linear regression model. Multiple linear regression is a model that finds out relationship between two or more independent variables

and a dependent variable by fitting a linear equation to observed data. Value of every independent variable (x) is correlated with the dependent variable (y). The population regression line for *p* independent variables $x_1, x_2, \ldots\ldots, x_p$ is defined to be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon \quad (3.4)$$

Here,

β_0 = intercept

x_i = independent variable

β_i = parameter

y = dependent variable

ε = error

After assuming the error term zero, the equation of Multiple Linear Regression would be

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots. + \beta_p x_p \quad (3.5)$$

Multiple Linear Regression take sample data and estimate the dependent variable, $\hat{y}$. So, estimated multiple regression equation would be

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ ... + b_p x_p \quad (3.6)$$

$b_0, b_1, b_2, \ldots. b_p$ are the estimates of $\beta_0, \beta_1, \beta_2 \ldots. \beta_p$ and $\hat{y}$ is the predicted value of dependent variable.

Estimated multiple regression equation gives an indication of how the mean response $\hat{y}$ changes with the independent variables. [32].

3.3.3 Advantage

The most essential benefit of multiple linear regression is that it aids in the understanding of relationships between variables in a dataset. So, this method includes a more exact and precise understanding of the relationship between each individual aspect (independent variables) and the outcome (dependent variable).

3.3.4 Disadvantage

As multiple linear regression deals with several independent variables two problem may arise:

- Overfitting
- Multicollinearity

Overfitting occurs when too many independent variables are added to a model; they account for more variance but do not add anything to the model.

Multicollinearity happens when some of the independent variables are correlated with each other.

To avoid overfitting and multicollinearity independent variables are selected carefully.

3.3.5 Reason Behind Selection of Multiple Linear Regression Model

Multiple linear regression allows to create a relation between all potentially important factors within one model.

3.3.6 Data Description

The objective is to forecast Clean Module Short Circuit Current (mA) and Dusty Module Short Circuit Current (mA) by using multiple linear regression. For forecasting two different dependent variable multiple linear regression was ran several times separately.

In first case, the target variable is Clean Module Short Circuit Current (mA). There are 7 independent variables available from the given dataset that were used to produce forecasts. These are:

- Clean Module Temperature (℃)
- Dusty Module Temperature (℃)
- Wind Speed (m/s)
- Humidity (%)
- Air Pressure (hPa)
- Tr (Hour)
- T (Difference)

In second case, the target variable is Dusty Module Short Circuit Current (mA) and 7 independent variables that were used to produce forecasts are

- Clean Module Temperature (℃)
- Dusty Module Temperature (℃)
- Wind Speed (m/s)
- Humidity (%)
- Air Pressure (hPa)
- Tr (Hour)
- T (Difference)

Chapter 4

Software Setup and Methodology

4.1 Software Setup (Random Forest)

Google colab was used to test the program's code in order to implement the specified methodology. In the model algorithm r2_score function has been called from Sklearn module for predicting accuracy, which is written using the Python programming language [19]. At first, a notebook is created in Google colab and then import all the necessary libraries from Sklearn to use those libraries while creating the algorithm.

4.1.1 Methodology

4.1.1.1 Training Set and Testing Set

The training dataset is used to train the machine learning model, whereas the testing dataset is used to test its learning abilities. When a learning system is trained on a dataset that spans a long period of time, it is common for it to forecast more accurately.

- Training dataset 1: 1st November 2019 to 29th November 2019.
- Training dataset 2: 1st November 2019 to 29th December 2019.
- Training dataset 3: 1st November 2019 to 30th January 2020.
- Training dataset 4: 1st November 2019 to 26th February 2020.
- Training dataset 5: 1st November 2019 to 30th March 2020.

The data from each month's last days is used to create the testing dataset

- Testing dataset 1: 29th November 2019.
- Testing dataset 2: 31st December 2019.

- Testing dataset 3: 31st January 2020.
- Testing dataset 4: 27th February 2020
- Testing dataset 5: 31st March 2020.

along with five weather parameters:

- Temperature.
- Wind Speed.
- Humidity.
- Air Pressure.
- Time.

for each clean and dusty module.

4.1.1.2 Model Implementation

Random forest (RF) is a recently designed data collection and forecasting technology that balances decision trees to minimize the danger of over-fitting. When compared to other machine learning algorithms, RF has a significant advantage when dealing with large datasets. RF has proven to be successful in both solar radiation forecasting and PV power forecasting in relevant studies. Here the predicted output is short circuit current, Isc(mA) accuracy for clean and dusty PV panel. The loss function is used for this model is 'mean_square_error' (MSE) and to get the expected prediction a minimum error is use as a base, The predicted values range between -10,000 to 10,000. Following that, the training and testing datasets are imported [14]. The data framing phase begins after the training and testing datasets have been imported. Here in this data framing, x_tarin and y_train data frames are created from training dataset. x_train contains all the weather parameters (such

as Wind Speed, Temperature, Humidity, Air Pressure, Time) and y_train contains the short circuit current (mA) for both clean and dusty modules.

Moreover, x_test is created with all the testing data sets parameters and y_test contains the short circuit current (mA) for both clean and dusty modules.

Then the model is created and then it is trained based on the x_train and y_train datasets. This model contains n estimator=1000 (The number of Decision trees) and the loss function is mean_square_error' (MSE). After training the model based on training datasets, a short circuit current is predicted for each month, for the last day of each month (data of five months) for both clean and dusty module.

Furthermore, calculate the accuracy for each month for both clean and dusty module. Afterwards compare this accuracy with the other two models.

On the next page, there is a flowchart of the model's operation.

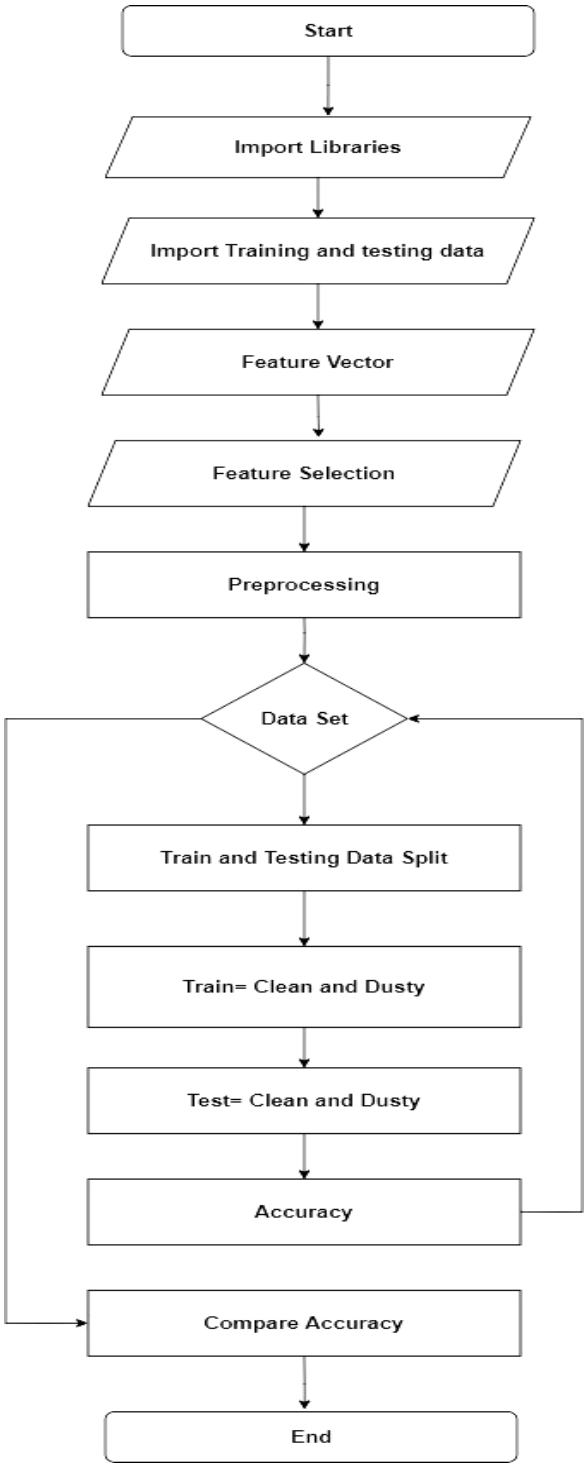


Figure 4.1: Flowchart of the procedure of Random Forest.

4.1.2 Software Setup (ANN)

To Develop Our Artificial Neural Network Model, we choose Goggle Colab and Python as coding Language to perform Comparative Data Analysis and PV system Considering Weather Parameters. We choose Google Colab instead of other device installed software because Colab has its personal Ram and Disk space which run faster and exclude other difficulties like ram or disk space shortage. Google Colab has personal integrated 12.69 GB ram and 107.77 GB of Disk space to run any program.

Here we import Keras in TensorfFlow library. TensorFlow is a machine learning platform that runs from inception to delivery and is open-source. It's an extensive and adaptable environment of tools, libraries, and other resources that provide high-level APIs for processes. The framework provides many layers of insights from which we may decide to construct and implement machine learning models. TensorFlow allows construction and train models at different levels of abstraction [24]. Whereas Keras is a Python-based deep learning API that runs on top of the TensorFlow machine learning framework. It had been developed with the intent of allowing for quick experimentation. It is essential to get it from conception to outcome as quickly as feasible when conducting research. Keras is a robust interface for handling significant machine learning problems, emphasizing recent artificial neural networks. It provides fundamental abstractions and building elements for designing and releasing elevated machine learning systems. Keras enables engineers and researchers to utilize TensorFlow's scalability and cross-platform features effectively [25]. Here We import Keras sequential architecture. The Sequential constructor, like any other layer or model in Keras, takes a name parameter. This is handy for giving functionally relevant names to Tensor Board graphs.

The Adam Algorithm is implemented by importing optimizer. Adaptive moment estimate is Adam's full Form. Adam needs comparatively low memory for the processing and also works fine even with relatively small parameters modification. We use an activation function right before deciding what the activation value should be during the calculations of the values for activations in each layer. We calculate a value for each activation in the next layer based on the previous activations, weights, and biases in each layer. However, we use an activation function to scale the value before sending it to the activations of the next layer. We import Leakyrelu as an activator and import some layer like Dropout and Dense that we can construct our model afterwards. Here Relu Activator is built in to Dense Layer. The Rectified Linear Unit is full form of Relu. In an Artificial neural network, this is the most frequently implemented activation function. If the function obtains any negative input, it delivers 0; however, if the function receives any positive value x, it provides that value back. It can be written as f(x) = max (0, x). graphically it represents as:

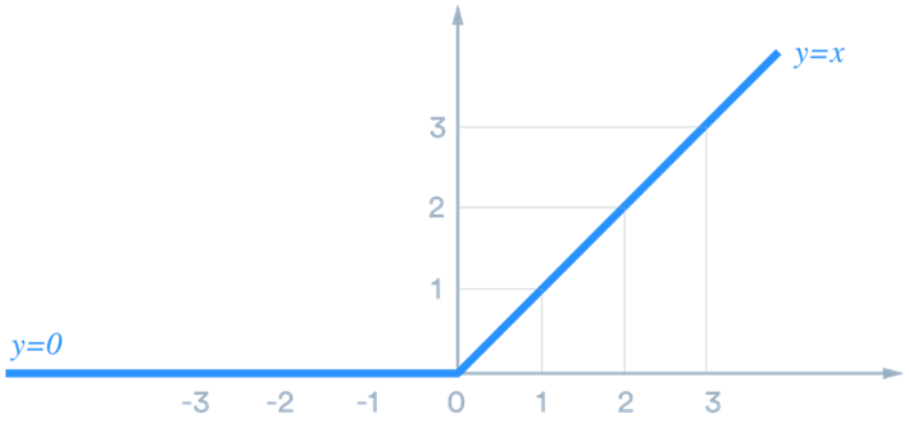


Figure 4.2: Graphical look of Relu Function.

It is very simple, widely acceptable and works fine in most of the models. It also helps our model at accounting for non-linarites and correlations [26]. LeakyRelu full form is Leaky Rectified Linear Unit. It is basically the activation function based on Relu Activation. However, instead of a smooth slope, it has a slight slope for negative values. The slope

coefficient is calculated just before training rather than being learned during training. This activation function is used in applications involving sparse gradients, such as training generative adversarial networks.

Function of LeakyRelu is F(x) = {0.01x for x< 0 and x for x => 0} [27].

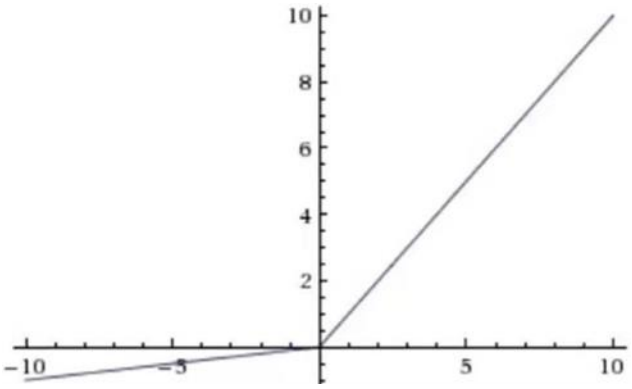


Figure 4.3: LeakyRelu graphical function.

Dropout is a layer we are using for developing our ANN model. Dropout is usually use for that our data should not over fit in our model. in this case if you do not add this layer it seems train our data well but when it comes for prediction it produces bad result for our model. Basically, The Dropout layer, which helps minimize overfitting, changes input units to 0 at random with a rated frequency at each step during training time. Inputs that are not set to 0 are scaled up by 1/ (1 - rate) such that the total sum of all inputs remains the same.

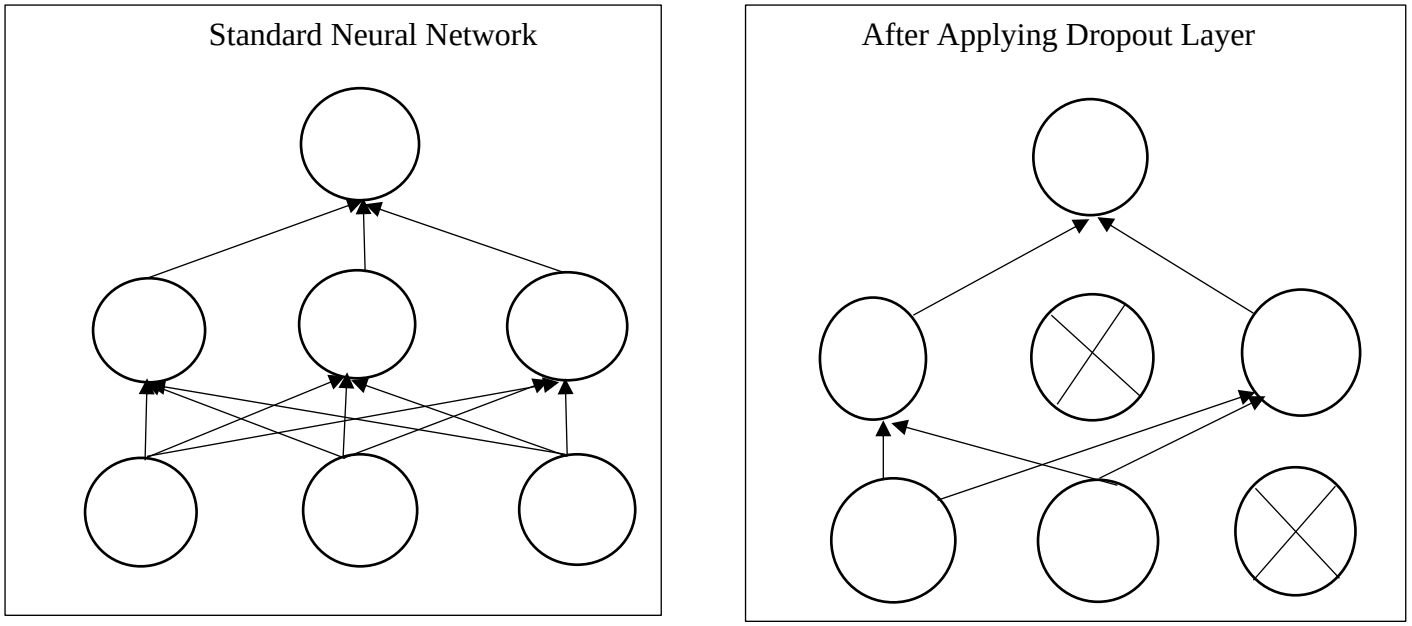


Figure 4.4: Working Figure of Dropout Layer

Here we use a layer named Scaled Exponential Linear Unit (SELU). The SELU activation function essentially multiplies the scale (>1) with the tf.keras.activation.elu output. The Scaled Exponential Linear Unit (SELU) activation function is defined as:

- if x > 0: return scale * x
- if x < 0: return scale * alpha * (exp(x) - 1)

Where pre-defined constants are alpha and scale (alpha=1.67326324, with scale=1.05070098) Alpha and scale values are selected such that the mean and variance of the inputs are retained across two subsequent layers, as long as weights are appropriately set and there is sufficient number of input units [28]. Choose Adam as Optimizer and Select Mean Squared Error as loss function. It Computes the mean of error squares between labels and predictions. Basically, it creates and define the mean squared error loss during the training epochs for both training and testing data. Here is the MSE Mathematical function:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - yi)^2 \quad (4.1)$$

Where n is the data point and Y is the vector of the variable observed values predicted and vector Y is the predicted values and n is the numbers of the training values [29]. Adam is an adaptive learning rate approach that calculates individual learning opportunities in various parameters. Its name is taken from adaptive moment estimation and is entitled because Adam utilizes first and second gradient assessments to modify learning rates for each weight of the neural network [30].

4.1.2.1 Flowchart for Software Setup of ANN Model

Here I am explaining the software setup process by the flowchart. Flowchart is given below:

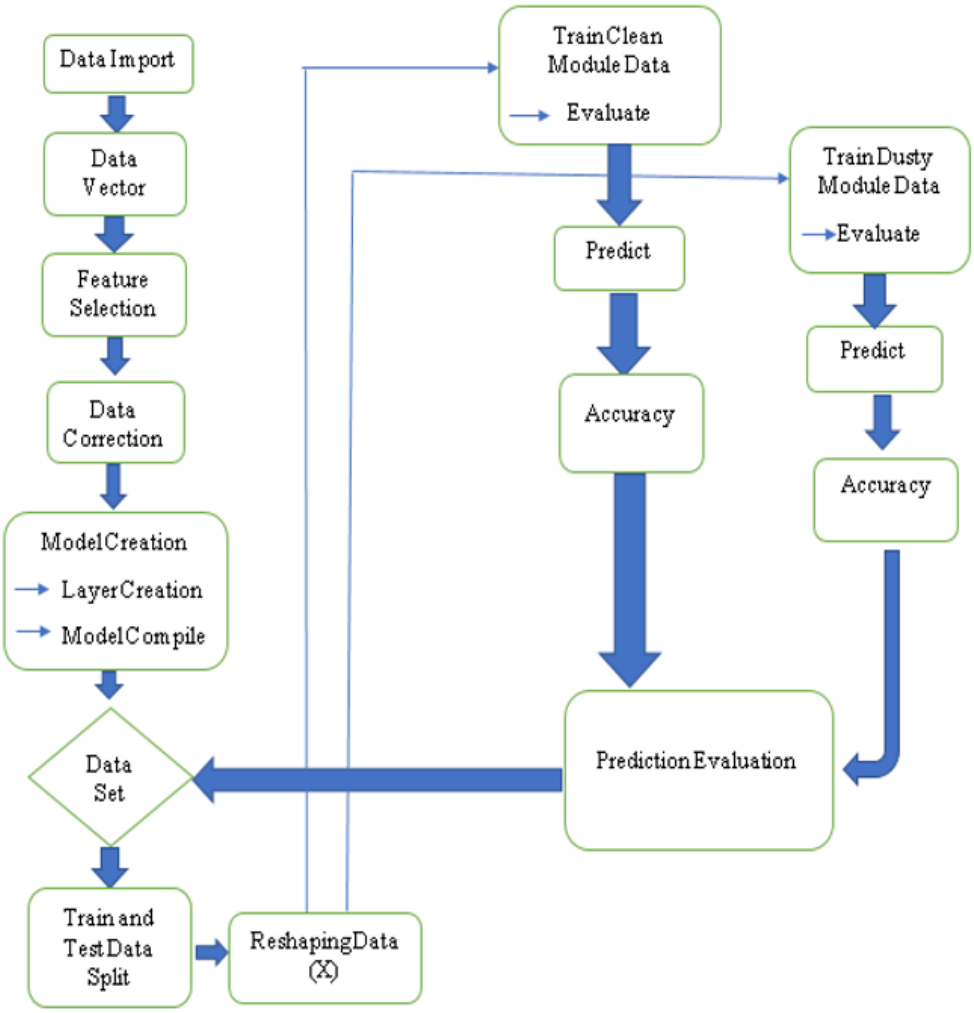


Figure 4.5: Flowchart of the Software Setup of ANN Model.

First of all, we import and mount the data. Then we enter the CSV file by vectorising the data. Then, when preparing data via Feature Selection, we choose the column for the X and Y axes. Then fix the data errors. Afterward, we design our prediction model and also define Layer and compile all The Layers. Prepare the train data set and test it in our model. Then the train and test data are partitioned. Then we transform the data we describe on the X-axis and provide the raw data of the Y-axis. Subsequently, we train the dusty and clean module data separately and then compare the accuracy with the test data we provide. Then the loop returned to the Data Set and continues till 5 data set is finished.

4.2.1 Methodology

4.2.1.1 Training Set and Testing Set

The testing dataset is intended to assess the machine learning model's learning skills, whereas the training dataset is used to train it. It is usual for a learning system to predict more correctly after being trained on a dataset that covers a significant period of time. Data Collection from the PV Module starting in November 2019 until 31st march. Here are five sets of training datasets are created:

- Training dataset 1: 1st November 2019 to 28th November 2019.
- Training dataset 2: 1st November 2019 to 30th December 2019.
- Training dataset 3: 1st November 2019 to 30th January 2020.
- Training dataset 4: 1st November 2019 to 26th February 2020.
- Training dataset 5: 1st November 2019 to 30th March 2020

To build the testing dataset, humidity, air pressure, wind speed, temperature of each Clean Module and Dusty Module, T(Difference), and the Time is utilized as input variables to create

the testing dataset. Here are five Testing Data Set giving below:

- Testing dataset 1: 29th November 2019.
- Testing dataset 2: 31st December 2019.
- Testing dataset 3: 31st January 2020.
- Testing dataset 4: 27th February 2020.
- Testing dataset 5: 31st March 2020

4.2.1.2 Full Process

Firstly, in the software install the tensor flow keras library in the google colab as its not preinstalled in google colab. After that import Sequential model which is basically an architecture of Artificial Neural Network. For building the Neural Network Model import the dense layer where relu activation is present in-built format also import dropout Layer. Import Adam as optimizer and Leaky Relu as another activation. Import pyplot as plot for plotting all the graph which will be generated from our prediction by our developed model Secondly Mount the google drive where we store all the data set we prepare and read all the CSV file and define the folder path where all the dataset is stored and before defining the train and test Data frame we define the name of the train and test file in the same manner of the csv file where they read all the train and test data and getting into this into a for loop which continues to read until all the data set is completed. Thirdly Print all the train and testing Data, to see that all are in good shape or not or is there any unusual error in the data frame or not. After all this thing we run a loop where if there is any string value in the column of Air Pressure (hPa) and Clean Module

Temperature (c), it converts that string value into float number. Fourthly, we develop our model. Model is presenting in block diagram below

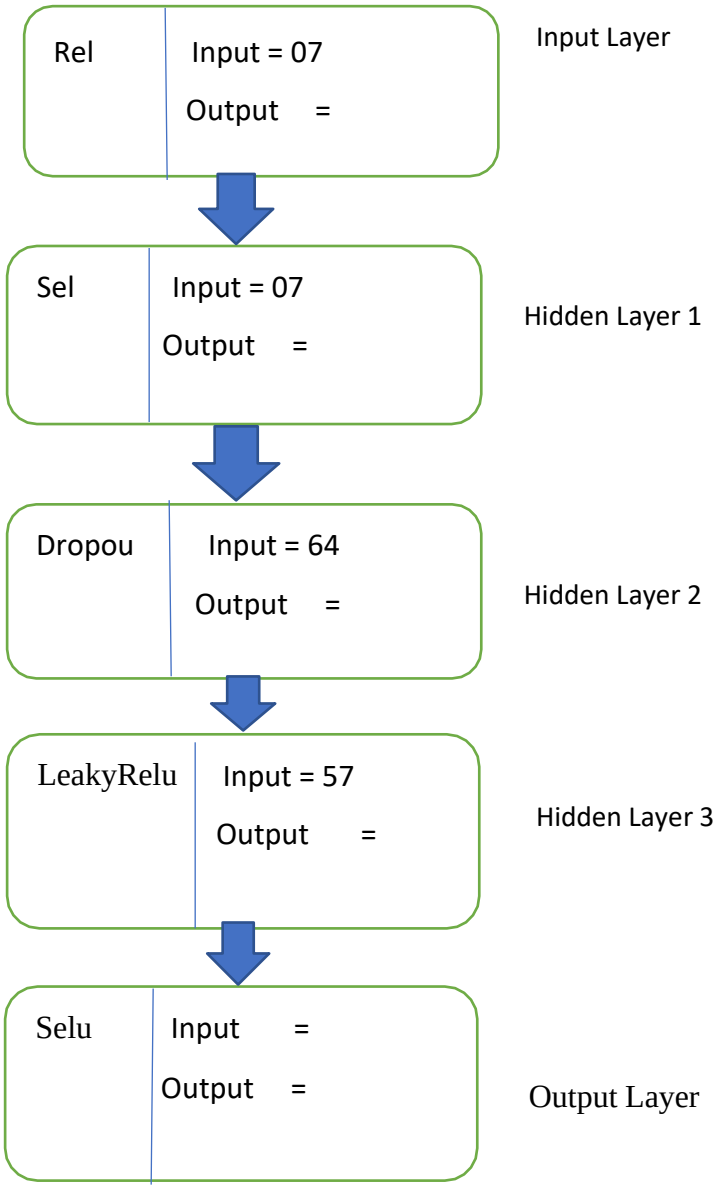


Figure 4.6: Sequential Model Block Diagram.

Here is the Basic Graphical Presentation of our Artificial Neural Network model which is also called as Deep Neural Network for its multiple Hidden Layers:

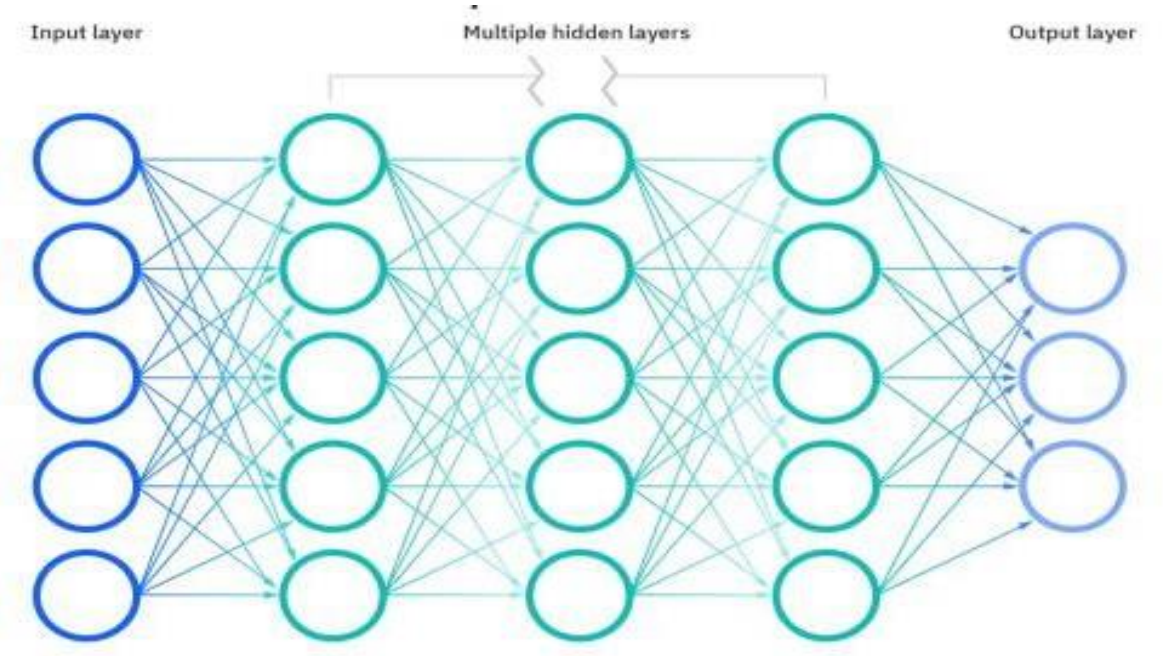


Figure 4.7: Developed Neural Network Model.

In our Developed Model There is One Input layer where our input data is 7 parameters which are: Clean Module Temperature (°C), Wind Speed (m/s), Humidity (%), Air Pressure (hPa), Tr (Hour) and T (Difference), Dusty Module Temperature (°C) and on that layer there is Relu activation incorporate with that layer and provide 7 neurons as same as input and after processing all this it sends it to second layer which is basically Hidden Layer 1. Selu is incorporate with that layer where input layer 7 parameter as output would be the input of this layer and on that Selu layer we provide 64 neurons with individual weight and which are all connected among each and every output of the first layer. Hidden layer 2 is basically Dropout Layer where data from the second layer will refine or drop up to 10% and send it to the Hidden

Layer 3. It actually helps our model not to over fit from our data. Hidden Layer 3 is Leaky Relu Activation where input is 57 after the dropout of data in hidden layer 2 and we provide the neuron number 128 on that layer with adjusted weights. They all are interconnected among all the neurons and after that it send it to the next layer which is Output Layer and the Layer which is also Selu activation layer and their input neuron number is 128 which is basically the output provided by Leaky Relu Activation Layer. After all the processing it gives an output in the output layer. For all the processing we use Adam as the Optimizer and Mean Squared Error as the Loss function. We set the metrics as accuracy means it can evaluate the result we get in output in terms of accuracy. We set epochs 100 and Batch size 100 that means it will take 100 Data in a single note and train it 100 times before make any result in this process it will continue in a loop until it completes all the data sets. Here we define the data frame name X train and X test and combine Clean Module Temperature (c)', 'Dusty Module Temperature (c)', 'Wind Speed (m/s)', 'Humidity (%)', 'Air Pressure (hPa)', 'Tr (Hour)', ' T (Difference)' all this parameter in them and will reshape or scaling by Min Max scaler and also define data frame name Y train clean, Y test clean for clean module and Y train dusty, Y test dusty for dusty module and declare Clean Module Short Circuit Current (mA) and Dusty Module Short Circuit Current (mA) accordingly but we don't scale this Y train or testing data we just sent the raw data to compare in the case. All this data split train and test in loop when i= o is the first Data set and When i will be 1 it will work for second data set, when i will be 2 it will work for third dataset, when i will be 3 it will work for fourth data set and lastly when i= 4, it will work for fifth data set for both train and testing Data set. After all this process we are separately train for clean module and the dusty module and after that our model predict the value in terms of accuracy in compare with the testing data what we provided. After that, we used the r2 score to calculate accuracy by comparing our predicted value to the actual value, and then we printed

all of the graphs and bar charts for accuracy in comparison of real and predicted values for both clean and dusty modules.

4.3 Software Setup and Methodology

4.3.1 Software Setup

The main objective was to forecast the probable output short circuit current, Isc(mA) for a specific day of the year by using machine learning model. Colaboratory was used for writing and executing code throughout the whole process. Colaboratory is a web-based IDE for python which runs fully on cloud. It has 12.69 GB dedicated RAM and 107.77 GB of disk space. Various libraries were imported to execute multiple linear regression. ‘panda’, ‘seaborn’, ‘matplotlib’and ‘sklearn’ are example of such libraries. Each and every library is used for executing the code smoothly. These libraries have some useful features.

4.3.1.1 Panda’s Features

- High level data structures, example: DataFrame.
- Works with tabular data and has a rich time series functionality.
- Can manipulate Numpy and SciPy functions

4.3.1.2 Seaborn

Seaborn is used for data visualization. One of the key feature of seaborn is generating heatmap. Heatmap is a clear visualization of how different variables affect each other. Furthermore, seaborn's functions are extremely powerful, capable of producing a variety of graphical plots

(histograms, pie charts, scatter plots, box plots, and so on) with only a few lines of code.

4.3.1.3 Matplotlibrary

Matplotlibrary library is used for plotting graphs. Line graph were generated by matplotlibrary. Matplotlibrary's pyplot function was utilized. With the help of pyplot (a set of functions) matplotlib behaves like MATLAB.

4.3.1.4 Sklearn

Sklearn is also known as Scikit-learn. Multiple linear regression model was built with Sklearn. R2_score function is used from sklearn library. R2_score function determines the accuracy of the modeling by setting side by side the predicted value with the original one [33].

After that, the model is put to the test to see if it can predict the output of short circuit current, Isc (mA), separately for Clean and Dusty Modules, using a probable dataset of humidity, temperature, wind speed, air pressure, and time of a specific day that is obviously different from its training dataset. The flowchart below depicts the entire approach for forecasting the short circuit current, Isc(mA), using the machine learning algorithm:

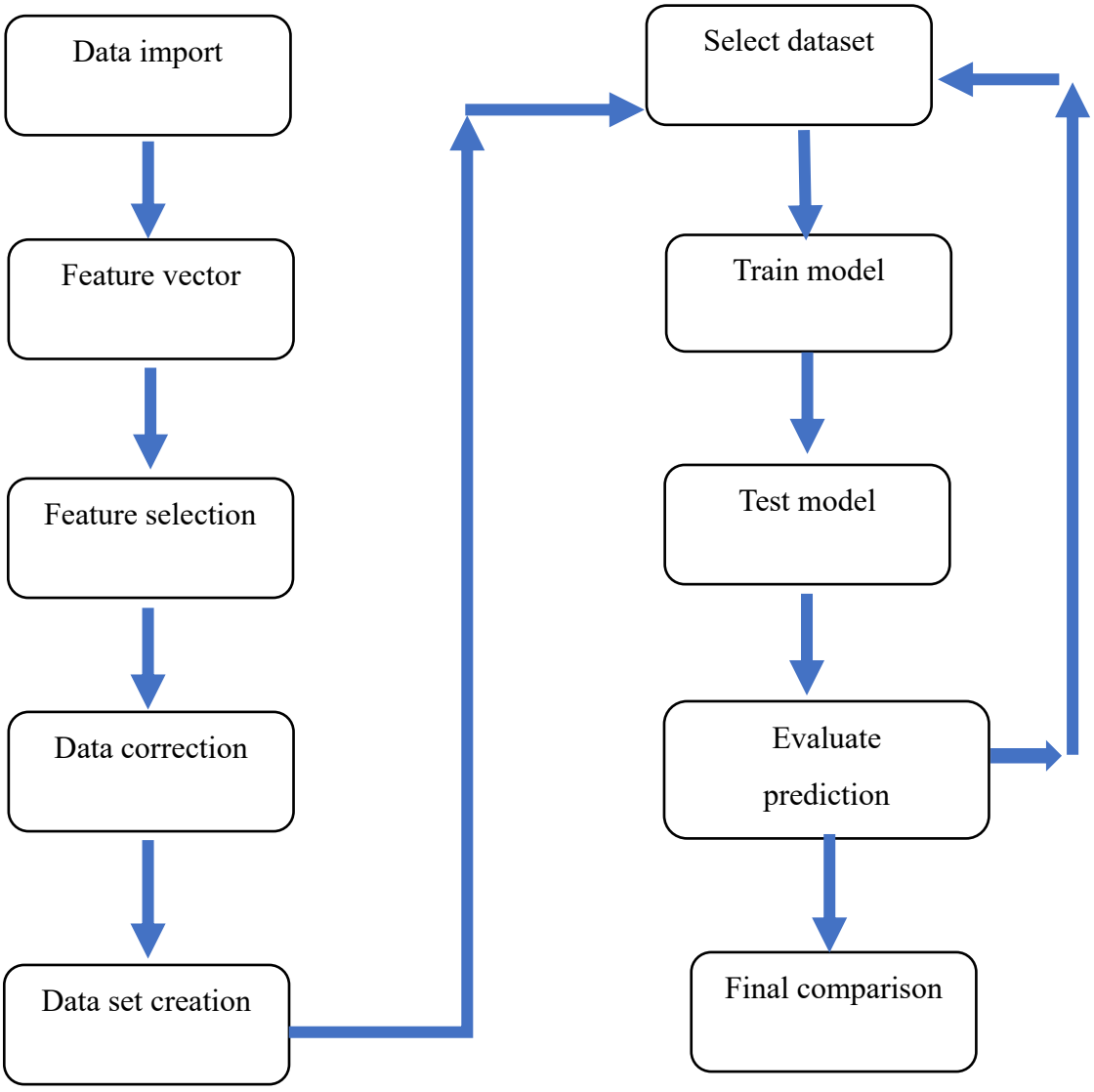


Figure 4.8: Flowchart of the working procedure of algorithm.

4.3.2 Methodology

4.3.2.1 Training Set and Testing Set

The training dataset is used to train the machine learning model, while the testing dataset is used to test its learning abilities. When a learning system is trained on a dataset that spans a long period of time, it is common for it to forecast more accurately. From November 2019 onwards, data from the sensors included in the hardware configuration will be collected.

As a result, five sets of training datasets are created:

- Training dataset 1: 1st November 2019 to 28th November 2019.
- Training dataset 2: 1st November 2019 to 30th December 2019.
- Training dataset 3: 1st November 2019 to 30th January 2020.
- Training dataset 4: 1st November 2019 to 26th February 2020.
- Training dataset 5: 1st November 2019 to 30th March 2020

Humidity, air pressure, wind speed, temperature of each Clean Module and Dusty Module, T(Difference) as well as the time of March 31st 2020 are used as input(independent) variables to create the testing dataset.

Five testing datasets are formed using data of input(independent) variables. They are:

- Testing dataset 1: 29th November 2019.
- Testing dataset 2: 31st December 2019.
- Testing dataset 3: 31st January 2020.
- Testing dataset 4: 27th February 2020.
- Testing dataset 5: 31st March 2020.

4.3.2.2 Prediction Analysis with different training dataset

The short circuit current, Isc (mA), of the 31st of March 2020 is predicted using five distinct training datasets. The short circuit current, Isc (mA), is predicted individually for Clean and Dusty Modules and shown against time in the x-axis and Isc (mA) in the y-axis.

At first given datasets were imported from google drive in csv format. Then all the five datasets (Training dataset 1, Training dataset 2, Training dataset 3, Training dataset 4, Training dataset 5) were taken accordingly. All the datasets were read by a variable named df_train, df_train merges all the five datasets into one csv file.

Following that is feature selection. For forecasting the short circuit current, Isc (mA) selected feature will be Clean Module Temperature (℃), Dusty Module Temperature (℃), Wind Speed (m/s), Humidity (%), Air Pressure (hPa), Tr (Hour)and T (Difference); these solar parameters will be used as independent variable and short circuit current, Isc (mA) for clean and dusty module will be considered as dependent variable. For training the model all the solar parameters of the dataset were splitted by two variables named X_train, y_train. Where X_train contains Clean Module Temperature (℃), Dusty Module Temperature (℃), Wind Speed (m/s), Humidity (%), Air Pressure (hPa), Tr (Hour)and T (Difference). Furthermore, y_train_clean contains 'Clean Module Short Circuit Current (mA)' for clean module and y_train_dusty contains 'Dusty Module Short Circuit Current (mA)' for dusty module. For testing purpose two more variables were created; X_test and y_test. Where X_test contains Clean Module Temperature (℃), Dusty Module Temperature (℃) , Wind Speed (m/s), Humidity (%), Air Pressure (hPa), Tr (Hour)and T (Difference) and y_test_clean contains 'Clean Module Short Circuit Current (mA)' for clean module Short circuit current, Isc forecasting and y_test_dusty contains 'Dusty Module Short Circuit Current (mA)' for dusty module short circuit current, Isc (mA) prediction.

After splitting training and test dataset all null values were deleted from dataset to get a clean dataset. Then LinearRegression() function used on the clean dataset to predict the short circuit current, Isc (mA). LinearRegression() function was ran two times. Firstly, linearRegression.predict(X_test) ran to predict short circuit current, Isc (mA) for clean module by using function linearRegression.fit(X_train, y_train_clean).Then, linearRegression.predict(X_test) ran to predict short circuit current, Isc (mA) for dusty module by using function linearRegression.fit(X_train, y_train_dusty).Afterwards, 'r2_score'was used to calculate the accuracy of predicted short circuit current, Isc(mA)

Chapter 5

Result Analysis

5.1 Steps of Analysis

In first section, prediction for both clean solar panel and dusty solar panel short circuit current Isc (mA) will be performed based on dataset from 1st November to 28th November (training dataset 1) to predict the 29th November (testing dataset 1) short circuit current, Isc(mA). Parameters will be used in this segment is Clean Module Temperature (℃), Dusty Module Temperature (℃), Wind Speed (m/s), Humidity (%), Air Pressure (hPa), Tr (Hour) and T (Difference)

In second section, prediction for both clean solar panel and dusty solar panel short circuit current Isc(mA) will be performed based on dataset from 1st December to 30th December (training dataset 2) to predict the 31st December (testing dataset 2) short circuit current, Isc(mA). Parameters will be used in this segment is Clean Module Temperature (℃), Dusty Module Temperature (℃), Wind Speed (m/s), Humidity (%), Air Pressure (hPa), Tr (Hour) and T (Difference).

In third section, prediction for both clean solar panel and dusty solar panel short will be performed based on dataset from 1st November to 30th January (training dataset 3) to predict the 31st January (testing dataset 3) short circuit current, Isc(mA). Parameters will be used in this segment is Clean Module Temperature (℃), Dusty Module Temperature (℃), Wind Speed (m/s), Humidity (%), Air Pressure (hPa), Tr (Hour) and T (Difference).

In fourth section, prediction for both clean solar panel and dusty solar panel short will be performed based on dataset from 1st November to 26th February (training dataset 4) to predict the 27th February (testing dataset 4) short circuit current, Isc(mA). Parameters will

be used in this segment is Clean Module Temperature (°C), Dusty Module Temperature (°C), Wind Speed (m/s), Humidity (%), Air Pressure (hPa), Tr (Ho ur) and T (Difference).

In fifth section, prediction for both clean solar panel and dusty solar panel short will be performed based on dataset from 1st November to 30th March (training dataset 5) to predict the 31st March (testing dataset 5) short circuit current, Isc(mA). Parameters will be used in this segment is Clean Module Temperature (°C), Dusty Module Temperature (°C), Wind Speed (m/s), Humidity (%), Air Pressure (hPa), Tr (Hour) and T (Difference).

5.2 Prediction Analysis

5.2.1 Analysis of Random Forest

For the Training dataset 1: 1st November 2019 to 29th November 2019 (29 days) for Clean and Dusty Module.

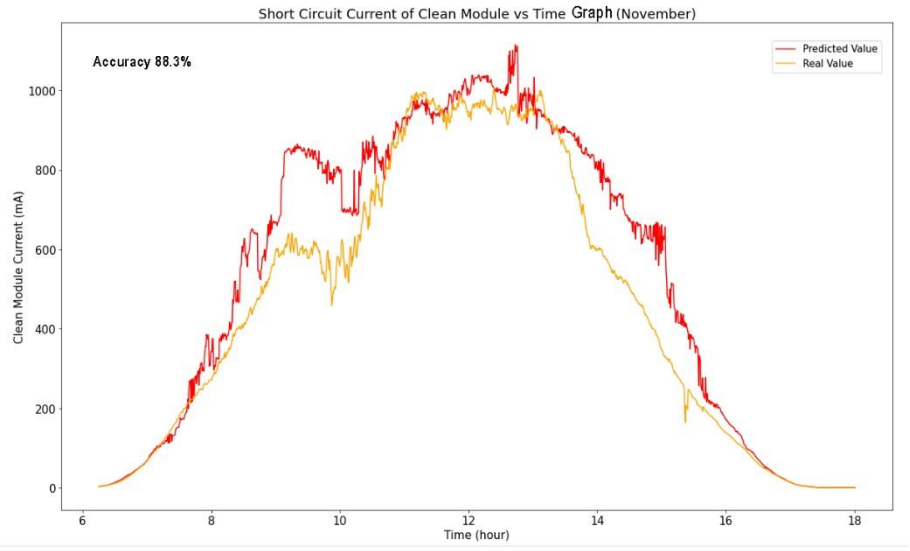


Figure 5.1: Plot of predicted value and real value for November (Clean Module). Here the accuracy of the model is 88.3%.

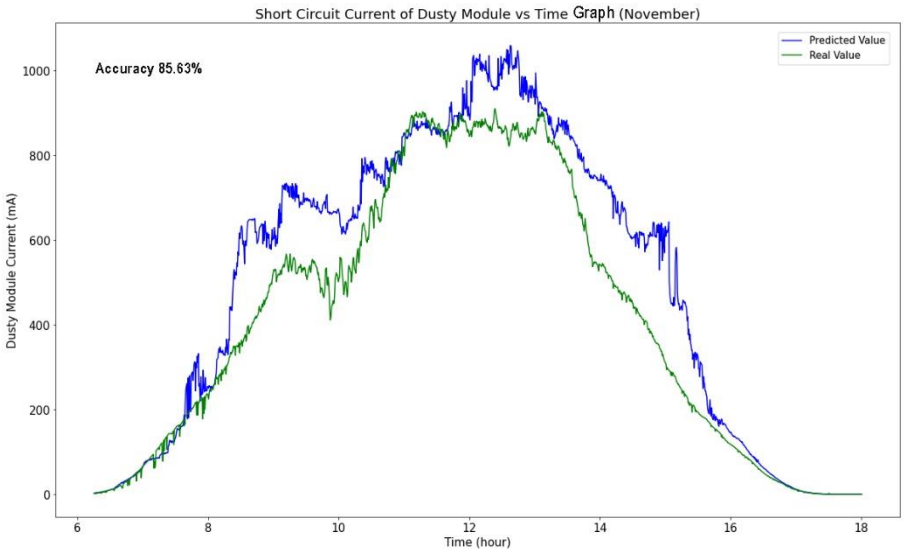


Figure 5.2: Plot of predicted value and real value for November (Dusty Module)

Here the accuracy of the model is 85.63%.

Training dataset 2: 1st November 2019 to 29th December 2019 (59days) for clean and Dusty Module

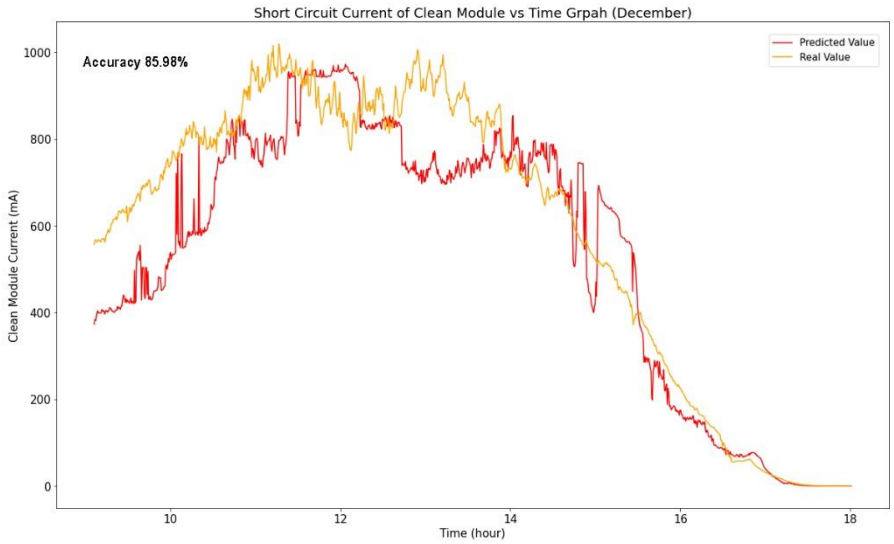


Figure 5.3: Plot of predicted value and real value for December (Clean Module) Here the accuracy of the model is 85.98%.

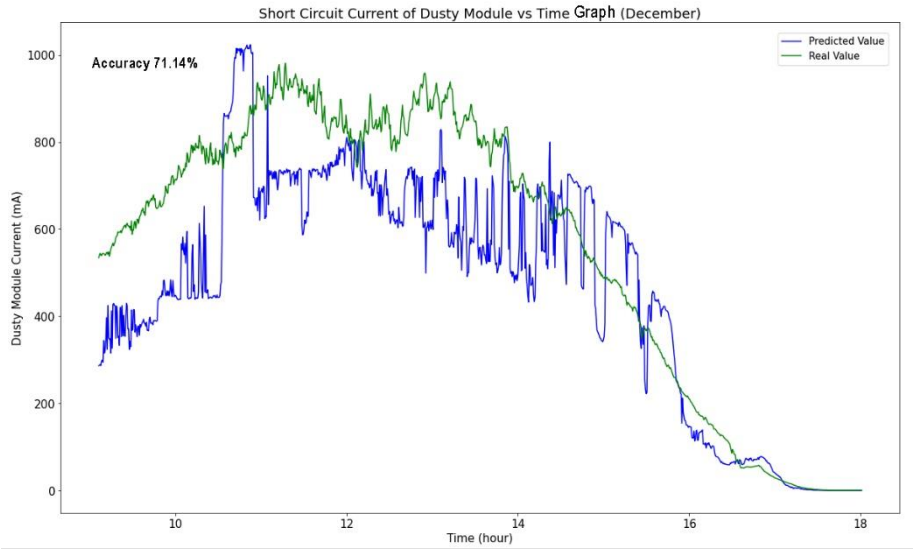


Figure 5.4: Plot of predicted value and real value for December (Dusty Module). Here the accuracy of the model is 71.14%.

Training dataset 3: 1st November 2019 to 30th January 2020 (89 days) for Clean and Dusty Module

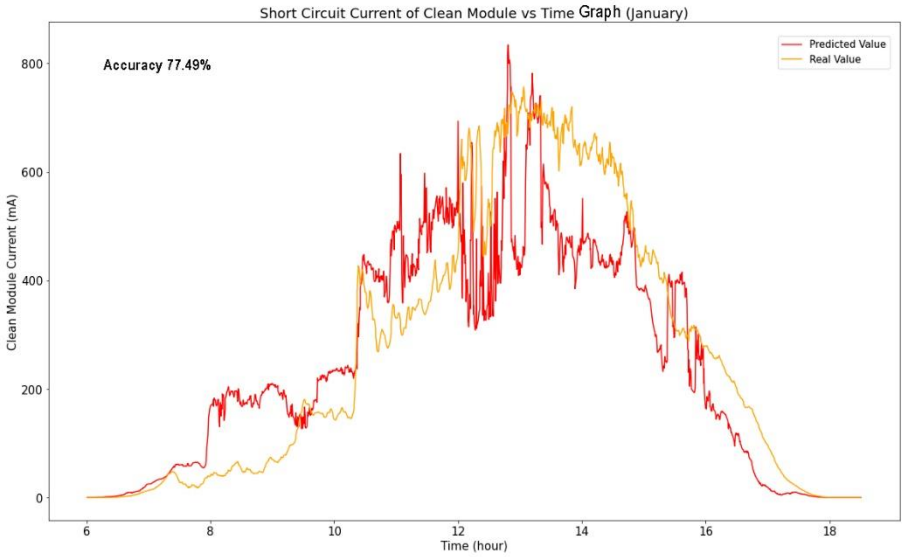


Figure 5.5: Plot of predicted value and real value for January (Clean Module). Here the accuracy of the model is 77.49%.

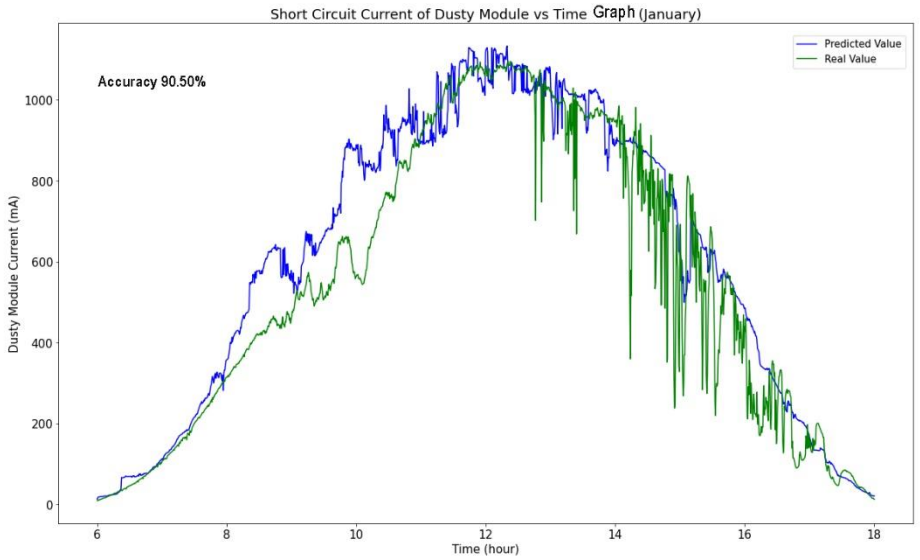


Figure 5.6: Plot of predicted value and real value for January (Dusty Module). Here, the accuracy of the model is 90.50%.

Training dataset 4: 1st November 2019 to 26th February 2020. (115 days) for Clean and Dusty Module

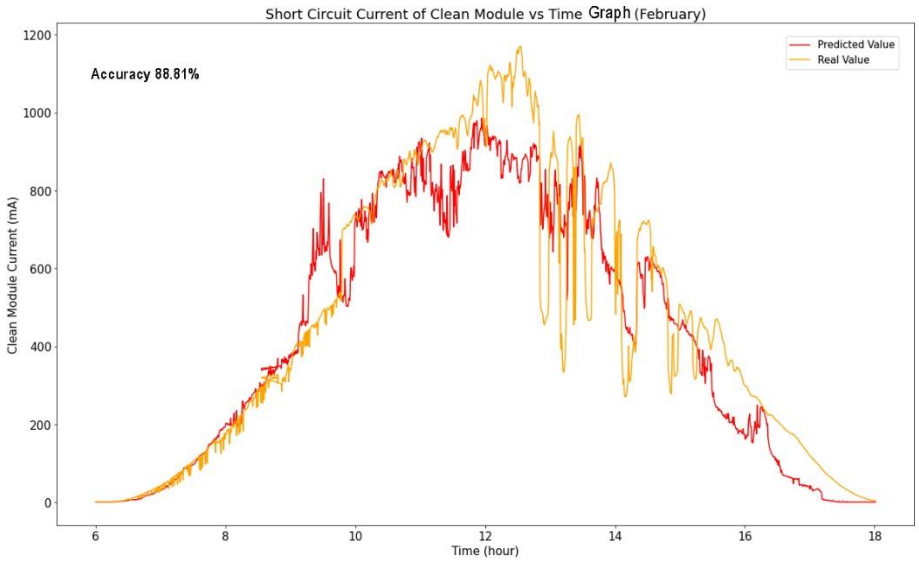


Figure 5.7: Plot of predicted value and real value for February (Clean Module). Here the accuracy of the model is 88.81%.

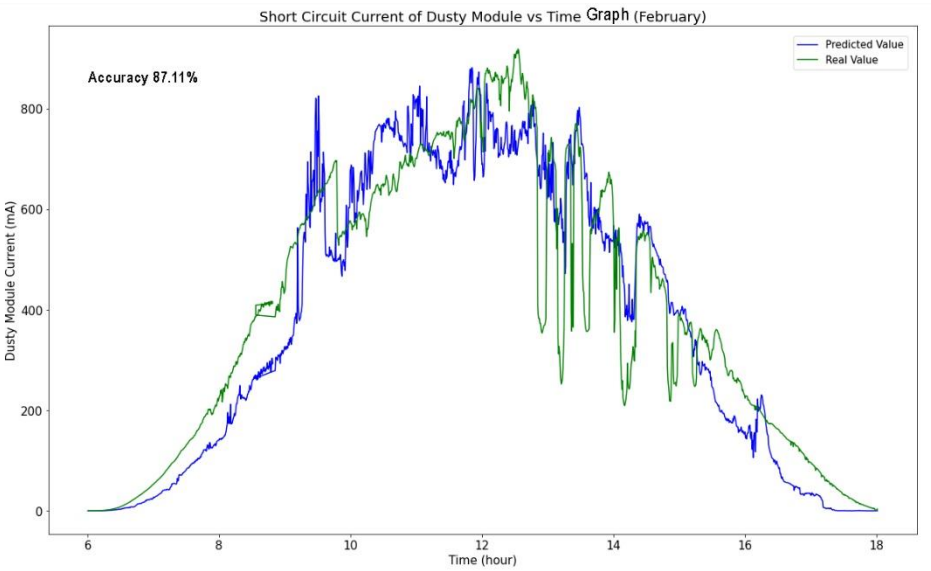


Figure 5.8: Plot of predicted value and real value for February (Dusty Module). Here the accuracy of the model is 87.11%.

Training dataset 5: 1st November 2019 to 30th March 2020. (145 days) for clean and Dusty Module

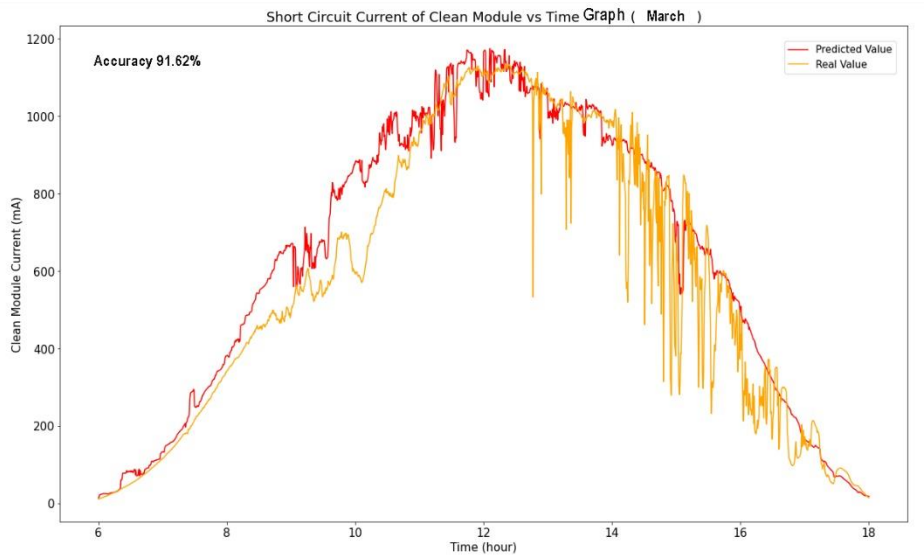


Figure 5.9: Plot of predicted value and real value for January (Clean Module). Here the accuracy of the model is 91.62%.

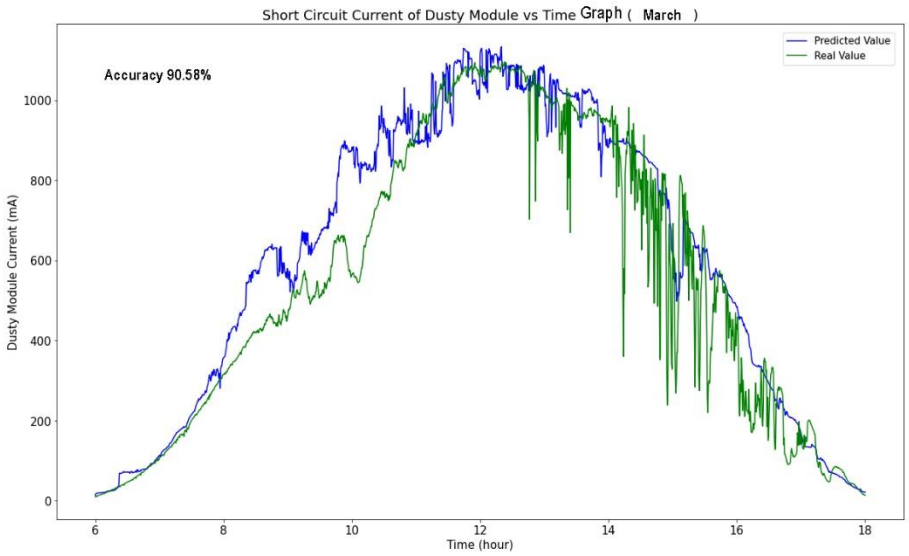


Figure 5.10: Plot of predicted value and real value for March (Dusty Module). Here the accuracy of the model is 90.58%.

5.2.1.1 Bar chart representation

Here is a bar chart portrayal of a Predicted Short Circuit Current for Clean Module and Dusty Module Accuracy (%)

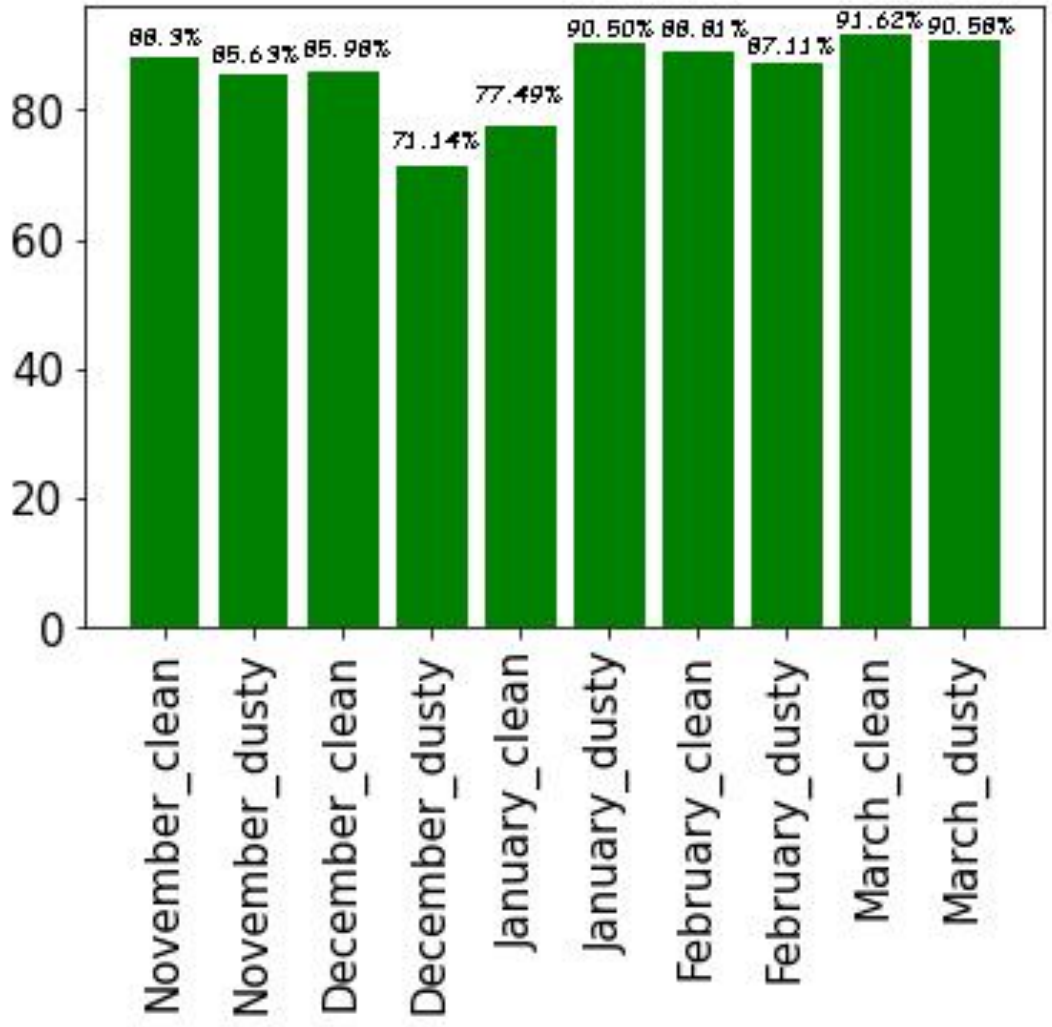


Figure 5.11: Bar chart of Predicted Short Circuit Current for Clean Module and Dusty Module Accuracy (%).

Table 5.1: Comparison of results between Clean and Dusty Module. (Random Forest).

Dataset	**Predicted Short Circuit Current for Clean Module Accuracy (%)**	**Predicted Short Circuit Current for Dusty Module Accuracy (%)**
1) Training dataset 1: 1st November 2019 to 28th November 2019	88.3	85.63
2) Training dataset 2: 1st November 2019 to 30th December 2019	85.98	71.14
3) Training dataset 3: 1st November 2019 to 30th January 2020	77.49	90.50
4) Training dataset 4: 1st November 2019 to 26th February 2020	88.81	87.11
5) Training dataset 5: 1st November 2019 to 30th March 2020	91.62	90.58

5.2.2 Prediction analysis for Artificial Neural Network (ANN)

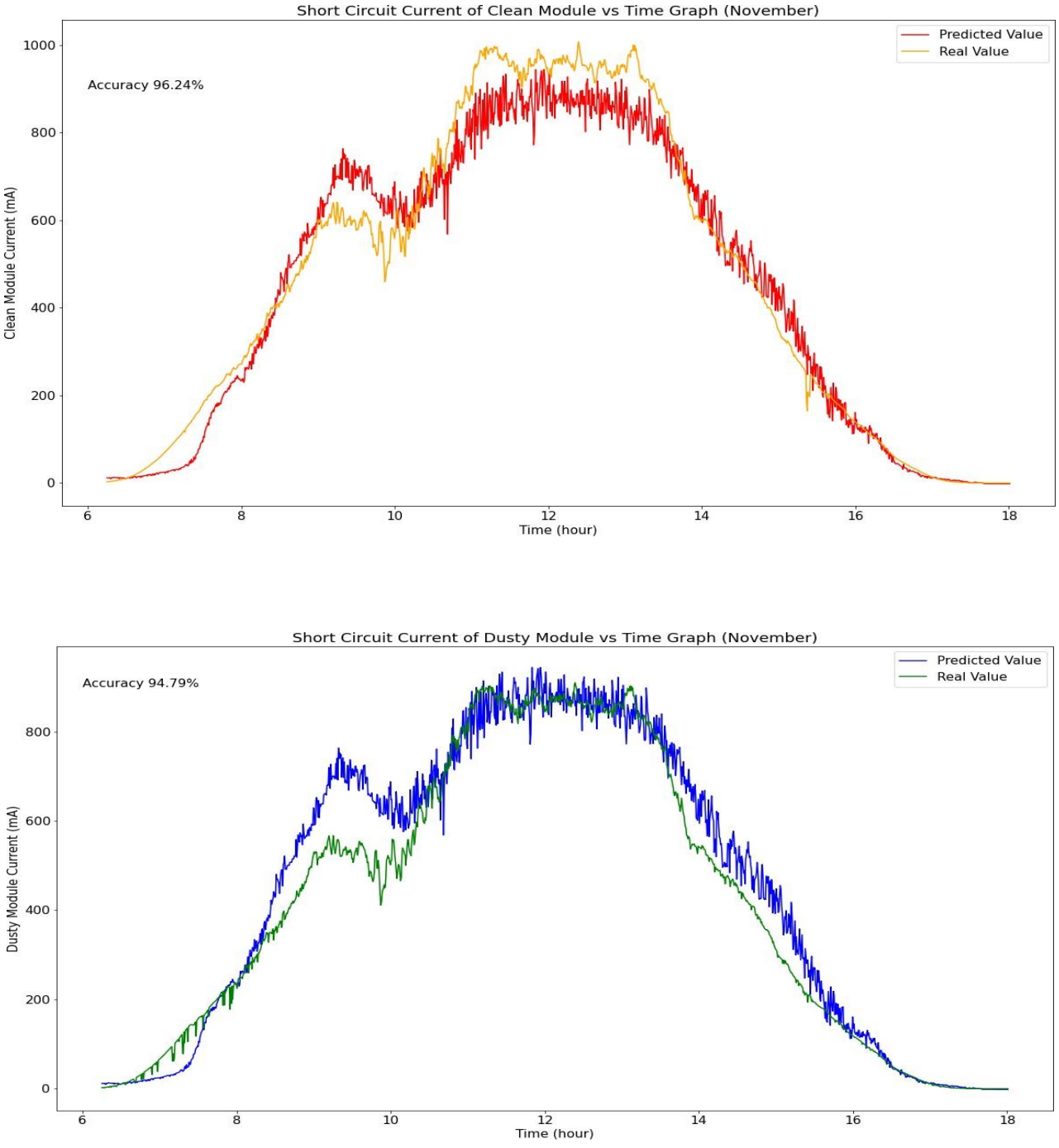
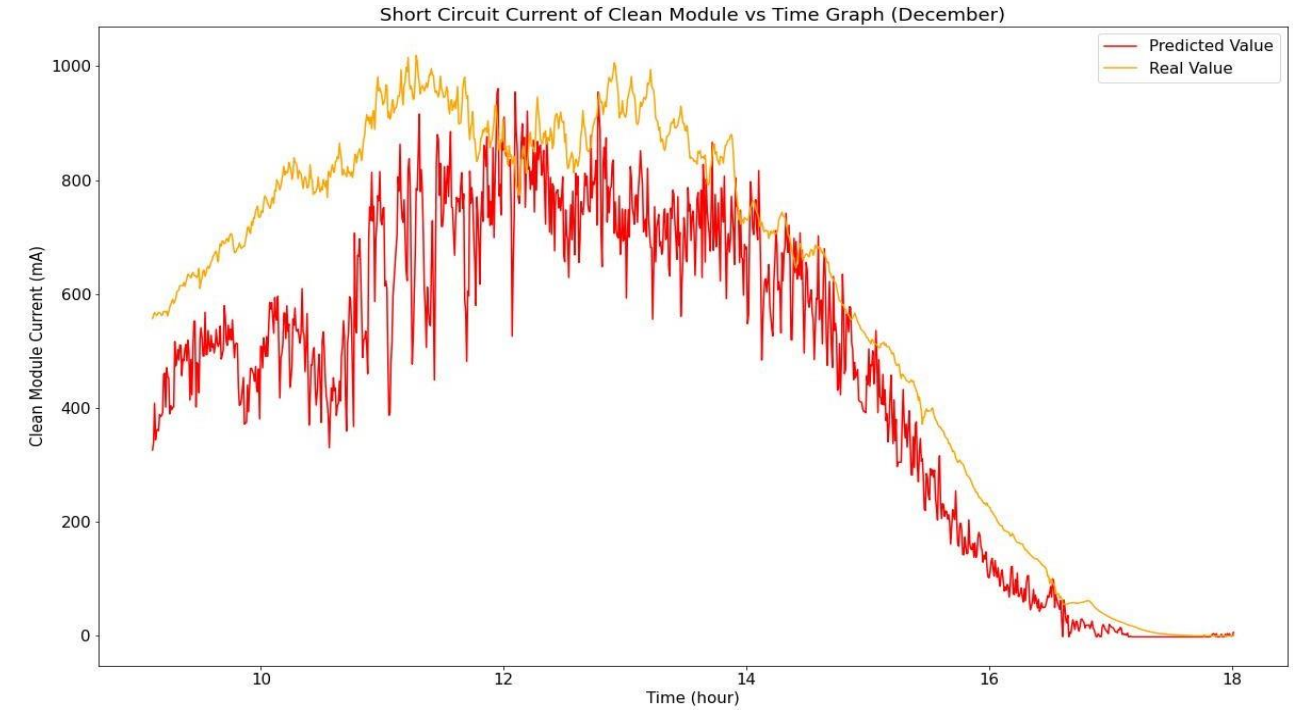


Figure 5.12: Real value and Predicted value plot for short circuit current estimated of upper oneclean module and lower one dusty module, using Training dataset 1.



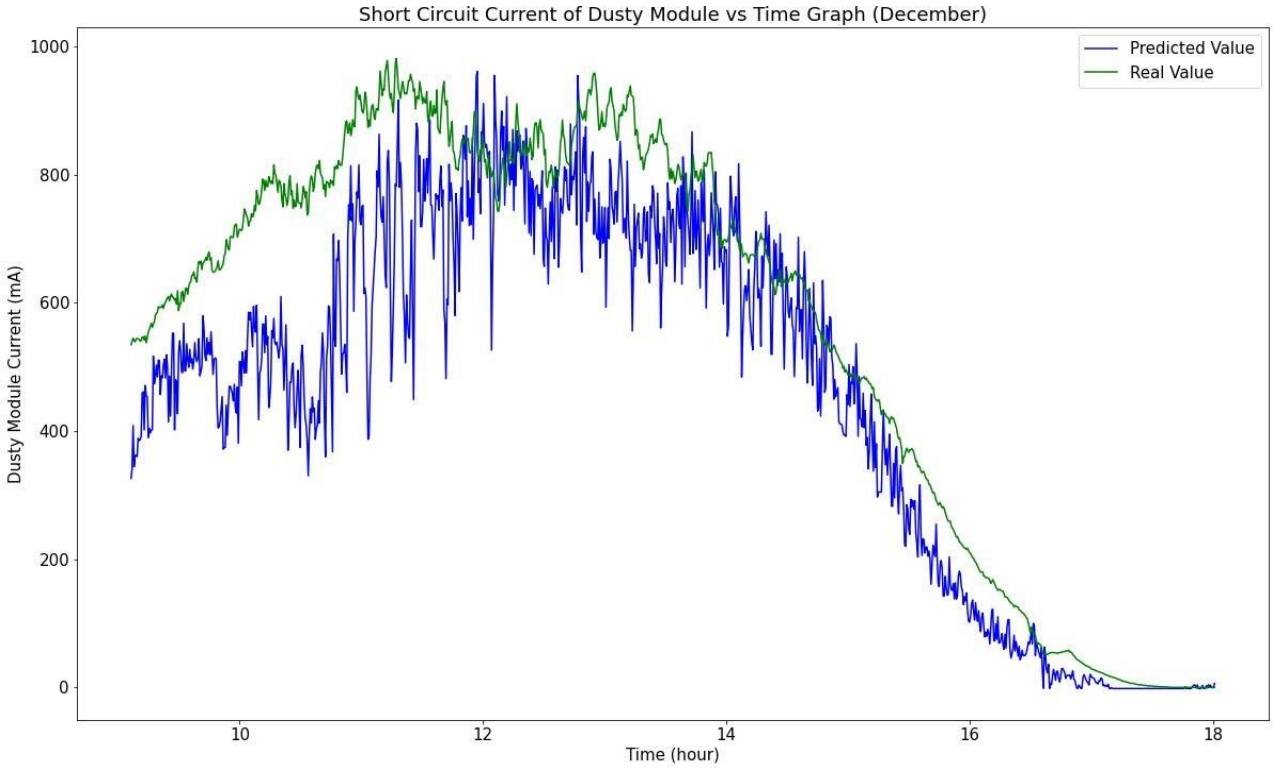


Figure 5.13: Real value and Predicted value plot for short circuit current estimated of upper oneclean module and lower one dusty module, using Training dataset 2.

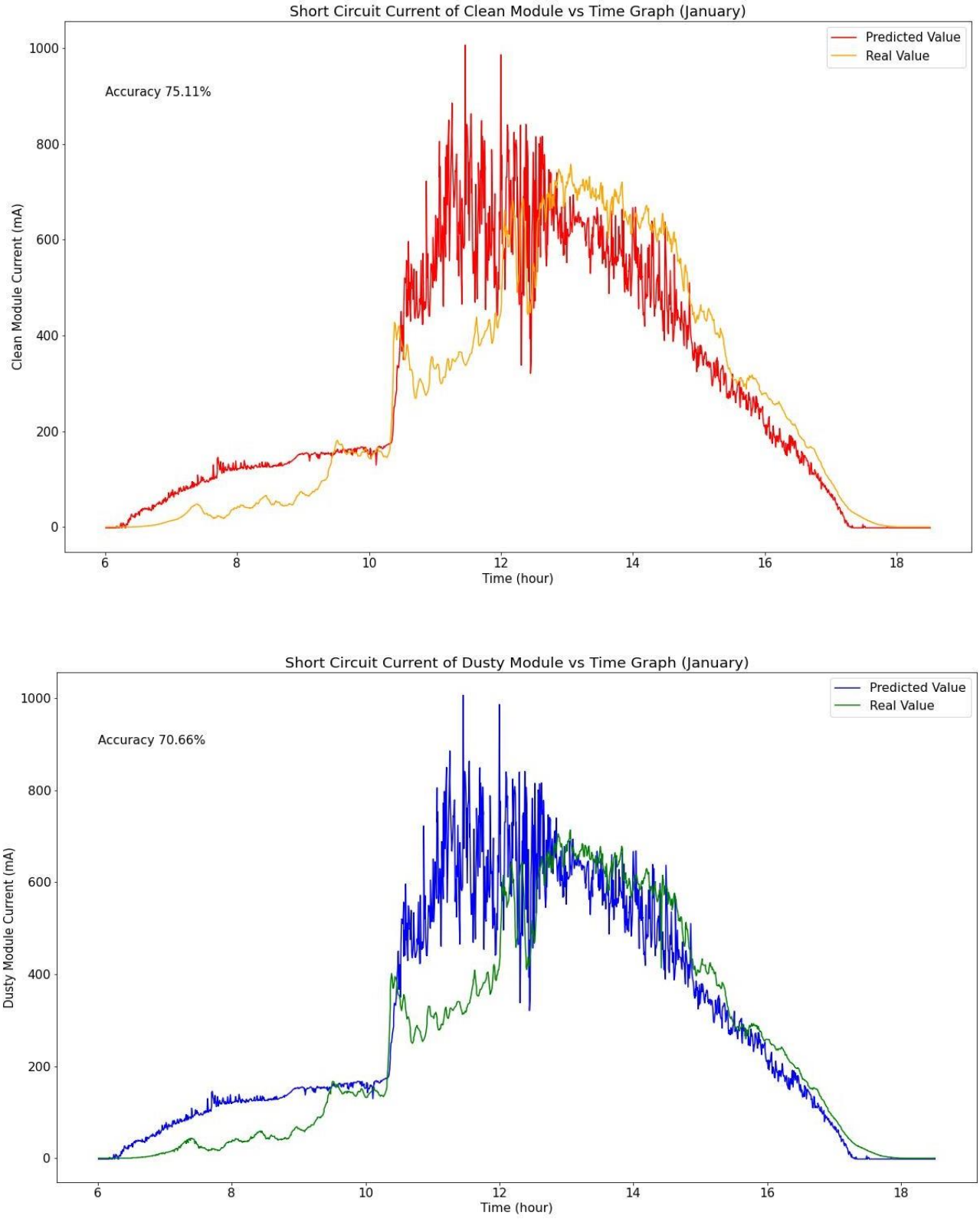


Figure 5.14: Real value and Predicted value plot for short circuit current estimated of upper oneclean module and lower one dusty module, using Training dataset 3.

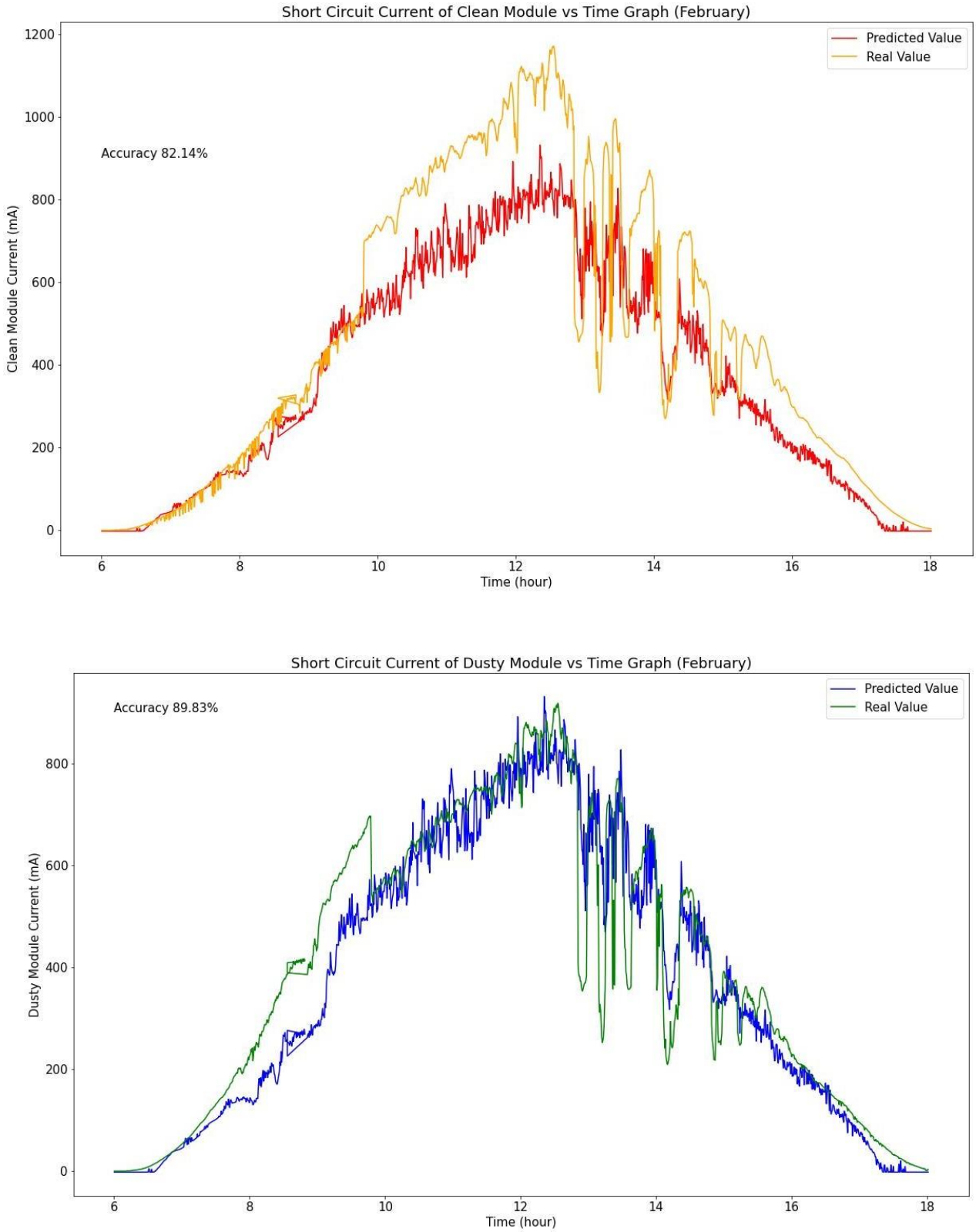
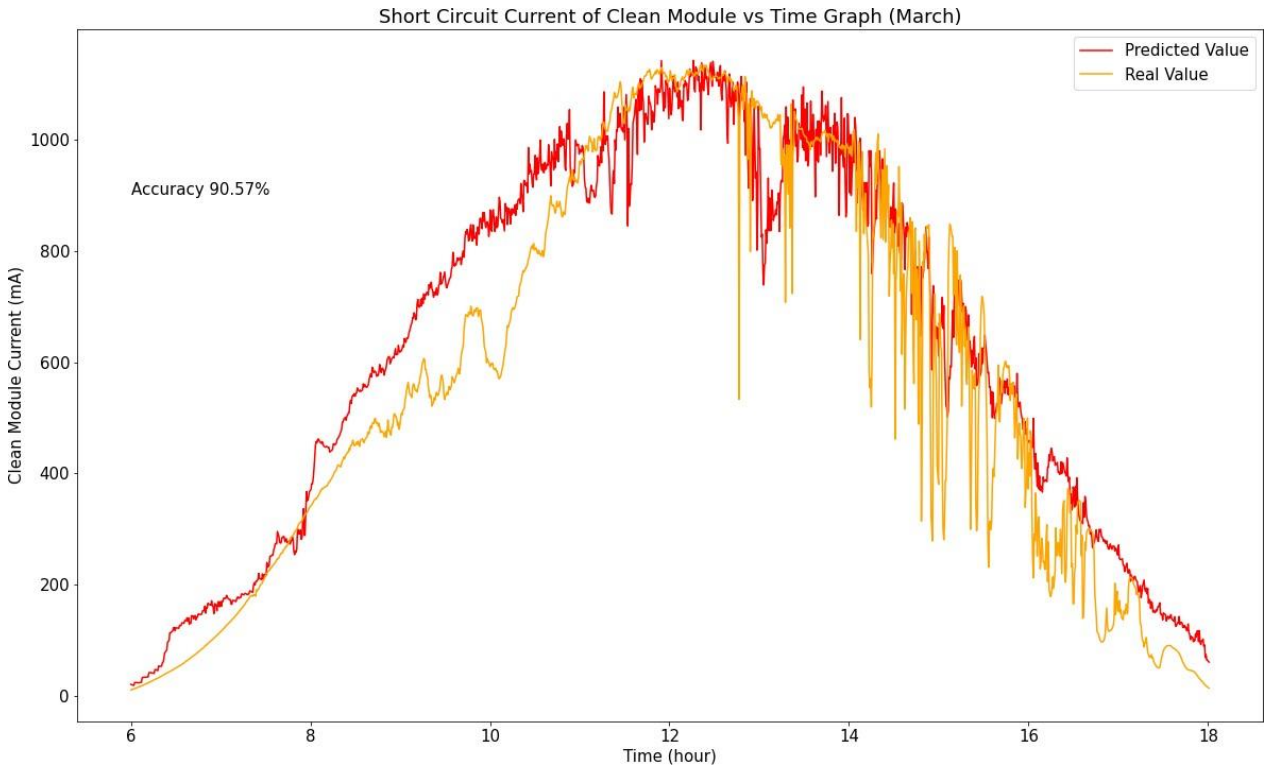


Figure 5.15: Real value and Predicted value plot for short circuit current estimated of upper one cleanmodule and lower one dusty module, using Training dataset 4.



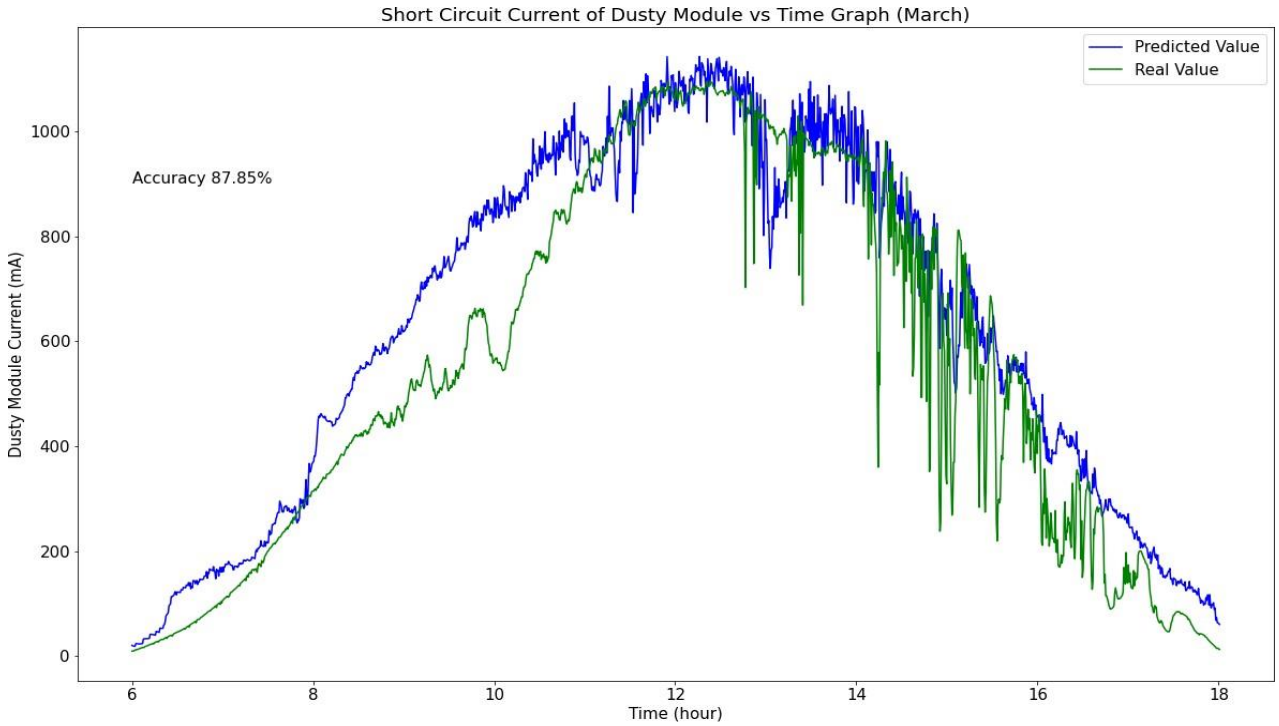


Figure 5.16: Real value and Predicted value plot for short circuit current estimated of upper one clean module and lower one dusty module, using Training dataset 5.

Table 5.2 Comparison of Results Between Clean and Dusty Module.

Dataset	Predicted Short Circuit Current for Clean Module Accuracy (%)	Predicted Short Circuit Current for Dusty Module Accuracy (%)
1) Training dataset 1: 1st November 2019 to 28th November 2019	96.24	94.79
2) Training dataset 2: 1st November 2019 to 30th December 2019	73.38	78.42
3) Training dataset 3: 1st November 2019 to 30th January 2020	75.11	70.66
4) Training dataset 4: 1st November 2019 to 26th February 2020	82.14	89.83
5) Training dataset 5: 1st November 2019 to 30th March 2020	90.57	87.85

5.2.2.1 Bar chart representation

Here is a bar chart portrayal of a Predicted Short Circuit Current for Clean Module and Dusty Module Accuracy (%)

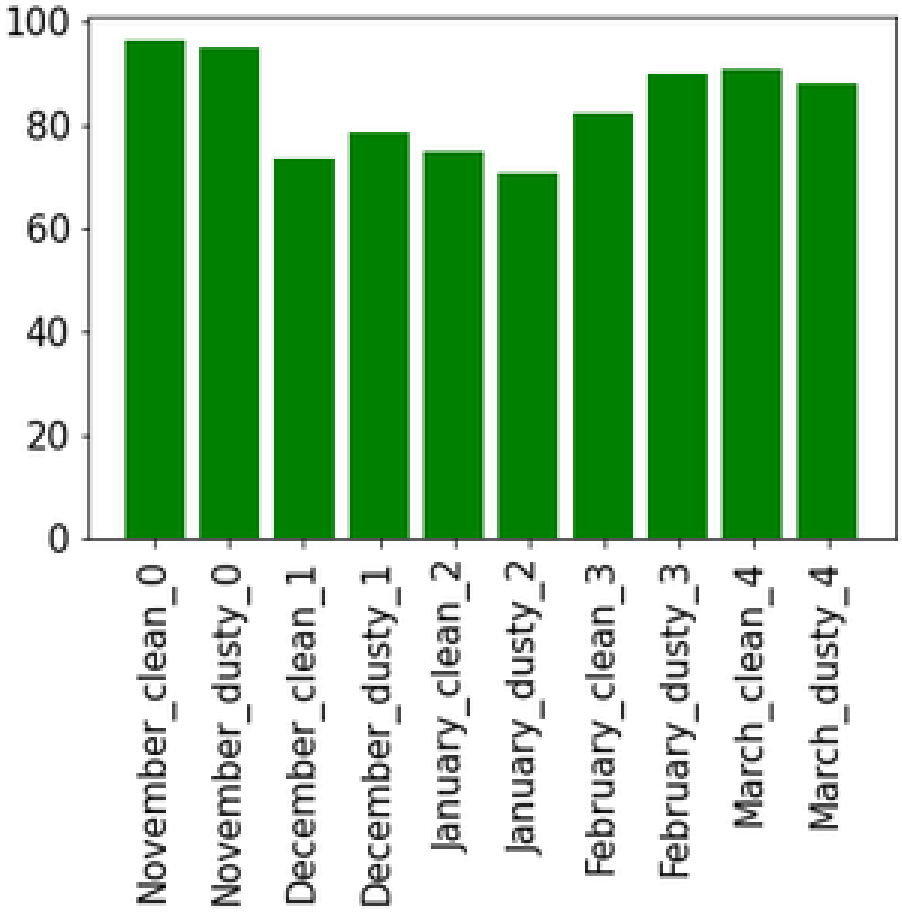


Figure 5.17: Bar chart of Predicted Short Circuit Current for Clean Module and Dusty ModuleAccuracy (%).

5.2.3 Prediction analysis for Multiple Linear Regression (MLR)

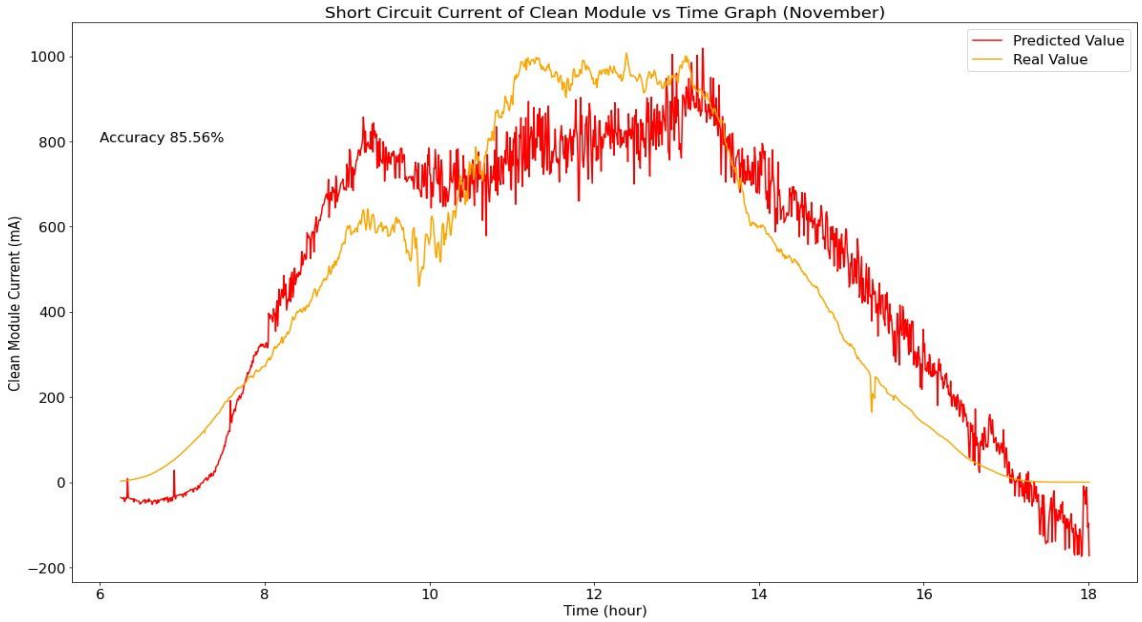


Fig 5.18: Predicted value vs real value for February (Clean Module).

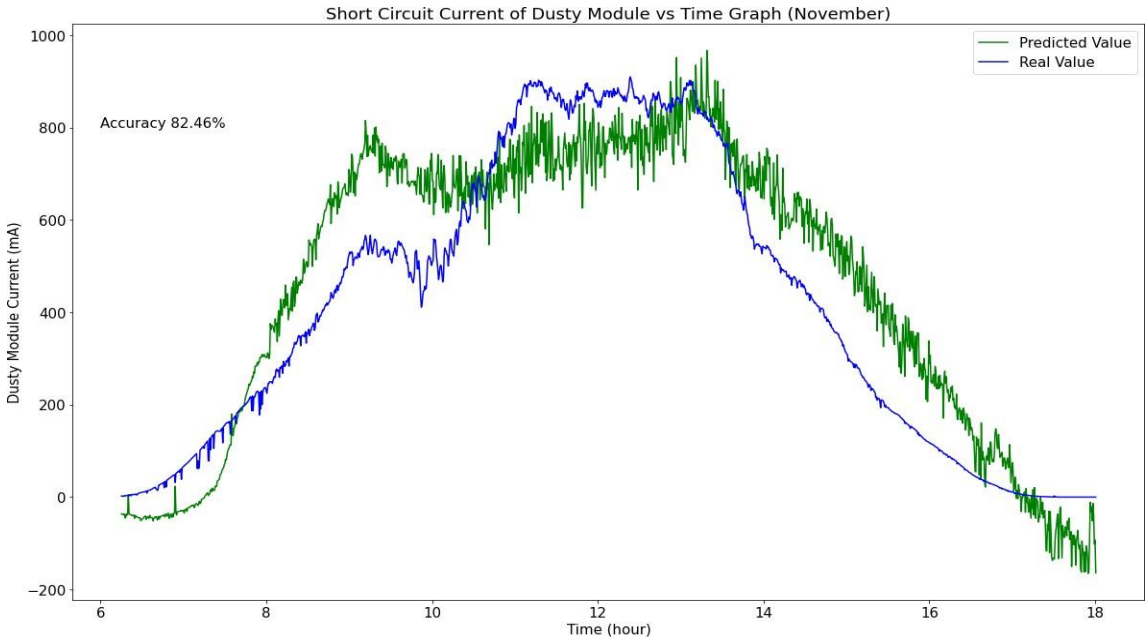


Fig 5.19: Predicted value vs real value for February (Dusty Module).

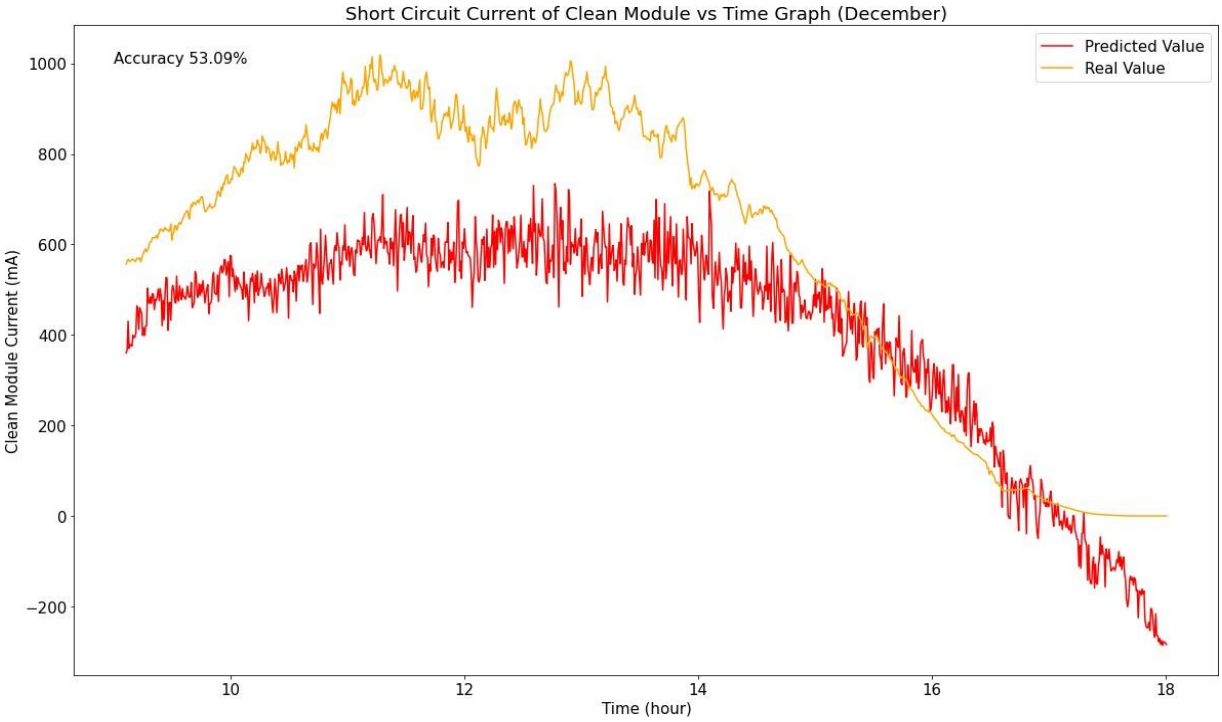


Fig 5.20: Predicted value vs real value for February (Clean Module).

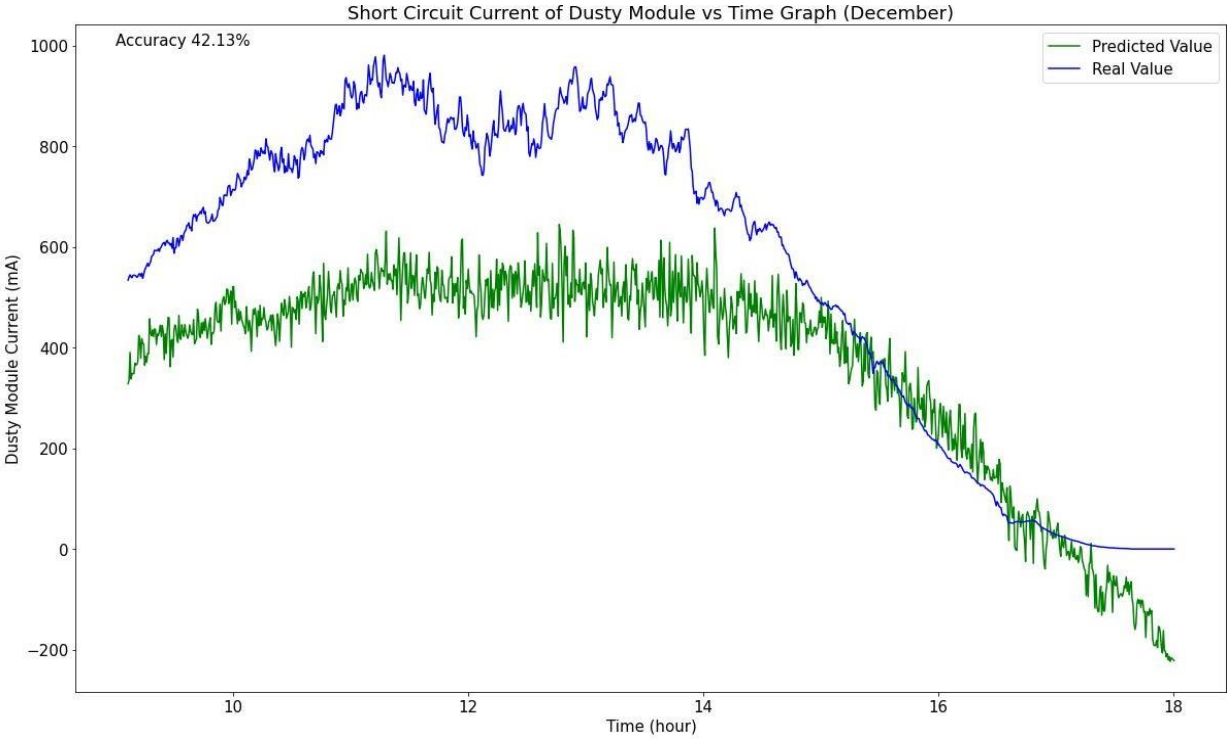


Fig 5.21: Predicted value vs real value for February (Dusty Module).

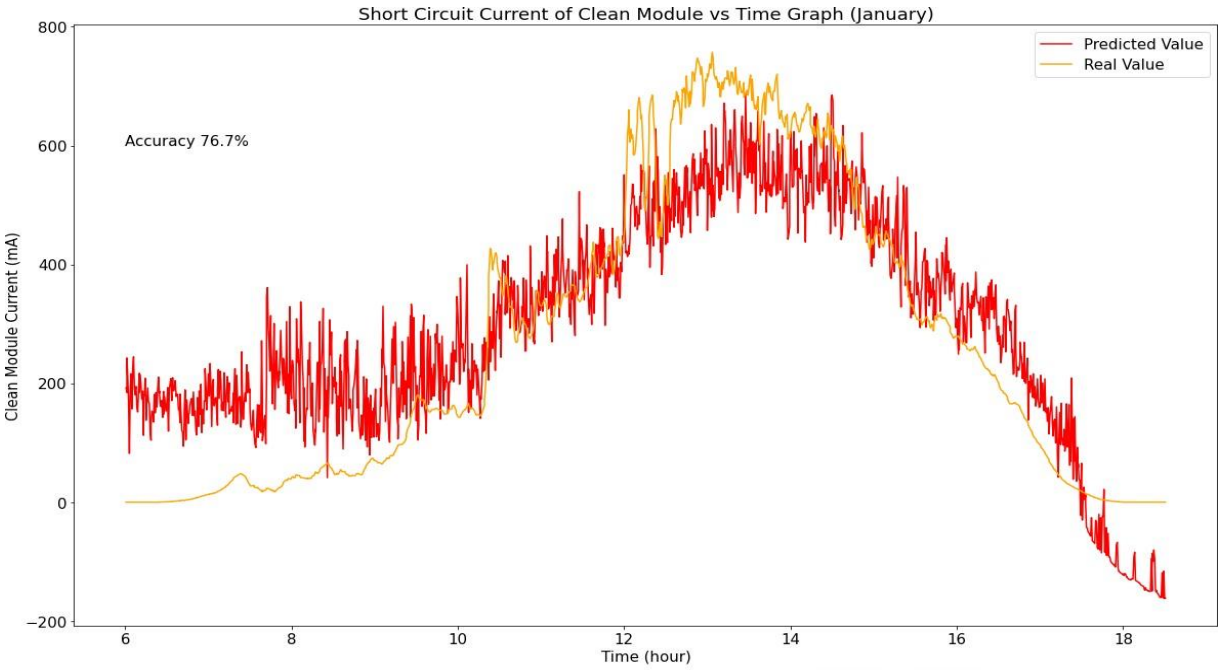


Fig 5.22: Predicted value vs real value for February (Clean Module).

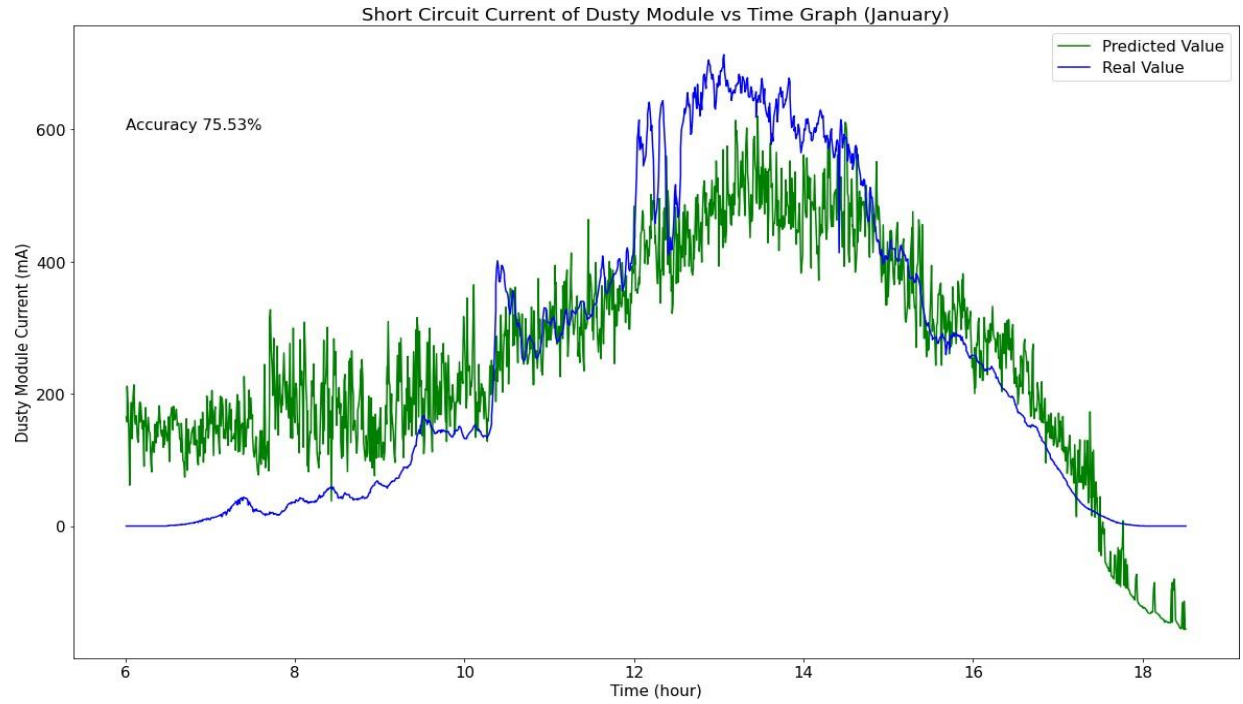


Fig 5.23: Predicted value vs real value for February (Dusty Module).

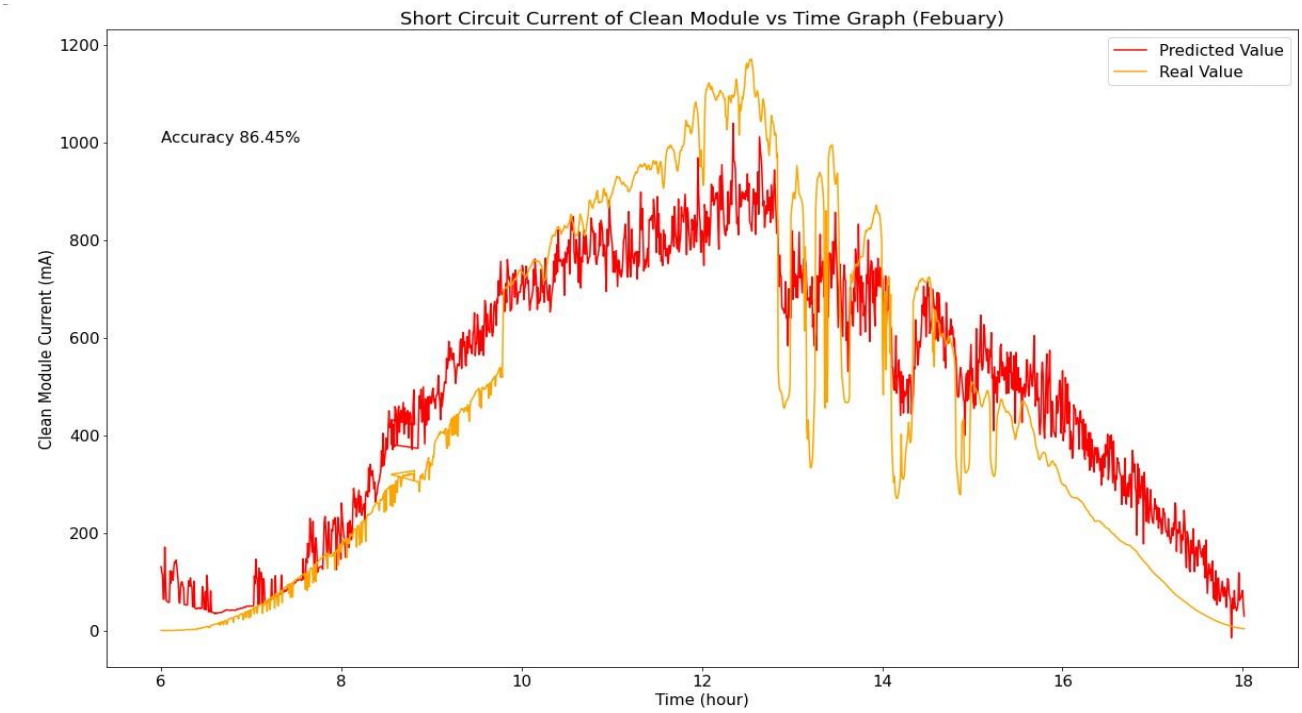


Fig 5.24: Predicted value vs real value for February (Clean Module).

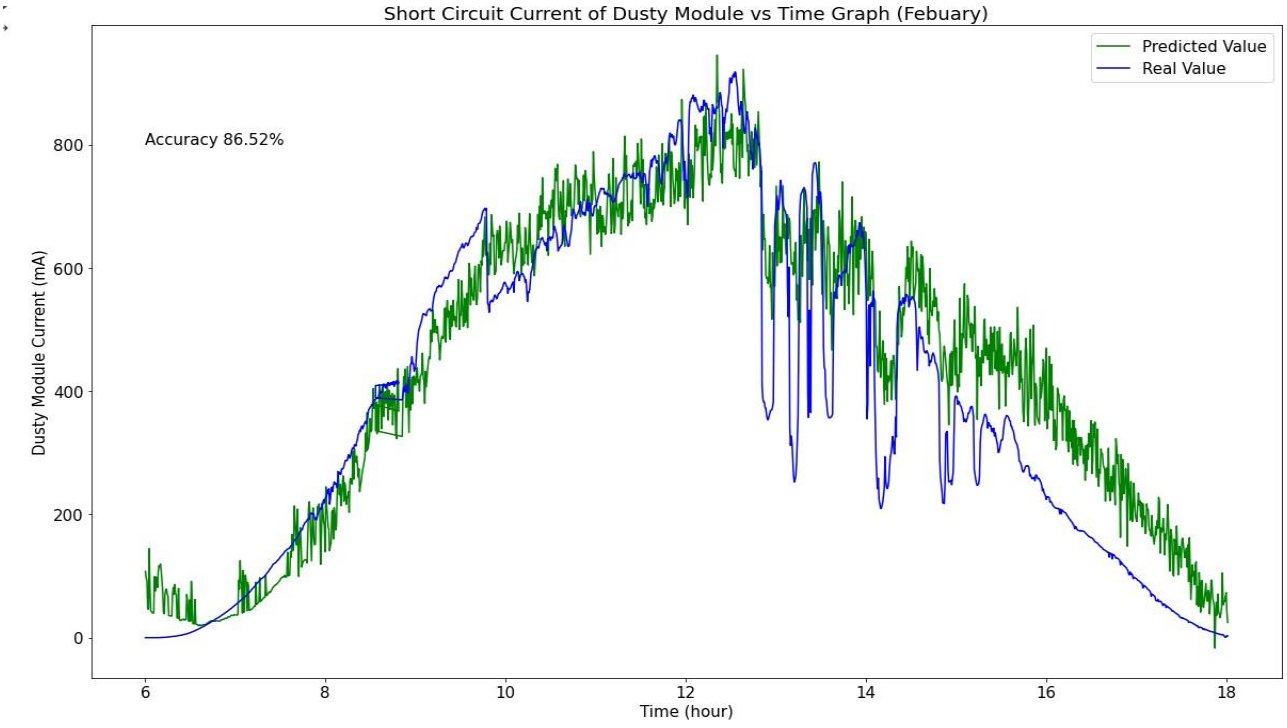


Fig 5.25: Predicted value vs real value for February (Dusty Module).

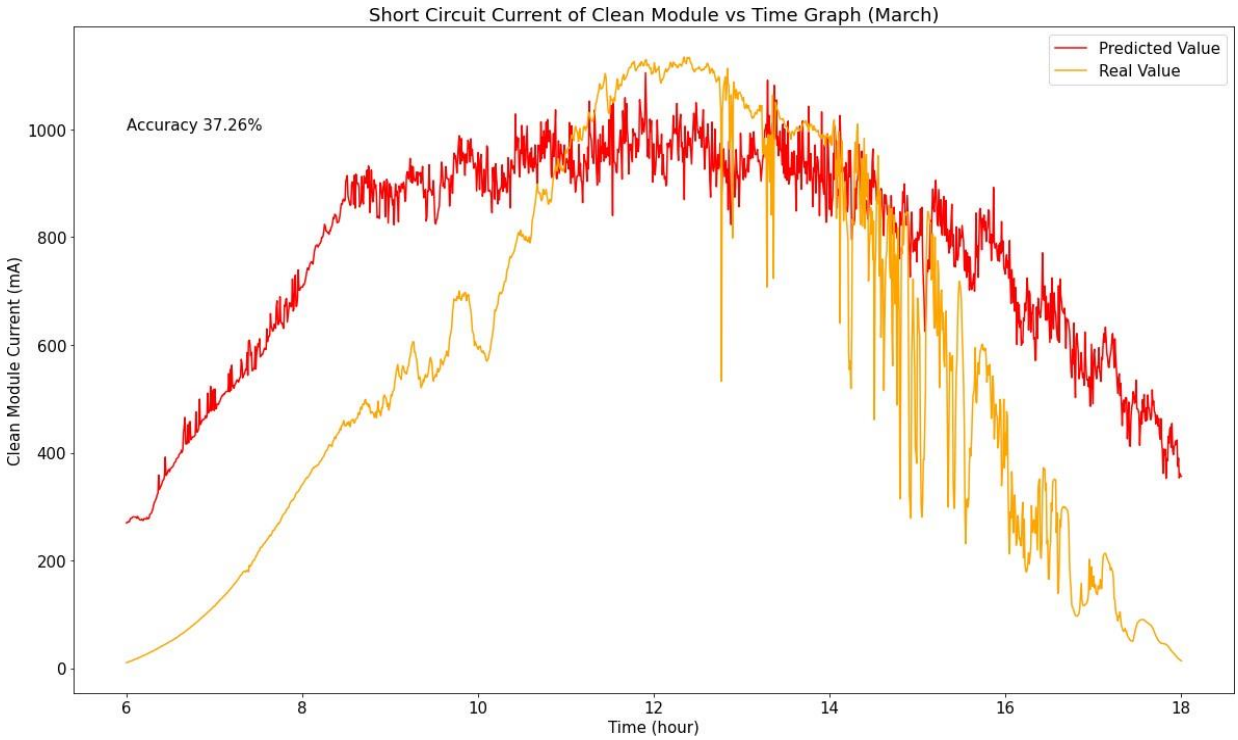


Fig 5.26: Predicted value vs real value for March (Clean Module).

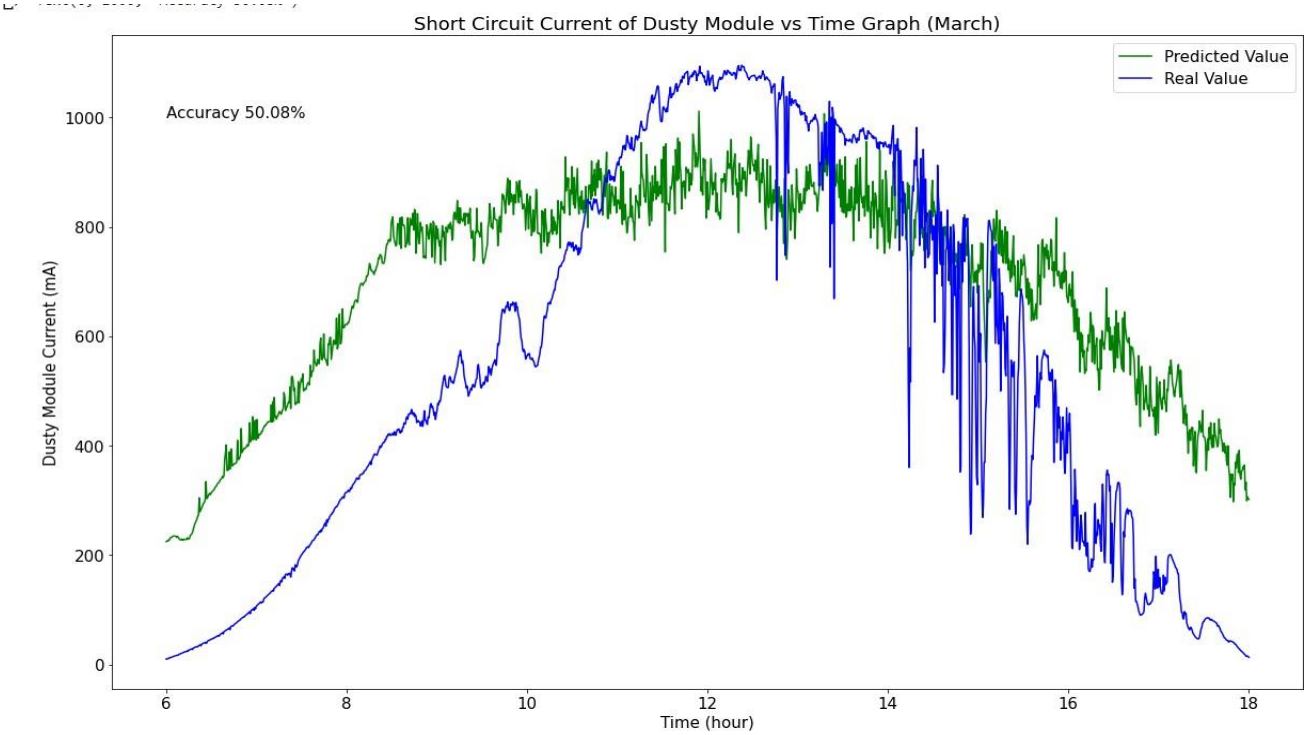


Fig 5.27: Predicted value vs real value for March (Dusty Module).

5.2.3.1 Representation of Bar-chart

Here is a bar chart portrayal of a Predicted Short Circuit Current for Clean Module and Dusty Module Accuracy (%)

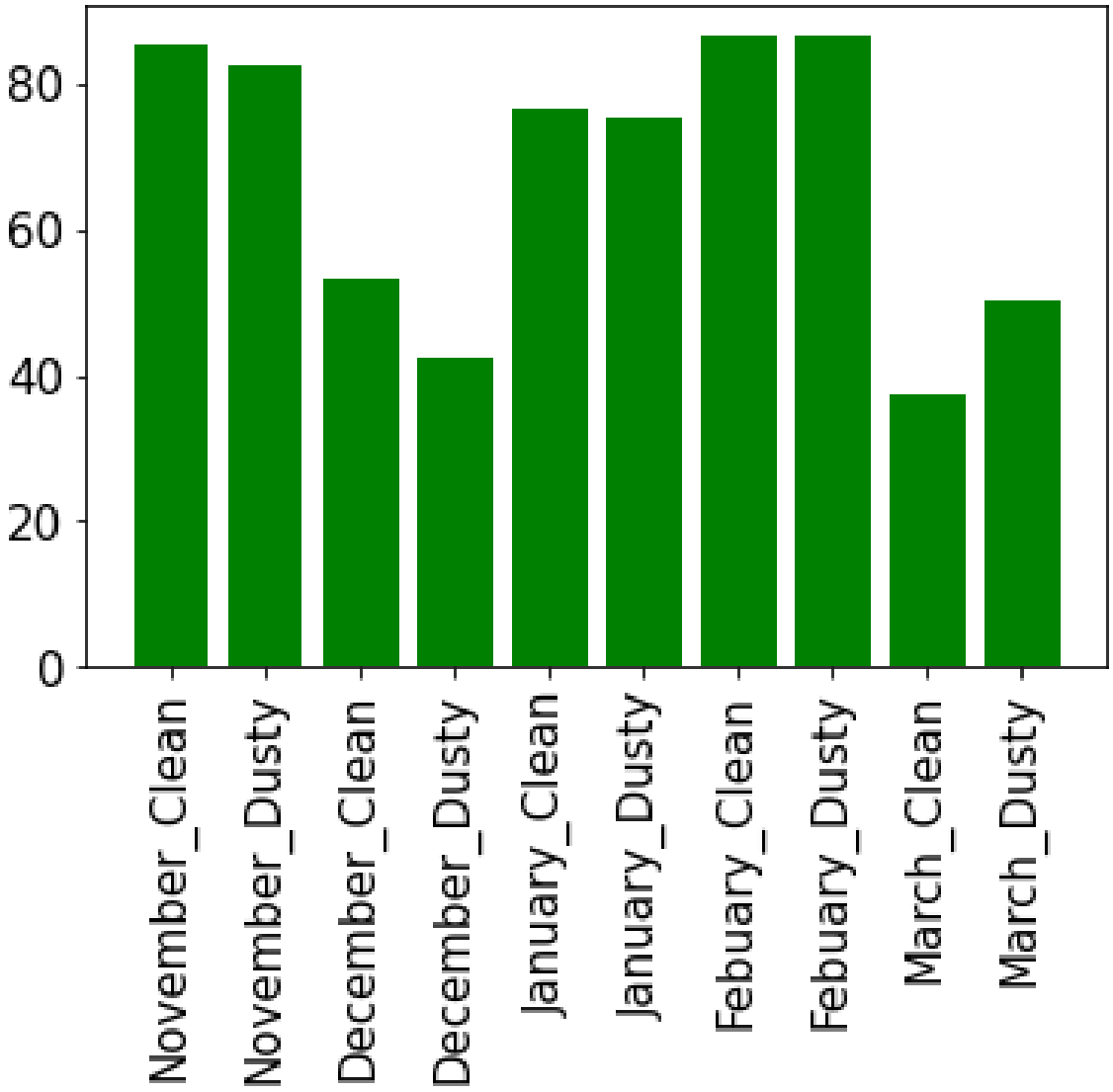


Fig 5.28: Accuracy (%) In Bar-Chart.

Table 5.3 Comparison of Results Between Clean and Dusty Module.

Dataset	**Predicted Short Circuit Current for Clean Module Accuracy (%)**	**Predicted Short Circuit Current for Dusty Module Accuracy (%)**
1) Training dataset 1: 1st November 2019 to 28th November 2019	85.56	82.46
2) Training dataset 2: 1st November 2019 to 30th December 2019	53.09	42.13
3) Training dataset 3: 1st November 2019 to 30th January 2020	76.7	75.53
4) Training dataset 4: 1st November 2019 to 26th February 2020	86.54	86.52
5) Training dataset 5: 1st November 2019 to 30th March 2020	37.26	50.08

Chapter 6

Comparative Analysis

6.1 Comparison of three Machine Learning Methods

For training dataset 1, Artificial neural Network model performed best for clean module and dusty module. For training dataset 2, Random Forest model performed best for clean module and Artificialneural network model performed best for dusty module. For training dataset 3, Random Forest modelperformed best for both clean and dusty module. For training dataset 4, Random Forest model performed best for clean module and Artificial Neural Network model performed best for dusty module. For training dataset 5, Random Forest model performed best for both clean and dusty module.

The sensors integrated in clean PV module as well as dusty modules have dispensed data of short circuit current for varying weather parameters. With the exploitation of several machine learning algorithms and analyzing the outcome, we have been piloted that, other than using linear regression method or artificial neural network, random forest method provides considerably superior accuracy.

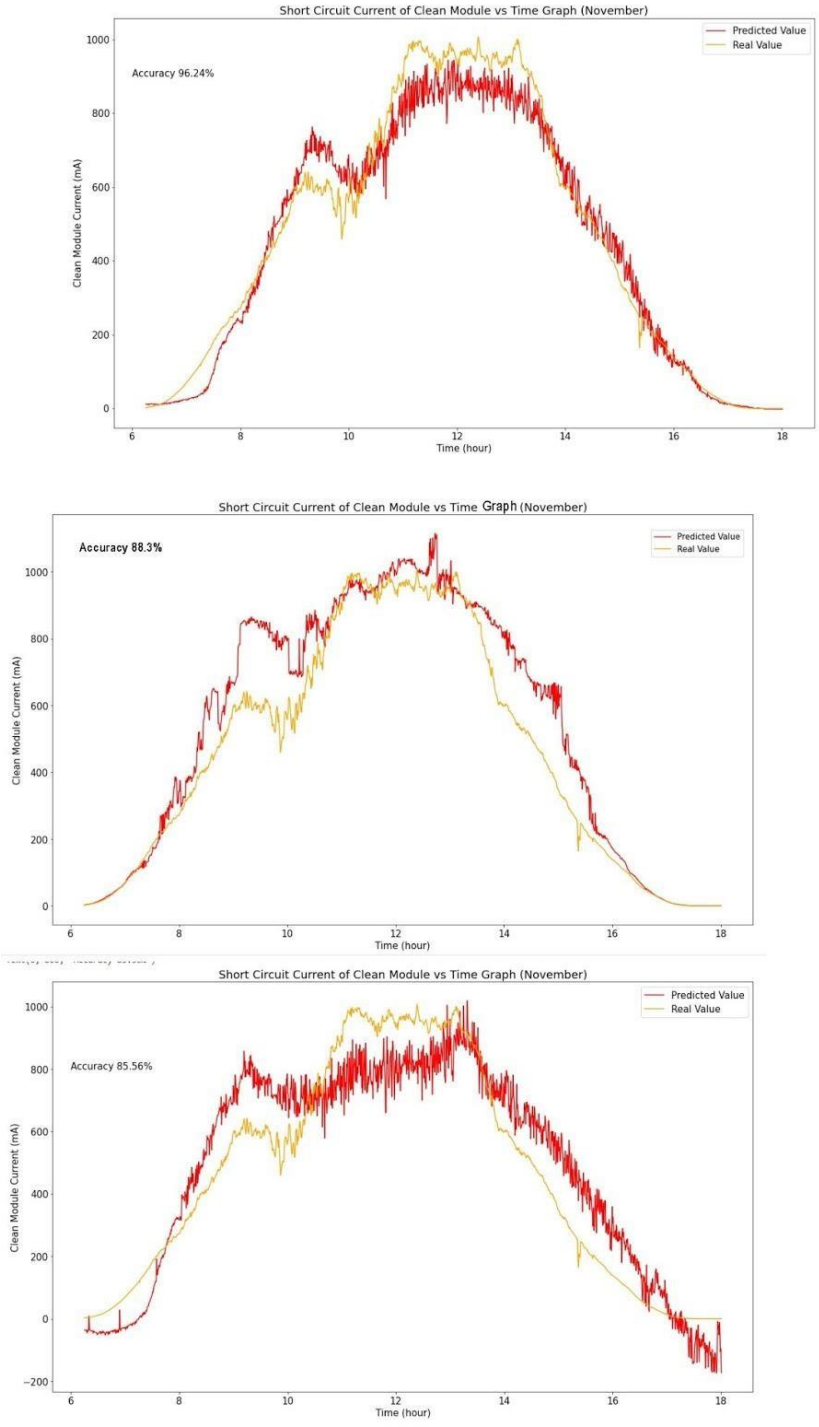


Figure 6.1: Clean module Comparison for dataset 1.

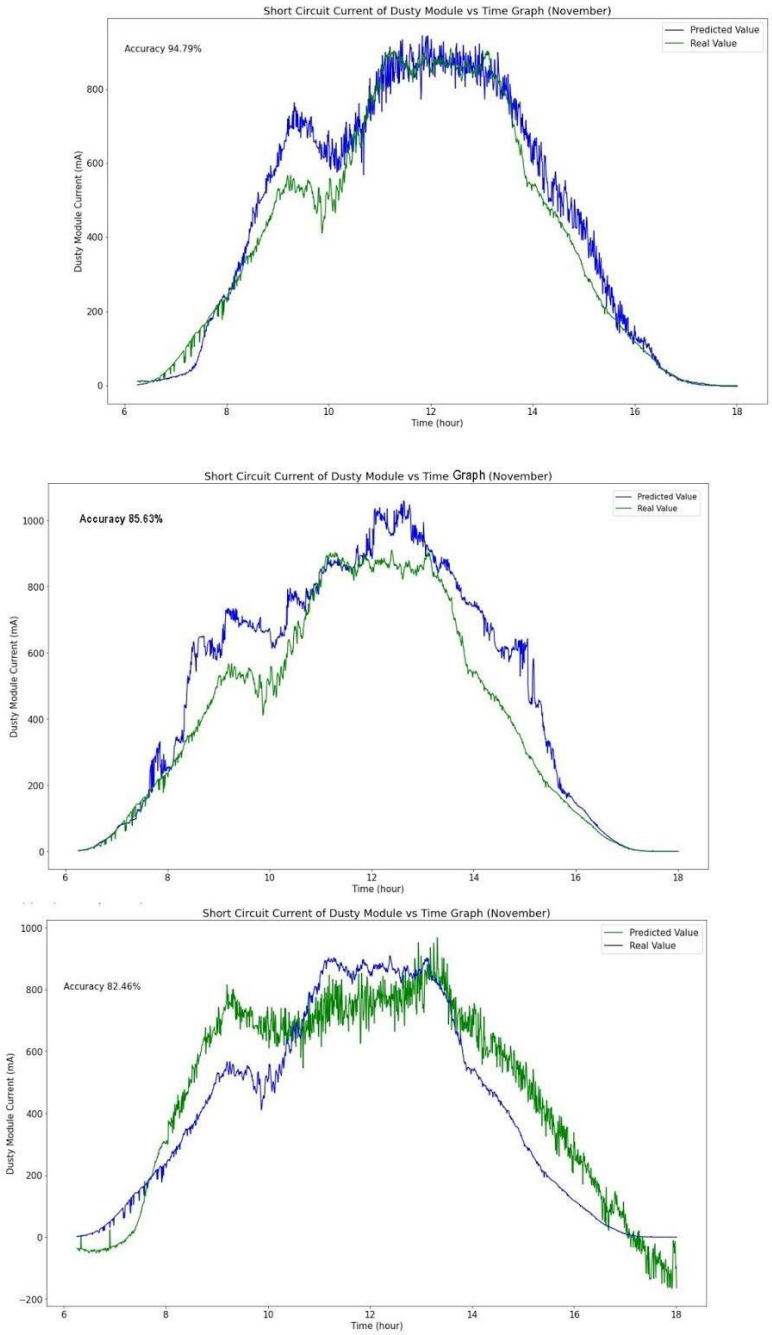


Figure 6.2: Dusty module Comparison for dataset 1.

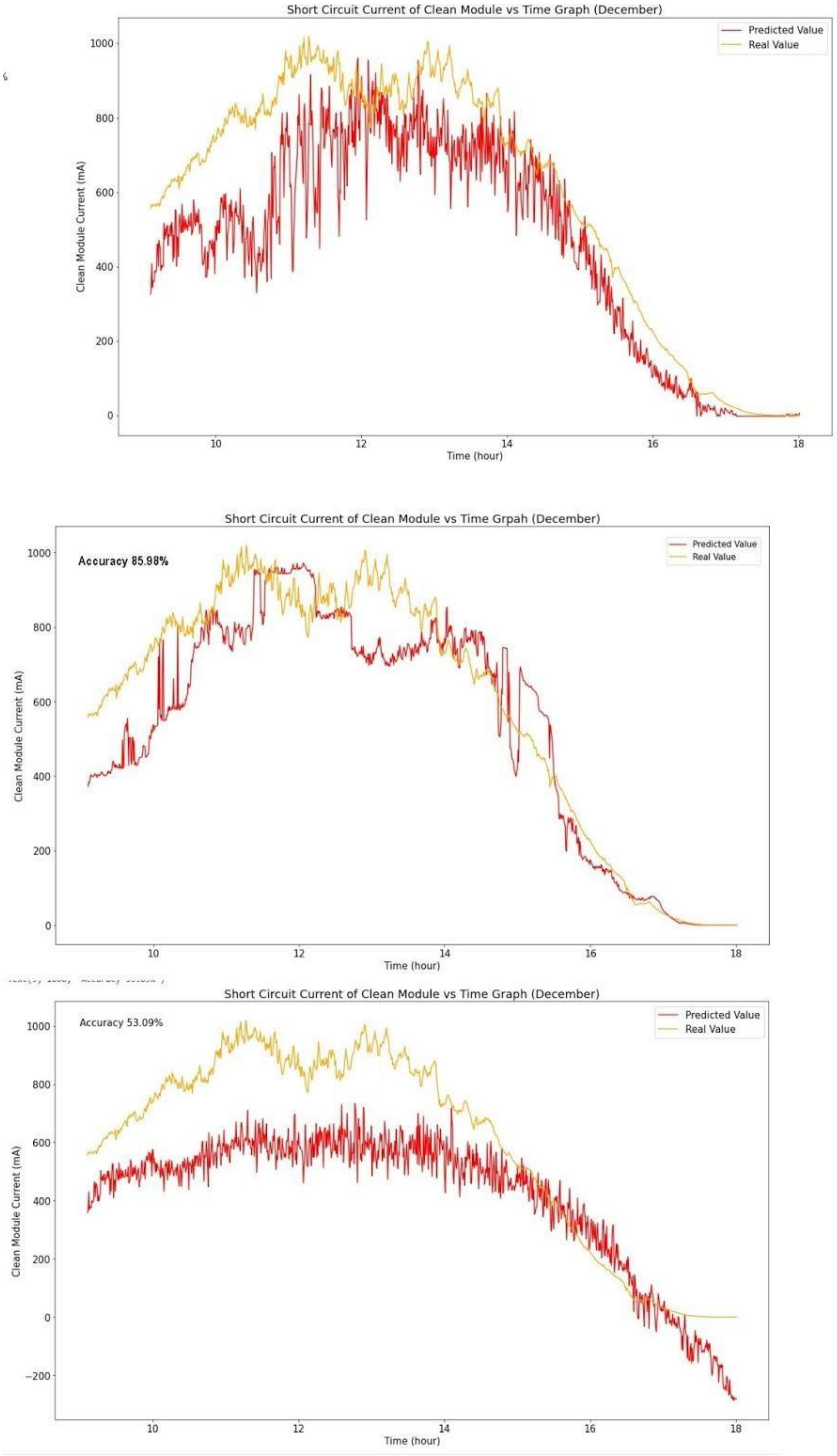


Figure 6.3: Clean Module Comparison for Dataset 2.

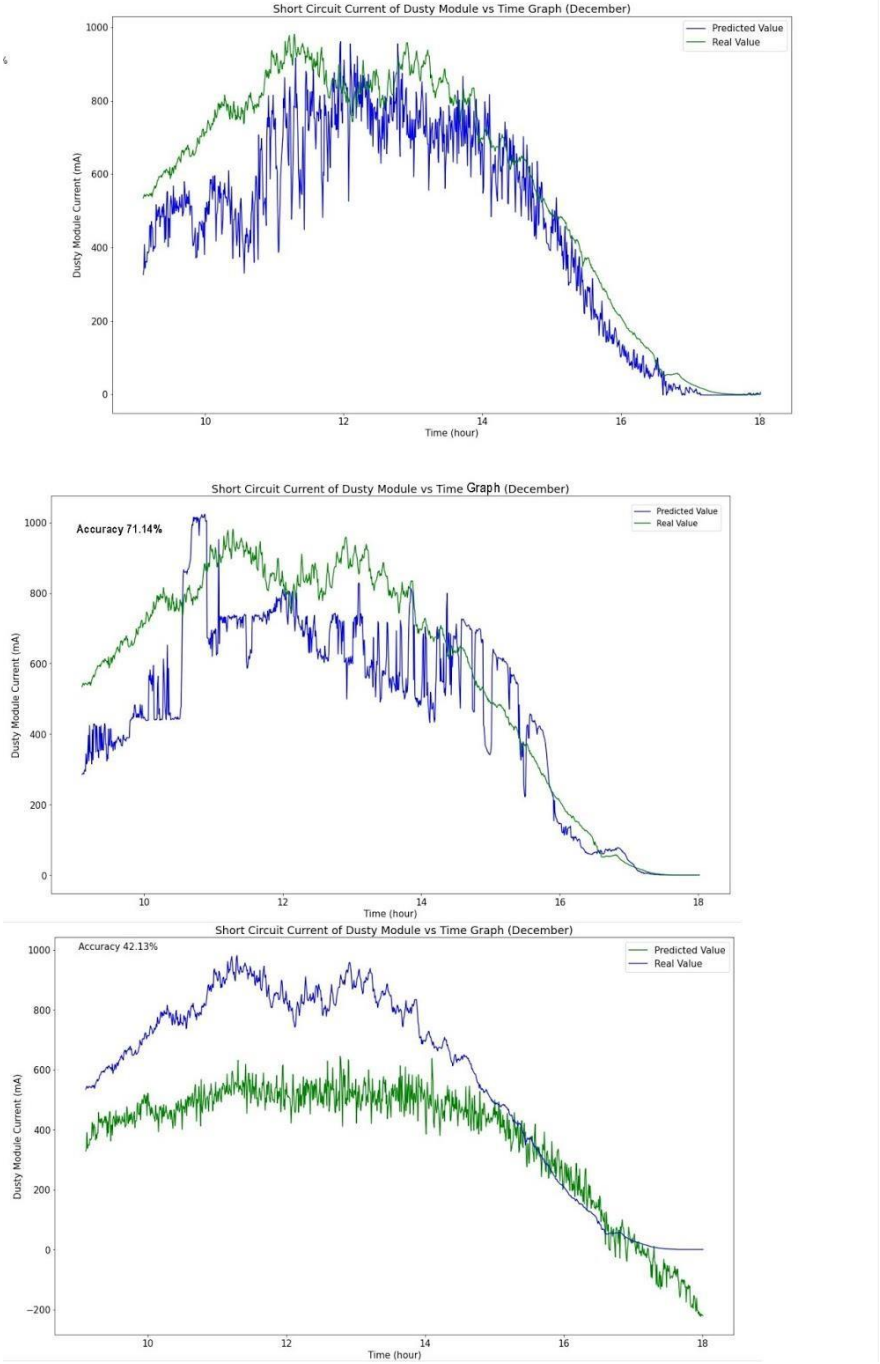


Figure 6.4: Dusty Module Comparison for Dataset 2.

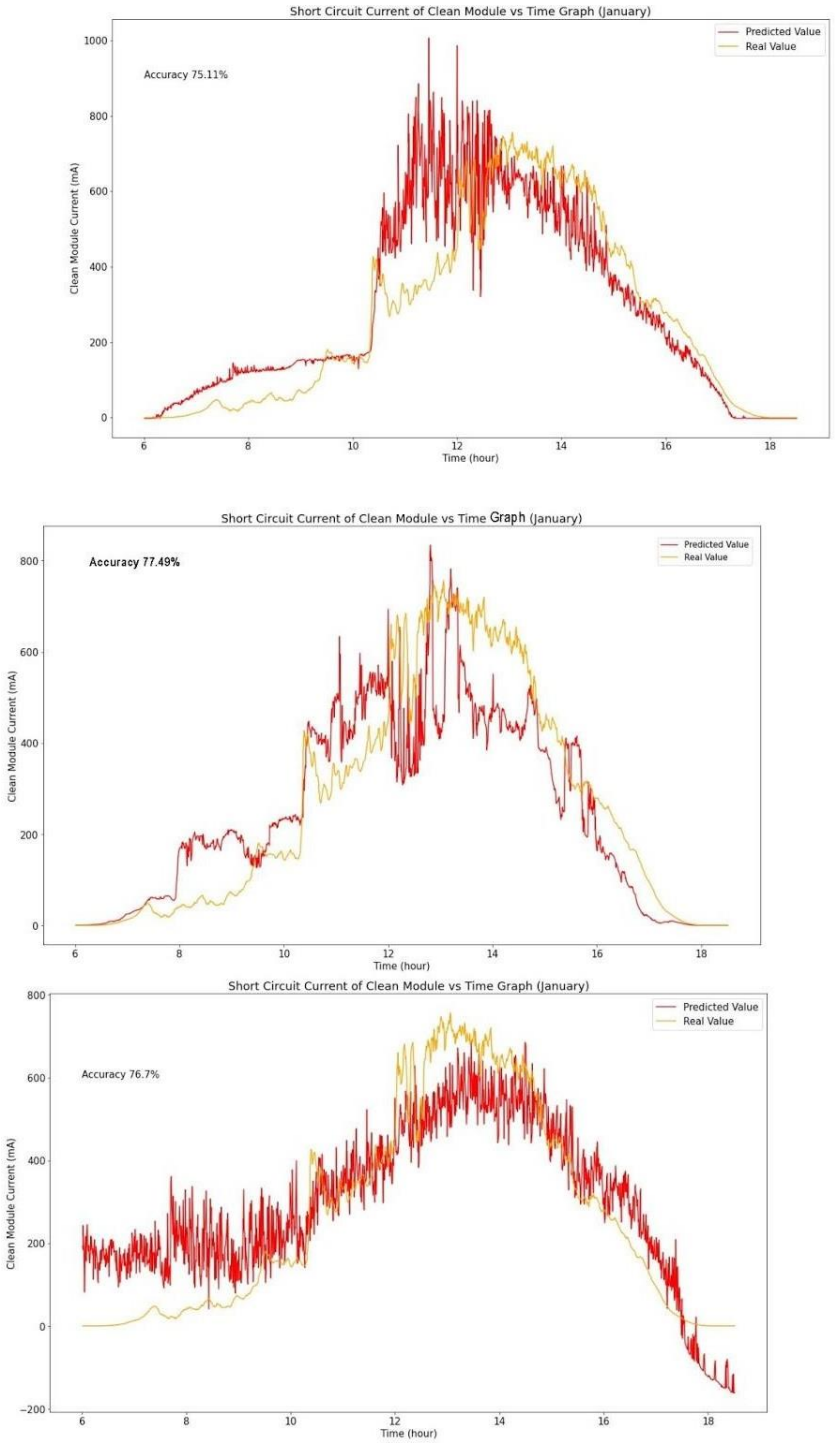


Figure 6.5: Clean Module Comparison for dataset 3

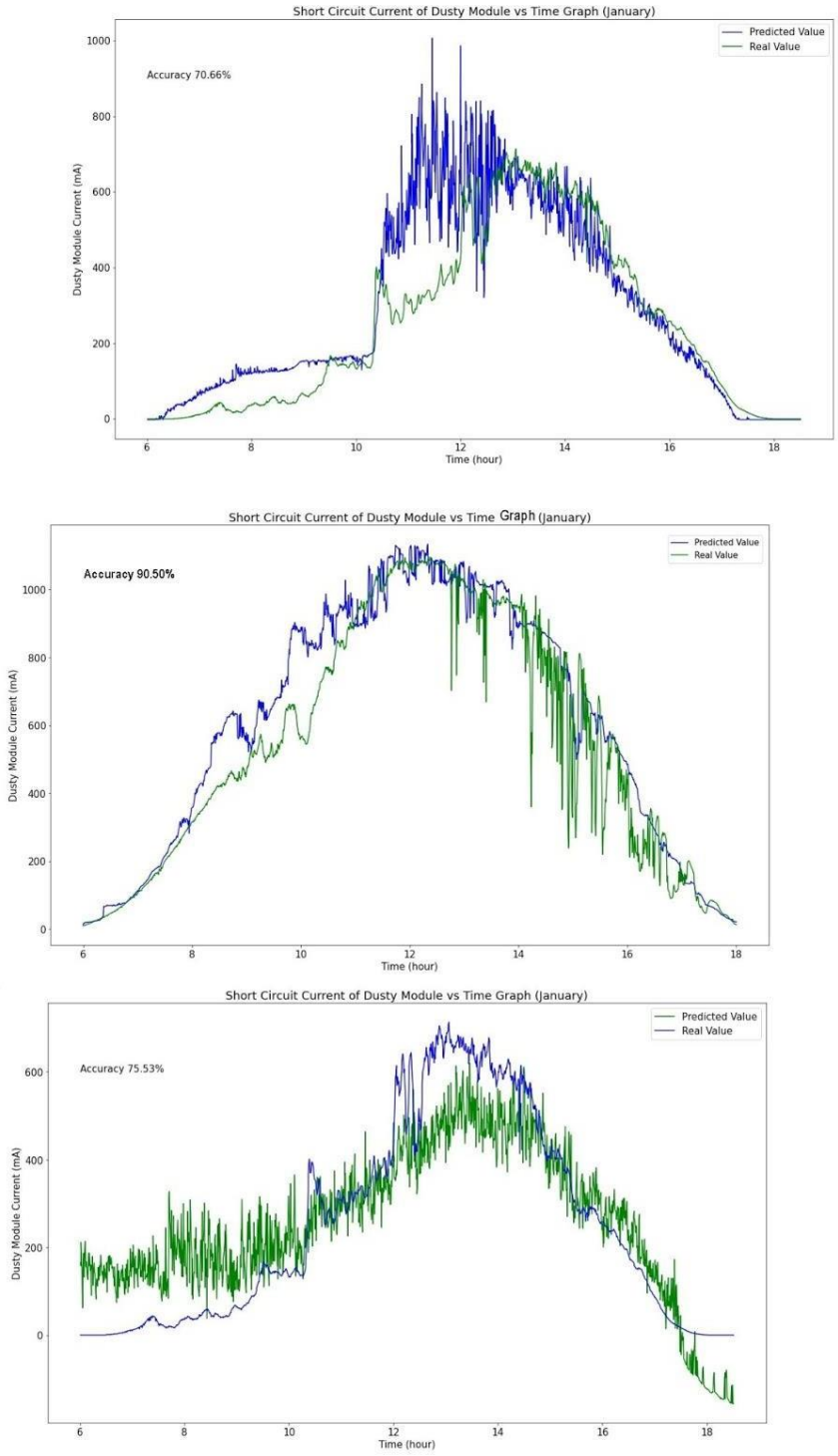


Figure 6.6: Dusty Module Comparison for dataset 3

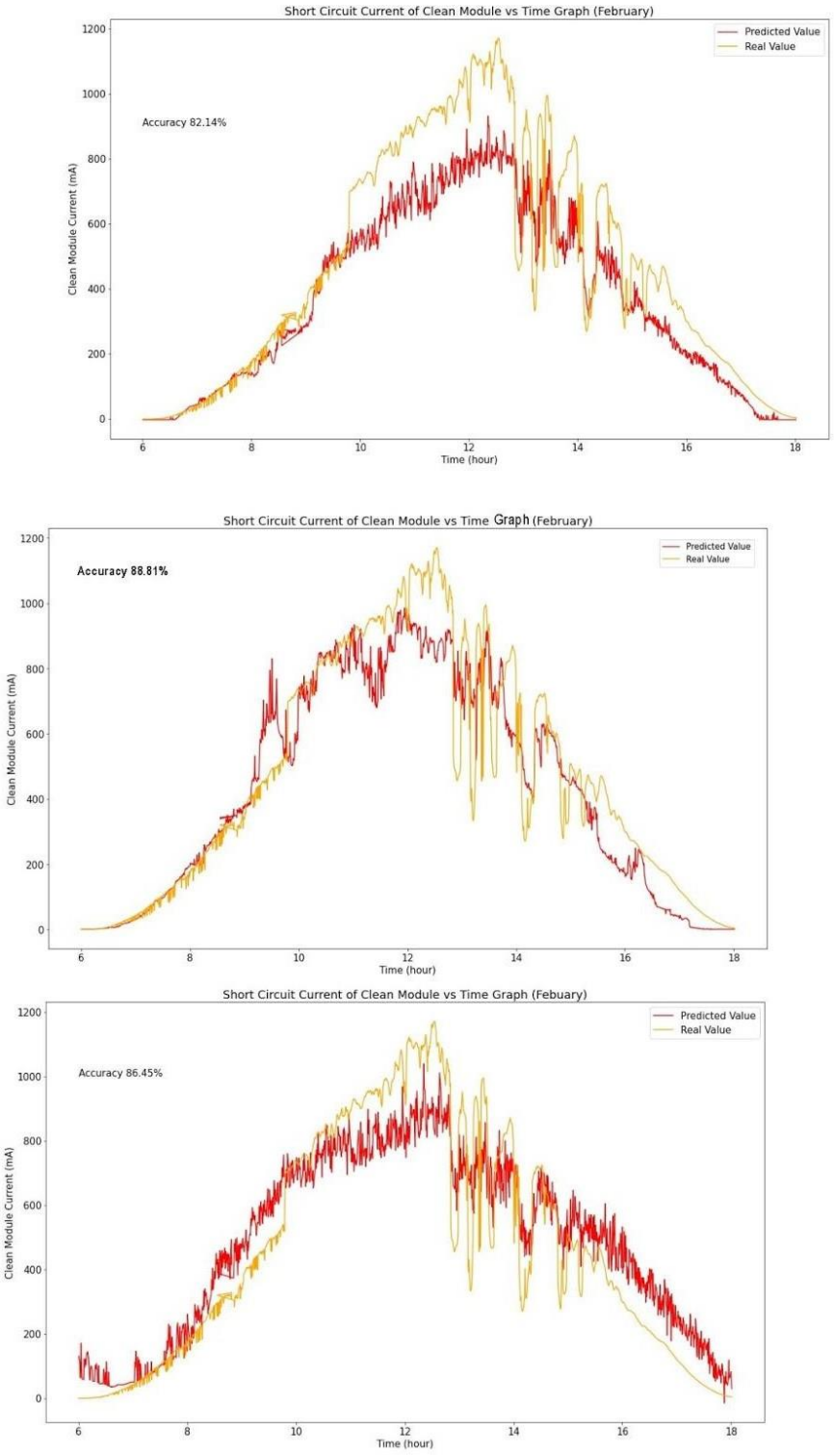


Figure 6.7: Clean Module Comparison for dataset 4

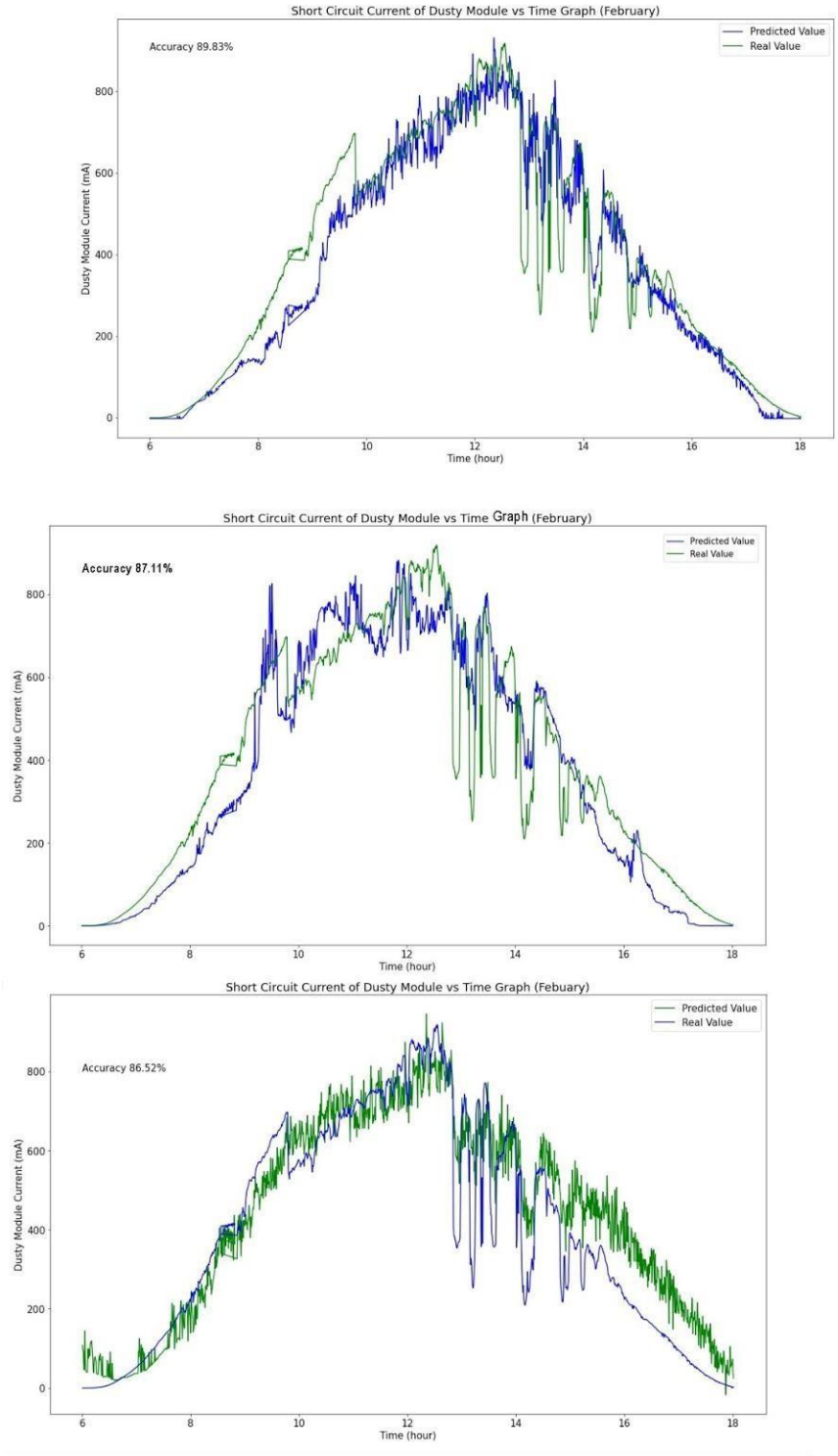


Figure 6.8: Dusty Module Comparison for dataset 4.

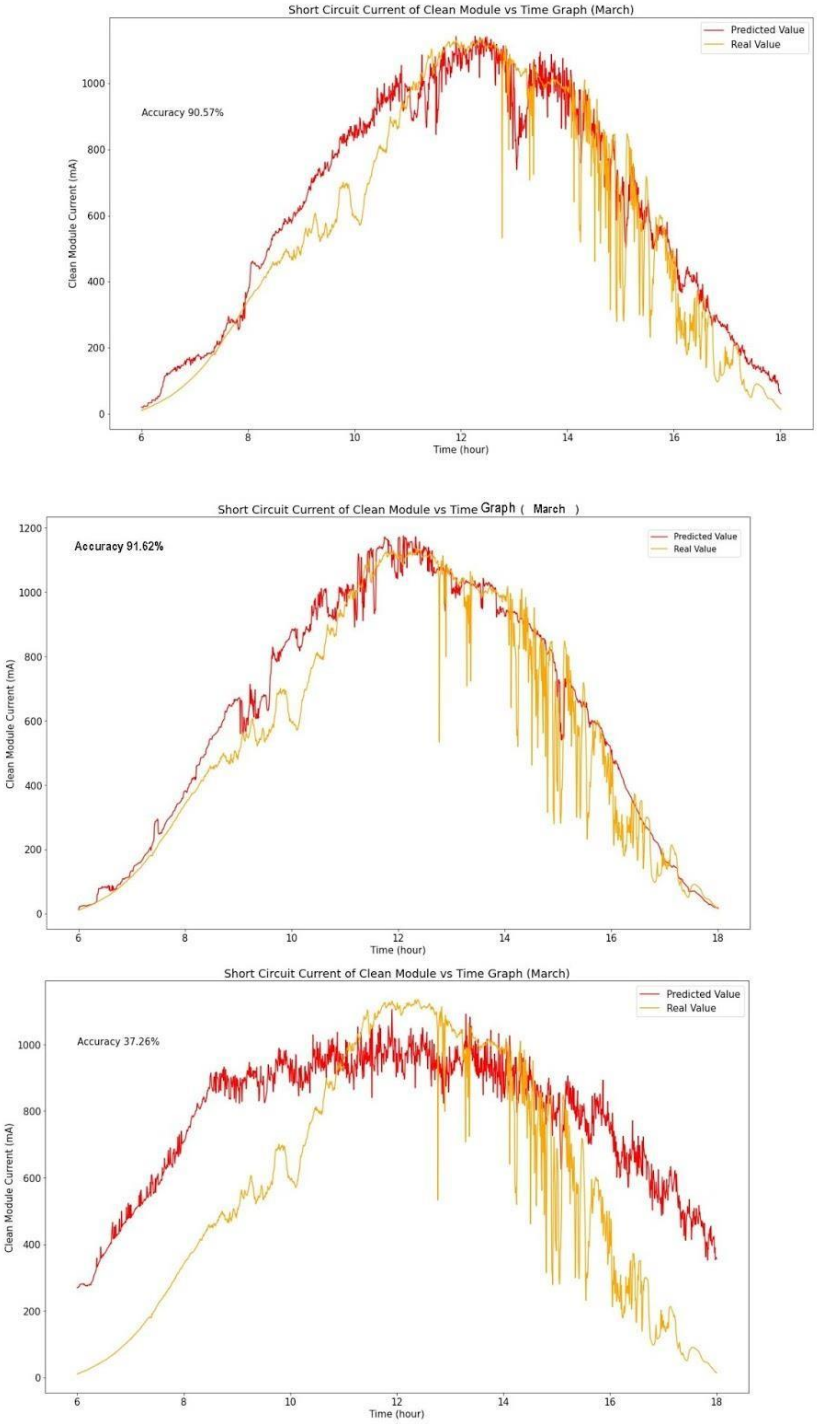


Figure 6.9: Clean Module Comparison for dataset 5.

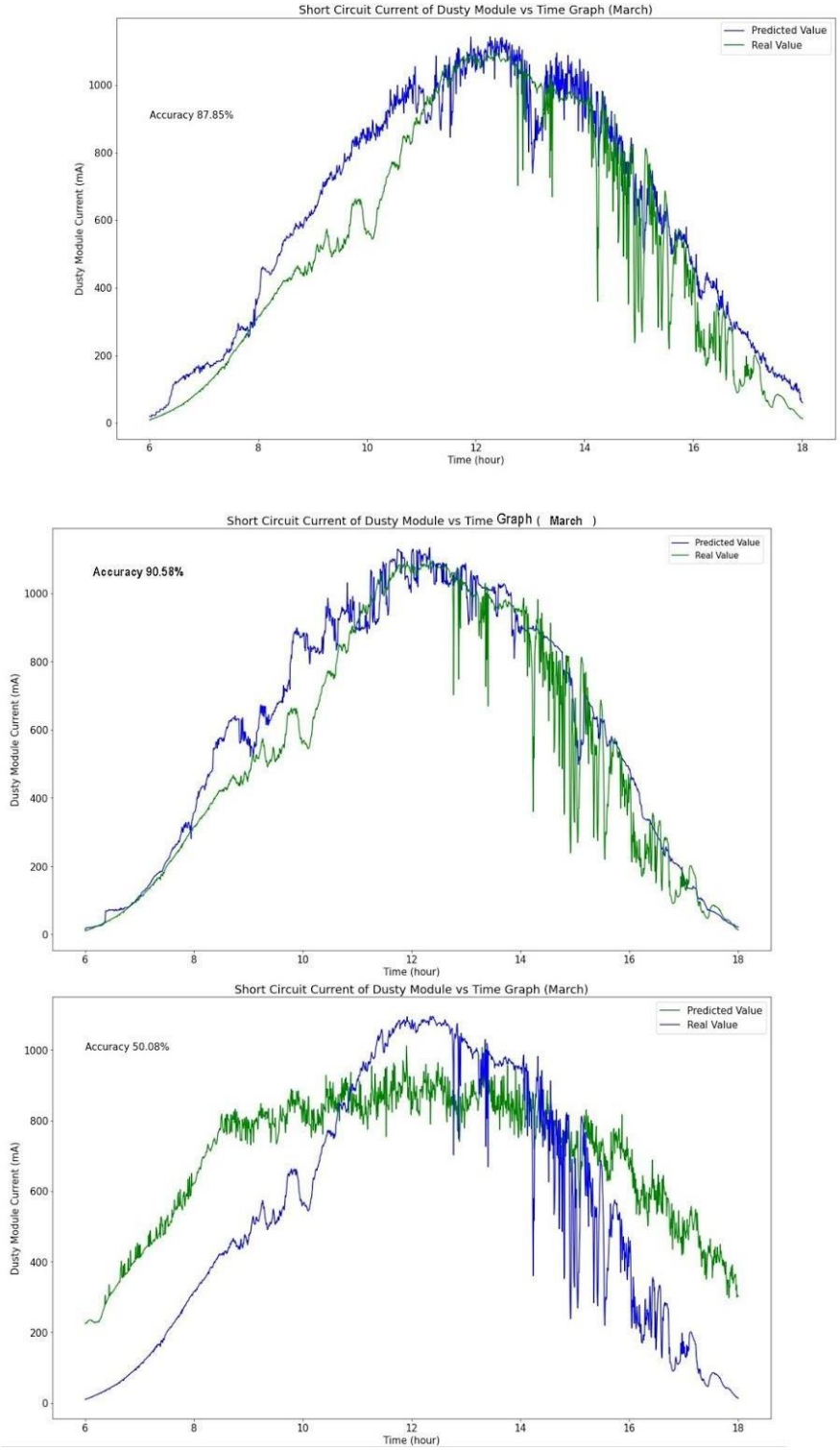


Figure 6.10: Dusty Module Comparison for dataset 5.

For training dataset 1, Artificial neural Network model performed best for clean module and dusty module. For training dataset 2, Random Forest model performed best for clean module and Artificial neural network model performed best for dusty module. For training dataset 3, Random Forest model performed best for both clean and dusty module. For training dataset 4, Random Forest model performed best for clean module and Artificial Neural Network model performed best for dusty module. For training dataset 5, Random Forest model performed best for both clean and dusty module. The sensors integrated in clean PV module as well as dusty modules have dispensed data of short circuit current for varying weather parameters. With the exploitation of several machine learning algorithms and analyzing the outcome, we have been piloted that, other than using linear regression method or artificial neural network, random forest method provides considerably superior accuracy.

For training dataset 1, Artificial neural Network model performed best for clean module and dusty module.

For training dataset 2, Random Forest model performed best for clean module and Artificial neural network model performed best for dusty module.

For training dataset 3, Random Forest model performed best for both clean and dusty module.

For training dataset 4, Random Forest model performed best for clean module and Artificial Neural Network model performed best for dusty module.

For training dataset 5, Random Forest model performed best for both clean and dusty module. The sensors integrated in clean PV module as well as dusty modules have dispensed data of short circuit current for varying weather parameters. With the exploitation of several machine learning algorithms and analyzing the outcome, we have been piloted that, other than using linear regression method or artificial neural network, random forest method provides considerably superior accuracy.

Table 6.1: Comparative analysis of the models (RF, ANN, MLR).

	Random Forest		**Artificial Neural Network**		**Multiple Linear Regression**	
Dataset	**Clean Module Accuracy(%)**	**Dusty Module Accuracy(%)**	**Clean Module Accuracy(%)**	**Dusty Module Accuracy(%)**	**Clean Module Accuracy(%)**	**Dusty Module Accuracy(%)**
1) Training dataset 1:1st November 2019 to28th November 2019	88.3	85.63	96.24	94.79	85.56	82.46
2) Training dataset 2:1st November 2019 to30th December 2019	85.98	71.14	73.38	78.42	53.09	42.13
3) Training dataset 3: 1st November 2019 to30th January 2020	77.49	90.50	75.11	70.66	76.7	75.53
4) Training dataset 4:1st November 2019 to26th February 2020	88.81	87.11	82.14	89.83	86.54	86.52
5) Training dataset 5:1st November 2019 to30th March 2020	91.62	90.58	90.57	87.85	37.26	50.08

Chapter 7

Conclusion

This thesis is collated and analyzed by the dataset we got in the effect of the weather parameters. In this work we predicted the shot circuit current for clean PV module and dusty PV module in terms of some weather parameters like temperature, wind speed, humidity, air pressure by ANN, RF, MLR this three Machine learning algorithm and then investigates all three methods which one's prediction shows better outcomes. In our study we saw in every training dataset Random Forest performs the best for both clean and dusty PV modules. On the other side, ANN is somewhat less accurate than RF, whereas MLR is the weakest.

7.1 Future work

In the study we collected the 5 months' data and we predicted the accuracy. In the result some variation has been seen as the data sit under three season and we saw some variation. In future if bigger dataset is given then the results can be more accurate for the machine learning models. A longer period dataset might help with the study of dust on both panels over the course of the year in the future. For the prediction of short circuit current on a given day, the ML model requires independent variables. As a result, in the future, collecting datasets over a longer time period will be immensely helpful to this ML model to get the higher accuracy. Not only for data analysis of weather predictions we can also use this process in medical study, financial institutions etc. we can use it as example in Share Market, future Price of the product like this in many sectors which could be more useful.

References

1. A. Khalyasmaa et al., "Prediction of Solar Power Generation Based on Random Forest Regressor Model," 2019 International Multi-Conference on Engineering, Computer and Information Science (SIBIRCON), 2019, pp. 0780-0785,
2. M. Meng and C. Song, “Daily Photovoltaic Power Generation Forecasting Model Based on Random Forest Algorithm for North China in Winter,” Sustainability, vol. 12, no. 6, p. 2247, Mar. 2020.
3. H. Shareef, A. Mutlag, A. Mohamed et al., "Random Forest-Based Approach for Maximum Power Point Tracking of Photovoltaic Systems Operating under Actual Environmental Conditions", Computational Intelligence and Neuroscience, vol. 2017, Article ID 1673864, 17 pages, 2017.
4. M. Abuella and B. Chowdhury, "Solar power probabilistic forecasting by using multiple linear regression analysis," SoutheastCon 2015, 2015, pp. 1-5,
5. Y. Kologlu, H. Birinci, S. Ilgaz, B. Ozyilmaz, (2018, July). A Multiple Linear Regression Approach for Estimating the Market Value of Football Players in Forward Position [Online]. Available: https://www.researchgate.net/publication/326171665_A_Multiple_Linear_Regression_Approach_For_Estimating_the_Market_Value_of_Football_Players_in_Forward_Position
6. “Types of solar panels,” July 15,2020. Accessed on: May. 15, 2021 [Online]. Available: https://www.energysage.com/solar/101/types-solar-panels/#:~:text=There%20are%20three%20major%20types,property%20and%20desired%20system%20characteristics.
7. Solar Cell I-V Characteristic, [Online]. Available: https://www.alternative-energy-tutorials.com/photovoltaics/solar-cell-i-v-characteristic.html . [Accessed May 10,2021].

8. Effect of Temperature, [Online]. Available: https://www.pveducation.org/pvcdrom/solar-cell-operation/effect-of-temperature. [Accessed May 16, 2021].
9. M. Meng, C. Song, “Daily Photovoltaic Power Generation Forecasting Model Based on Random Forest Algorithm for North China in Winter” 2020.
10. A. Khalyasmaa, S. A. Eroshenko, T. P. Chakravarthy, V. G. Gasi, S. K.Y. Bollu, R. Caire, S. K. R. Atluri, S. Karrolla “Prediction of Solar Power Generation Based on Random Forest Regressor Model” 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), 2019.
11. H. Shareef, A.M.Mutlag, A. Mohamed “Random Forest-Based Approach for Maximum Power Point Tracking of Photovoltaic Systems Operating under Actual Environmental Conditions”, Computational Intelligence and Neuroscience, 2017.
12. K. Fukunaga K “A Criterion and an Algorithm for Grouping Data” IEEE Transactions on Computers, C-19(10).
13. J. Brownlee “How to Develop a Random Subspace Ensemble” Available: https://machinelearningmastery.com/random-subspace-ensemble-with-python.
14. P. Grover “5 Regression Loss Functions All Machine Learners Should Know” Available: https://heartbeat.fritz.ai/5-regression-loss-functions-all-machine-learners-should-know.
15. N. Liberman “Decision trees and Random Forests” Available: https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991.
16. J. Brownlee “Bagging and Random Forest Ensemble Algorithms for Machine Learning” Available: https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning.
17. P. Sarkar “Bagging and Random Forest in Machine Learning” Available:

https://www.knowledgehut.com/blog/data-science/bagging-and-random-forest-in-machine-learning.

18. Bootstrap Aggregation, Random Forests and Boosted Trees, [Online], Available: https://www.quantstart.com/articles/bootstrap-aggregation-random-forests-and-boosted-trees/.
19. sklearn.ensemble.RandomForestRegressor,[Online], Available: https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
20. Decision Tree Classification Algorithm, [Online], Available: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.
21. R. Bala and D. Kumar, "Classification Using ANN: A Review", International Journal of Computational Intelligence Research, India, pp. 1811-1814, 2017.
22. B. Dickson, "What are artificial neural networks (ANN)?", TechTalks, 2021. [Online]. Available: https://bdtechtalks.com/2019/08/05/what-is-artificial-neural-network-ann.
23. Maad M. Mijwil, “Artificial Neural Networks Advantages and Disadvantages”, Iraq, pp. 1-2, 2018.
24. A. Choudhury, "TensorFlow vs Keras: Which One Should You Choose", Analytics India Magazine, 2019. [Online]. Available: https://analyticsindiamag.com/tensorflow-vs-keras-which-one-should-you-choose/.
25. Keras documentation: About Keras, [Online]. Available: https://keras.io/about/.
26. Rectified Linear Units (ReLU) in Deep Learning, [Online]. Available: https://www.kaggle.com/dansbecker/rectified-linear-units-relu-in-deep- learning.
27. Papers with Code- Leaky ReLU Explained, [Online]. Available: https://paperswithcode.com/method/leaky-relu.
28. Keras documentation: Layer activation functions, [Online]. Available:

https://keras.io/api/layers/activations/#tanh-function.

29. Mean squared error, [Online]. Available: https://en.wikipedia.org/wiki/Mean_squared_error.
30. Adam—latest trends in deep learning optimization, [Online]. Available: https://towardsdatascience.com/adam-latest-trends-in-deep-learning-optimization-6be9a291375c.
31. Linear Regression, [online]. Available: http://www.stat.yale.edu/Courses/1997- 98/101/linreg.htm . [Accessed April 15, 2021].
32. Multiple Linear Regression, [online]. Available: http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm . [Accessed April 15, 2021].
33. K. Jain, Scikit-learn(sklearn) in Python – the most important Machine Learning tool I learnt last year, January 5, 2015. Accessed on: April, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/.