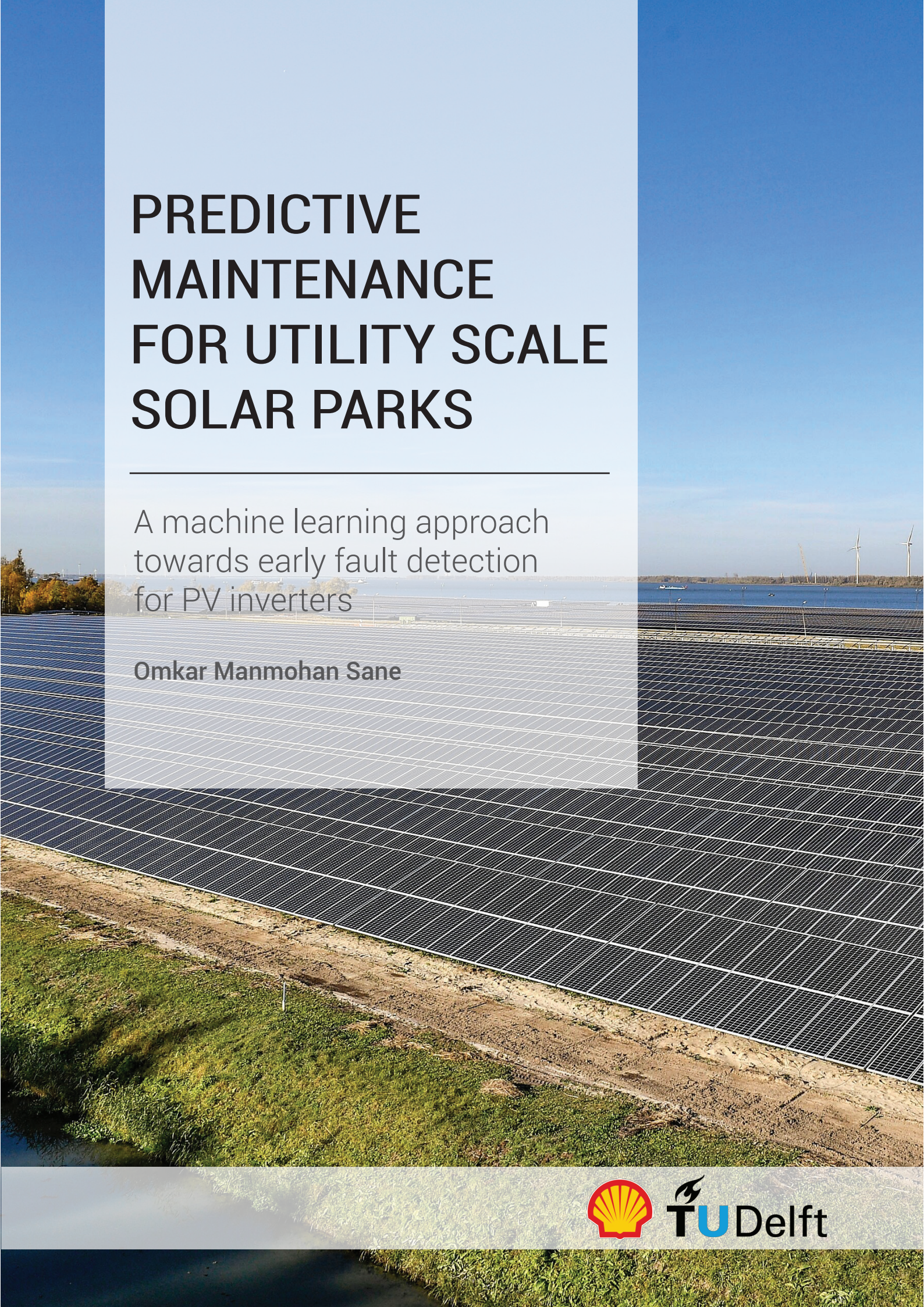

PREDICTIVE MAINTENANCE FOR UTILITY SCALE SOLAR PARKS

A machine learning approach towards early fault detection for PV inverters

Omkar Manmohan Sane

TU Delft

Predictive maintenance for utility scale solar parks

A machine learning approach towards early fault detection for PV inverters

by

Omkar Manmohan Sane

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday August 25, 2020 at 1:30 PM.

Student number:	4775163	
Project duration:	December 14, 2019 – August 25, 2020	
Thesis committee:	Dr. Olindo Isabella,	Head PVMD group-TU Delft,
	Dr. Hesan Ziar,	Assistant Professor-TU Delft,
	Dr. Tatiana Kozlova,	Researcher-Shell NERT,
	Mr. Vasileios Giagkoulas ,	Senior Data Scientist-Shell ,
	Dr. Simon Tindemans ,	Assistant Professor-TU Delft

This thesis is confidential and cannot be made public until December 31, 2022.

TU Delft

Preface

Delft, August 2020

"If you thought that science was certain - well, that is just an error on your part."- Dr Richard Feynmann. This statement has constantly pushed me to introspect on my understanding of science. I began this research with the aim of furthering my understanding of artificial intelligence(AI), particularly machine learning(ML). I believe that ML is a powerful tool which can help in accelerating the energy transition by drawing meaningful inferences from the data, that has been generated over the last few decades.

I'm extremely grateful to Dr Olindo Isabella and Dr Hesan Ziar, my supervisors from TU Delft, for their valuable insights and constructive recommendations towards this research. They were always ready to help and extended their support throughout this project. I would also like to thank Dr Simon Tindemans for our discussions on anomaly detection of inverters, even though the conversation was for a limited time, it steered my research in the right direction.

I would like to express my deep gratitude towards Dr. Tatiana Kozlova, my supervisor from Shell. She constantly guided and motivated me throughout my thesis. It has been my privilege to collaborate with her for this research. She not only helped me academically but was also very supportive of my personal growth.

I am also very grateful to Ir. Vasileios Giakoulas, a senior data scientist at Shell, for his guidance throughout the development of the algorithm. He has been extremely patient with answering all my questions. Our conversations on applicability of ML towards new energies has further bolstered my motivation to research towards combining ML and new energies.

I would like acknowledge the assistance of Ir. Benjamin Gaskill from the Shell asset management team . He answered my questions regarding the documentation of Moerdijk with utmost patience, and was always open to discussions. Our conversation on the mapping of faults was a turning point for this research.

A big thanks to the entire team of NERT for giving me a platform to showcase my results, for it helped me hone my presentation skills.

I very much appreciate the constant support of all my friends who stood by me under these testing times. I personally want to thank Vignesh Balasubramaniam for having faith in me whenever I doubted myself and encouraging me to push myself towards completion of this research. I would also like to thank Malavika Krishnan for designing the cover page.

Finally I could not have undertaken this research without the support of my parents who encouraged me to move to the Netherlands to pursue my masters at TU Delft.

Abstract

The growing demand and improvements in manufacturing capabilities, supported by government subsidies, has allowed the increase in the installed capacity of utility scale solar parks. Due to the remoteness in their location, the costs associated with dispatching personnel for maintenance is extremely high. A major contribution towards unscheduled downtime of these plants is due to the inverter faults. Currently, reactive and preventive maintenance are the most prevailing methods to identify and fix inverter faults. The presumption that the components will not under-perform or fail until the scheduled visit, leads to a significant loss of production and revenue. To deal with the disadvantages of current maintenance methods, the solar industry is very keen on understanding the possibility of early detection of inverter faults by implementation of predictive maintenance.

This research assessed the applicability of Machine Learning (ML) towards early signal detection of inverter faults in order to generate predictive maintenance alerts. The data for building the ML algorithms was acquired from a Shell owned 26.6 MWp utility scale solar park located in Moerdijk, The Netherlands. The early signal detection algorithm developed, was based on the comparison between the actual and the predicted active power. The model built to predict the active power was based on two supervised learning methods;Elastic Net and Gradient Boosting Machine (GBM) with quantile regression.

These models were capable of predicting the active power with a Mean Absolute Error (MAE) of 0.98kW & Root Mean Square Error (RMSE) of 1.8kW using Global Plane of Array irradiance (GPOA) and module temperature measurements available from the Moerdijk data. The early signal detection relied on differentiating between prediction error and actual error. A window was created to encompass the maximum extent of prediction errors to avoid any false positive signals. This window for elastic net was found to be $-\sigma$ on the lower side and 2σ on the upper side. Although when elastic net method was tested on 337 inverters- by looking at their residual variation 1-week prior to registered fault- it was found that the predictions suffered a periodic structural error. This was due to the erroneous predictions at times with extreme irradiance values. To mitigate, this the GBM with quantiles of 0.01 and 0.99 of GPOA was built to create a range of predictions giving rise to a wider range for normal operation. The results from both the algorithms indicated no early signals for inverter fault detection. This was partly due to data quality issues with fault tags in the Supervisory Control and Data Acquisition (SCADA) monitoring system; only 7 actual fault cases were identified. Additionally, the economic feasibility of implementing predictive maintenance was found to potentially reduce the current Operational Expenses (OPeX) by up to 10%. Despite the issues with data quality, an approach of using ML towards early fault detection for inverters in utility scale solar parks has been realised through this research.

Keywords: Machine Learning (ML) , fault signature, Elastic net, Gradient Boosting Machine (GBM), quantile loss, early signal, Supervisory Control and Data Acquisition (SCADA), fault tags, Operational Expenses (OPeX), predictive maintenance, utility scale,Artificial Intelligence (AI)

List of Figures

1.1 Methods for maintenance analysis[21] . . . 2
1.2 Process physics based flow chart . . . 2
1.3 Limit check for process values . . . 3

2.1 Failure values statistics over a span of 2 years from 350 PV systems in California,USA Golnas [18] 6
2.2 Failure values statistics over a span of 5 years from a 4.6 MWp plant in Arizona, USA [10] . . . 6
2.3 Failure values statistics over a span of 3 years from 202 PV systems in Taiwan [22] . . . 6
2.4 Types of maintenance . . . 7
2.5 Block diagram of simulation model based on state space averaging method[28] . . . 8
2.6 Flow chart for time series modelling [36] . . . 11
2.7 Schematic of a neural network ANN . . . 12
2.8 Flow chart for early inverter fault detection based on two years of operational data from a 6.2 MWp plant in Greece [11] . . . 13
2.9 Schematic of an ANN developed for early inverter fault detection using operational data from a 6.2 MWp plant in Greece [11] . . . 13
2.10 Hourly residuals generated based on the difference between actual and predicted values over a period of two years from a 6.2 MWp plant in Greece [11] . . . 14
2.11 Daily residuals upper and lower limits to determine the early fault detection of inverters using two years of operational data from a 6.2 MWp plant in Greece [11] . . . 14
2.12 Registered faults and predictive alerts from a 6.2 MWp plant in Greece [11] . . . 15
2.13 Unsupervised learning methodology towards early inverter fault detection adopted by Betti and Lo Trovato [7] . . . 15
2.14 KPI levels defined by Betti and Lo Trovato [7] for early fault detection of inverter . . . 16
2.15 Statistics of fault prediction for a single inverter Betti and Lo Trovato [7] . . . 16
2.16 Statistics of fault prediction for all inverters in the plant Betti and Lo Trovato [7] . . . 17

3.1 Drift fault [42] . . . 20
3.2 Malfunction fault [42] . . . 20
3.3 Simple linear regression . . . 21
3.4 Overfit and underfit . . . 22
3.5 k-fold cross validation [17] . . . 23
3.6 Decision trees . . . 23
3.7 Quantile regression . . . 24

4.1 Moerdijk Solar Park outline . . . 27
4.2 Schematic of electrical architecture . . . 28
4.3 PV panels . . . 28
4.4 Huawei inverter architecture . . . 29
4.5 Pyranometer mount . . . 30
4.6 Pt100 module temperature sensor . . . 30
4.7 SCADA . . . 30
4.8 Device drop-down . . . 31
4.9 Substation drop-down . . . 31
4.10 Fault log snip . . . 31
4.11 Count of faults in 2019 . . . 31
4.12 Moerdijk schematic . . . 32

5.1 Current annual OpeX cost comparison . . . 34
5.2 Snip from O&M service provider contract . . . 35

5.3 OpeX Cost distribution 36

6.1 Raw GPOA sensor data from a sensor located at 100TSR(s1) 40
6.2 Raw module temperature sensor data from a sensor located at 100TSR(s1) 41
6.3 Processed GPOA sensor data from all 5 GPOA sensors 42
6.4 Processed module temperature sensor data from all module temperature sensors 42
6.5 Major contributors for inverter downtime 43
6.6 Inverter fault taxonomy from Moerdijk with fault count in 2019 43
6.7 Count of *equipment failure* per inverter from the Moerdijk plant for the year 2019 45
6.8 Modified Count of *equipment failure* per inverter from the Moerdijk plant for the year 2019 . . . 45
6.9 Code snippet of allocation of inverters in substaions without sensors to a sensor cluster 47
6.10 Average monthly fit across all clusters to identify the choice of global/local, annual/monthly models 48
6.11 Selection of 50 inverters for modelling based on the MSE optimisation 48
6.12 Yearly variation of regression fit vs actual values 49
6.13 Predicted power values using linear regression vs actual power 49
6.14 Simple Linear regression annual residual variation 50
6.15 Feature weights in elastic net 51
6.16 Elastic net residuals 51
6.17 vVriation of actual, predicted power values during a day in a low irradiance month 52
6.18 Comparing annual and monthly Elastic Net models 52
6.19 Safe operating zone for inverters 53
6.20 *100EIN-0001-0001* Equipment failure timestamp with 24 hours historical variation showing an early signal 54
6.21 *102EIN-0003-0004* Equipment failure timestamp with 24 hours historical variation showing an early signal 54
6.22 *100EIN-0001-0001* Equipment failure timestamp with 24 hours historical variation 54
6.23 *102EIN-0003-0004* Equipment failure timestamp with 24 hours historical variation 55
6.24 GBM predictions based on quantile regression 55
6.25 *100EIN-0001-0001* Equipment failure timestamp with 24 hours historical variation 56
6.26 *100EIN-0001-0001* Equipment failure timestamp with 24 hours historical variation 56
6.27 Erroneous mapping of faults in the SCADA 57
6.28 Monthly variation of counts of equipment failure and AC fault 58
6.30 Actual equipment failure 58
6.29 Equipment failure preceded by AC fault shown by the dip in the frequency 59

List of Tables

2.1 Comparison of predictive maintenance methodologies for inverters 10

4.1 Fault taxonomy 32
4.2 Modified naming convention for substations 32

5.1 Current OPeX costs 33
5.2 Cost break down of O&M service provider contract 35
5.3 Add caption 36
5.4 Cost of deploying personnel 36
5.5 Dispatch Activation Matrix 36
5.6 MWh loss matrix 37

6.1 % of outliers 41
6.2 Explanation of fault taxonomy of SCADA fault tags used in the Moerdijk plant 44
6.3 Average P values obtained through two independent sample t-test for all inverter clusters 46
6.4 Count per cluster and Change in P value 47
6.5 Performance metrics of linear regression, elastic net, GBM with quantile regression 50

A.1 hyper-parameters for 0.01 quantile 66
A.2 hyper-parameters for 0.99 quantile 66

B.1 Library versions 67

Contents

List of Figures vii

List of Tables ix

Glossary xiii

1 Introduction 1
1.1 Background . . . 1
1.2 Research partners . . . 3
1.3 Research questions . . . 3
1.4 Report Outline . . . 3

2 Literature review 5
2.1 Failure history in PV sytems . . . 5
2.2 Inverter analysis . . . 6
2.2.1 Failure modes . . . 6
2.2.2 Current maintenance strategies . . . 7
2.3 Predictive maintenance . . . 8
2.3.1 Process physics based methods . . . 8
2.3.2 Signal-Processing based methods . . . 8
2.3.3 Process history based methods . . . 9
2.3.4 Comparative study of predictive maintenance methods for inverters . . . 10
2.4 Machine learning based predictive maintenance . . . 10
2.4.1 Forecasting techniques . . . 10
2.4.2 Previous research on predictive maintenance . . . 12

3 Machine learning 19
3.1 Basics of Machine learning . . . 19
3.2 Data pre-processing . . . 20
3.3 Modelling . . . 20
3.3.1 Regression methods . . . 20
3.3.2 Gradient Boosting . . . 23
3.4 Evaluation of models . . . 24
3.4.1 Coefficient of determination . . . 24
3.4.2 Mean Absolute Error (MAE) . . . 24
3.4.3 Mean Square Error (MSE) . . . 24
3.4.4 Mean Absolute Percentage Error (MAPE) . . . 25
3.5 Statistical tests . . . 25
3.5.1 t-Test . . . 25

4 Baseline data 27
4.1 Moerdijk . . . 27
4.1.1 Electrical architecture . . . 27
4.1.2 PV panel . . . 28
4.1.3 Inverter . . . 29
4.1.4 Measuring Instruments . . . 29
4.1.5 SCADA system . . . 30
4.1.6 Summary of Equipment's . . . 32

5 Economic analysis 33
5.1 Current operational expenses . . . 33
5.2 Economic feasibility of predictive maintenance . . . 34
5.3 Dispatch of personnel based on predictive alerts . . . 36

6 Results 39
6.1 Data Pre-processing. 39
6.1.1 Feature pre-processing. 39
6.1.2 Target pre-processing 42
6.2 Baseline Model 47
6.2.1 Linear Regression model 48
6.3 Advanced modelling results 50
6.3.1 Elastic Net 50
6.3.2 Early fault detection using Elastic Net 53
6.3.3 GBM with quantile loss function 55
6.3.4 Early fault detection using GBM 56
6.4 Advanced fault analysis 56

7 Conclusion & Discussion 61

8 Recommendations 63
8.1 Recommendations to Shell 63
8.2 Research Recommendations 63

A Hyper-parameter details 65
A.1 Hyper parameter optimisation 65
A.1.1 Elastic net hyper parameters 65
A.1.2 GBM hyper-parameters 65

B Code Base 67

Bibliography 69

Glossary

AC Alternating current. 6, 39

AI Artificial Intelligence. v

ANN Artificial neural network. vii, 12, 13

ANNOVA Analysis of Variance and Mean. 25

API Application Programming Interface. 63

AR Auto Regression. 10

ARIMA Auto Regression Integrated Moving Average. 10

ARMA Auto Regression Moving Average. 10

DC Direct current. 39

DLL Daily Lower Limit. 13

DUL Daily Upper Limit. 14

EWMA Exponentially Weighted Moving Average. 9

GBM Gradient Boosting Machine. v, viii, ix, xii, 20, 39, 50, 55, 56, 58, 59, 61, 65, 66

GHI Global Horizontal Irradiance. 29

GPOA Global Plane of Array irradiance. v, viii, 11, 29, 32, 39, 40, 42, 46–48, 50, 53, 61

IGBT Insulated-Gate Bipolar Transistor. 5–7

KPI Key Performance Indicator. vii, 15, 16

LASSO Least Absolute Shrinkage and Selection Operator. 22, 65

LCL Lower control limit. 2, 53, 59

MAE Mean Absolute Error. v, xi, 19, 20, 24, 50

MAPE Mean Absolute Percentage Error. xi, 25

ML Machine Learning. v, 2, 5, 9, 12, 17, 19, 20, 61, 64

MSE Mean Square Error. viii, xi, 24, 48

NERT New Energies Research & Technologies. 3

NWP Numerical Weather Prediction. 11

OEM Original Equipment Manufacturer. 28, 31

OPeX Operational Expenses. v, 33, 34, 61

PCB Printed Circuit Board. 6, 7

PV PhotoVoltaic. vii, xi, 1, 2, 5, 6, 8–12, 15–17, 62

RMSE Root Mean Square Error. v, 19, 20, 50, 52

RSS Residual Sum of least Squares. 21, 22

SCADA Supervisory Control and Data Acquisition. v, 17, 39, 43, 56, 63

SDM Supervisory Diagnostic Model. 15

SOM Self Organising Map. 15

TDR Time Domain Reflectometry. 8

UCL Upper control limit. 2, 53, 59

1 Introduction

1.1. Background

The American psychologist Abraham Maslow created a hierarchy pyramid of human needs. According to Maslow's pyramid food and water are the most fundamental needs for basic human survival. However, the modern man now also relies extensively on energy and more specifically electricity for his survival in the new social systems. Therefore electricity has become as essential as the other basic human needs. Electricity plays a major role in dictating the development of nations. Today, a significant portion of our energy comes from finite sources like fossil fuels. In addition to their scarcity in the recent years, the Paris agreement of 2015 aims to reduce the carbon emissions of the world by 40% . To overcome the future bottlenecks of such sources, solar energy has been explored in the last few decades to establish it as a mainstream source of electricity generation. For solar energy to compete with conventional sources, large scale systems such as utility parks are required. The installed capacity of such utility parks has increased from 40GW to 400GW in the last 10 years. As per a recent report by NREL Fu et al. [14], the price of utility solar plants during last decade have plummeted from 4.6 $/Wp to 1.06 $/Wp. This decrease in the price as well as governmental subsidies has led to an exponential growth in the size of utility scale PhotoVoltaic (PV) parks. In December 2019, a 2050 MWp park was commissioned in India[37], the park has 200,000 PV panels with 300 central inverters. The increased number of components and the remote locations of these large parks contributes to added risk. In addition, the lack of movable parts in a solar park reduces the value of maintenance amongst PV park developers. Therefore, in order to ensure a stable grid integration, reliability of PV parks becomes a crucial aspect.

Reliability is a function of failure modes, although PV parks can have a multitude of failures; majority of unscheduled downtime in solar parks can be attributed to inverters.Formica et al. [13] claims 43% of maintenance tickets correspond to inverters. This downtime contributes to a notional loss of 10% and the financial consequences are significant.

Typically solar parks have a maintenance scheduled every six months. During which the personnel inspect the components, and perform repairs and replacements if required. This approach is known as preventive maintenance, as components are evaluated or replaced based on predefined intervals. Additionally some components might fail before the scheduled maintenance visit and hence trigger a reactive maintenance to replace the component. **Therefore, solar park operators are keen on looking at solutions that could provide them with early fault notifications**. In principle predictive alerts should be complemented with the expected time horizon for the failure of the component, as this would enable the operators to evaluate the time horizon for the site visit of the personnel. This method is known as predictive maintenance as the maintenance requirements are generated based on actual performance rather than set rules. An overview of maintenance analytic is shown in figure 1.1. Descriptive methods are the simplest approach to maintenance analytics, as they provide information about *what has happened*. It includes reports on the failure events which have happened, which can be useful for reactive maintenance. Diagnostic methods take this approach further by providing information about why the component failed. Predictive methods provide insights into what will happen in the future instead of focusing on past events. However, they do not state the sequential actions that should be carried out when a fault is predicted. This is achieved by prescriptive methods that involve

specific knowledge of the component. The shift from descriptive methods towards prescriptive methods not only increases the reliability of the PV parks, but also paves the way for greater autonomy in the operations.

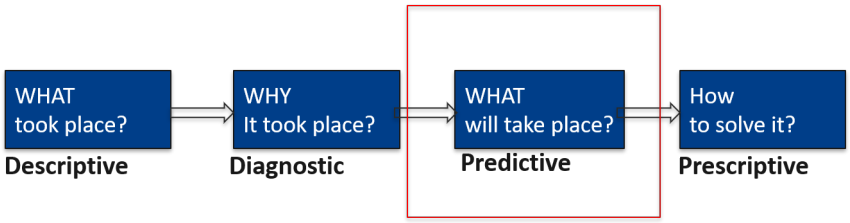

Figure 1.1: Methods for maintenance analysis[21]

Since prescriptive methods can only follow the availability of predictive alerts, this research focused on predictive maintenance. This thesis explored the various predictive maintenance methods such as Process physics based methods, Signal processing based method, Process history based methods of which **Process history based data driven methods involving machine learning, was chosen as the focus of this study.**

Predictive maintenance can be defined as a method where the occurrence of fault in a component is known prior to the actual occurrence. The probability of the occurrence depends on the accuracy of the algorithm. The equations which govern the physical process can be used to generate expected values of the output based on the inputs. These expected values can then be compared with the actual output to decide if the component is operating normally. This figure 1.2 shows the flow chart for this method. However, for an inverter to be modelled using equations, it could become very complex, because of the associated grid response . In addition to that, the equations require the measurement of all variables, but in practice only the major variables such as power are measured. Furthermore, it is difficult to obtain predictive alerts with this method. Therefore machine learning methods can help to model the normal operation of the inverter based on it's previous performance and predict the faults.

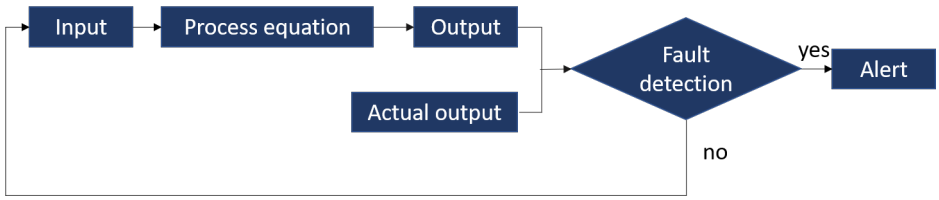

Figure 1.2: Process physics based flow chart

In order to implement a predictive maintenance using machine learning it is important to have fault data. The quality and quantity of the available data has a significant effect on the performance of the model. A performance parameter such as active power or voltage must be chosen as a decision variable for modelling. The anomalies in ML methods are detected by out of bound samples. The difference between actual and modelled values is known as the residual (res). To differentiate the modelling error from an actual fault an Upper control limit and Lower control limit are defined based on the distribution of the residuals. If the residual visibly crosses the limits before the timestamp of a failure, then a predictive alert can be generated. This is illustrated in figure 1.3, where the residual is monitored based on the UCL and LCL, if fault occurs at the point marked with a red dot then the algorithm would be able to detect the fault in advance because the residual crossed the UCL before the actual occurrence of the fault. This enables the algorithm to create a predicitve alert.

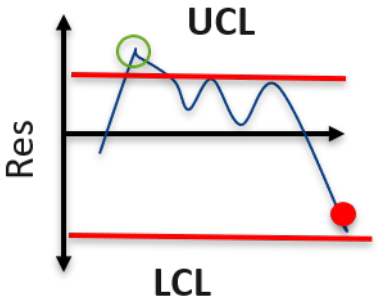

Figure 1.3: Limit check for process values

1.2. Research partners

This research project is carried out under PVMD research group of TU Delft, in collaboration with Shell New energies team. The solar business division of Shell is looking to increase solar projects portfolio and New Energies Research & Technologies (NERT) provides R&D support to the solar business. To increase the reliability of the existing and new parks that will be built, a study on predictive maintenance was required. The solar business provided the data from Moerdijk PV asset which is a Utility scale solar park owned by Shell. Additionally the digitalisation team was involved in supporting the development of the algorithm. Together with all internal stakeholders, the aim of this research was to create an in house understanding of predictive maintenance solutions for utility scale solar parks. The research was supported by Dr. Olindo Isabella and Dr. Hesan Ziar from PVMD group, Dr. Tatiana Kozlova and Ir. Vasileios Giagkoulas from Shell.

1.3. Research questions

The objective of this research was to develop a predictive maintenance algorithm using supervised machine learning methods in utility scale solar parks. The main research questions based on the research objectives are listed below:

- *For the given data should the model be global or local, additionally should it be annual or monthly?*
- *Which loss function is more suitable towards prediction of faults?*
- *Is predictive maintenance economically feasible to implement in the Moerdijk case?*

1.4. Report Outline

This report is structure into 8 chapters including the introduction. Chapter 2 discusses the available literature on maintenance and early fault detection. This is followed by the explanation of theoretical background on machine learning in chapter 3. Chapter 4 consists of the information regarding the baseline data of Moerdijk. It is followed by economic analysis of implementing predictive maintenance for Moerdijk in chapter 5. Chapter 6 presents the results of the two methods that were used. Chapter 7 presents the conclusions . Finally, chapter 8 presents the recommendations for Shell and for further research.

2

Literature review

This chapter explains the need to investigate the early failure detection for predictive maintenance of inverters. A PV farm consists of multiple components such as PV panels, inverters, Balance of System (BoS) and sensors . Consequently, it is important to attribute the failure to a specific component. The section 2.1 describes previous studies which have recorded the inverter to be the responsible component for majority of downtime over different plants. Based on the observations in section 2.1 the component of interest for this research was the inverter. Furthermore the section 2.2 describes the software and hardware related failure modes of inverters. Additionally, section 2.2 discusses the challenges associated with reactive and preventive maintenance, which are currently the prevailing maintenance strategies for PV parks.
To mitigate the drawbacks of current maintenance strategies, section 2.3 presents an alternative method of predictive maintenance. It elaborates on the possibility of predicting the failures beforehand, using different methods such as physics-based methods, signal processing methods and process history based methods. Building on the advantages of process history based ML methods, section 2.4 explains the methodology and the previous research conducted in this domain.

2.1. Failure history in PV sytems

The phases in the development of a PV park consists of installation, commissioning and operation[30]. Although failures can occur throughout the value chain, only the failures related to the operation phase are discussed in this section. Individual component failure during the operation phase differs in failure modes. The PV panels suffer from glass breakage, hotspots which may lead towards lower performance or ultimate failure of the panels [27]. Whereas inverters can fail due to the failure of Insulated-Gate Bipolar Transistor, monitoring system, or due to overheating. The established failure modes contribute significantly towards the down time during the operation of the system. A study was performed by Golnas [18] over 350 commercial systems through a course of 2 years to identify the major cause of downtime of the systems. In order to assess the effect of every component on the loss of production, a ticket was issued to record the cause and timestamp of the failure. The figure 2.1 lists the % of tickets corresponding to the % of production loss due to the individual components. It shows that 43% of the failure tickets correspond to inverters, making them the prime contributors to the downtime.

Failure Area	% of Tickets	% of kWh lost
Inverter	43%	36%
AC Subsystem	14%	20%
External	12%	20%
Other	9%	7%
Support Structure	6%	3%
DC Subsystem	6%	4%
Planned Outage	5%	8%
Module	2%	1%
Weather Station	2%	0%
Meter	1%	0%

Figure 2.1: Failure values statistics over a span of 2 years from 350 PV systems in California,USA Golnas [18]

A similar study was conducted by Collins et al. [10] on a 4.6 MWp plant over a span of 5 years and the % of downtime is shown in the figure 2.2. A supplementary study performed by Huang et al. [22] on 202 PV system over a span of 3 years and the results are tabulated in 2.3.

Failure Area	% of Tickets
Inverter	53%
AC Subsystem	14%
DC Subsystem	14%
Module	12%
Other (lightning)	7%

Figure 2.2: Failure values statistics over a span of 5 years from a 4.6 MWp plant in Arizona, USA [10]

Failure Area	% of Tickets
Inverter	60%
Balance of System Components	28%
PV Modules	12%

Figure 2.3: Failure values statistics over a span of 3 years from 202 PV systems in Taiwan [22]

The values of inverter failure at 43%, 53% and 60% from the respective research establishes the fact that the inverter is responsible for maximum downtime in a solar plant. This justifies the choice of inverter as a key research component. The failures in an inverter could occur due to a multiple reasons which are discussed in the next section.

2.2. Inverter analysis

The subsection 2.2.1 explains the failure modes linked to the sub components of an inverter. The subsection 2.2.2 elaborates on the current strategies for resolving the inverter faults.

2.2.1. Failure modes

The reliability of an inverter depends on the performance of each sub component and subsequently failure studies do not treat the inverter as a black box. The inverter components that usually suffer failure can be as simple as fans and contactors, but can also be more complex, like capacitors, Printed Circuit Board (PCB), AC fuse and Insulated-Gate Bipolar Transistor (IGBT). A grid-connected PV system may handle a high level

of power flow and operating temperature. This degrades the IGBT and PCB which decreases reliability due to an increased risk of failure. This assertion is supplemented by Kaplar et al. [23] and Zhang et al. [50], both the researches indicate that the most recurring failure modes for power electronics is related to higher exposure to electrical and thermal stress. On the other hand, contrary to claim made by Petrone et al. [35], capacitors do not fail as often as suspected unless there is a systemic design or manufacturing issue[18]. In addition to the hardware failures, there can be software related issues due to bugs in the control algorithm. As the in-depth understanding of the faults is beyond the scope of this research, a list with limited description of faults is provided.

Hardware failure modes

- Capacitor failure : The inverter can either have an electrolytic or a film capacitor. Electrolytic capacitors are notably more prone to failures than film capacitors [23].
  - The electrolytic capacitors fail due to poor sealing combined with high internal temperatures. It leads to the vaporization of the electrolyte causing an instant failure.
  - The film capacitors failure is very rare and occurs only under extremely high voltage spikes.

  Although film capacitors fail rarely, they aren't widely used because of their higher costs in comparison to electrolytic capacitors.
- IGBT failure : The electrical and thermal stress are the reasons for failure in IGBT. During these high stress conditions the operating conditions exceed the specifications provided by the manufacturer. The fatigue due to the electro-mechanical stress causes the solder joint and the bond wire to fail[33][34].

Software failure mode:
In practice, if the inverter operates after a manual restart then it is considered that the failure occurred due to a bug in the control software. Additionally, if the firmware is not updated as per the manufacturer's recommendation then it can lead to a communication failure, resulting to a shutdown of the inverter.

2.2.2. Current maintenance strategies

Maintenance is broadly classified into two groups reactive and proactive, this classification is shown in figure 2.4.The solar industry predominantly relies on reactive and preventative maintenance[9]. This is due to the prevailing notion that "since there are no moving parts, we can install it and forget it" [9].

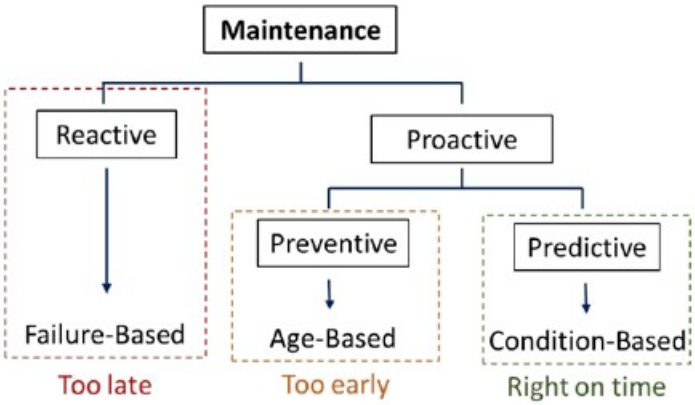

Figure 2.4: Types of maintenance

Reactive Maintenance

As the name suggests, the component is inspected after failure, and if it is completely damaged then the component is replaced. This method of maintenance may not be cost effective for remotely located, utility scale solar parks as the cost of dispatching personnel at the last moment is significantly higher. If there is an in-house team, instead of dispatching a team, then the operational expenses of the plant could increase by at least 25%, Kurtz et al. [25] . Additionally, the lost production significantly affects the revenue; if the plant has a contractual obligation to provide a certain amount of energy then failing to do so could result in significant fines.

Preventive maintenance

Preventive maintenance is a part of proactive maintenance which is based on a set of rules. The simplest rule applicable in the PV farms is the scheduled visual inspection of all components over a period of 6 months. Subsequently, a preventive maintenance approach also gives room to schedule maintenance based on the current monitored performance of the inverter. According to a research conducted by Lillo-Bravo et al. [27] 4.26% of all energy losses is due to a complete failure, whereas 20% of energy losses is due to operational inefficiencies. The periodicity of 6 months fails to capture this inefficiency and hence can be detrimental for the lifetime of the plant.

Although literature[46],[10] claims that the maintenance costs for PV plants are minimal , it does not take into account the loss of revenue due to inefficiencies and last minute deployment. To mitigate this problem predictive maintenance can be a paramount tool. It is discussed elaborately in the section 2.3.

2.3. Predictive maintenance

In contrast to conventional solar maintenance strategies, predictive maintenance aims to detect the occurrence of faults well before their actual occurrence. This is usually done by comparing the actual power output to an expected output used as a reference point. The fault diagnosis under predictive maintenance can further be split into process physics based methods, signal processing based methods, history based. methods[12].

2.3.1. Process physics based methods

In physics based modelling, the main idea is to compare actual operation data with the reference values. Accurate reference data is computed by available equations which leverage the knowledge of characteristic parameters of the components. For modelling of inverters, state space averaging models are created. A simple representation of state space equation is shown in equation 2.3.1 . These are highly complicated, and therefore in- depth understanding is not presented in this research. Information on the state space modelling can be found in [28]. The figure 2.5 shows variables involved in analytical modelling of inverters.

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t)\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \quad (2.1)$$

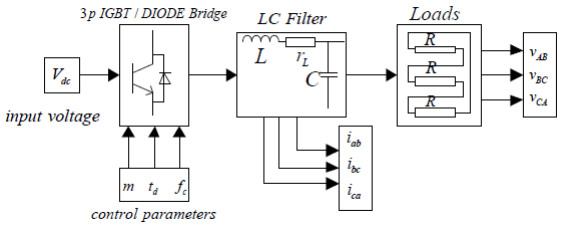

Figure 2.5: Block diagram of simulation model based on state space averaging method[28]

In addition to the complexity of the equations, the availability of a continuous data of every parameter is also a major constraint. These reasons make analytical or physics based modelling a sub optimal choice for generating reference values.

2.3.2. Signal-Processing based methods

The commonly available method of fault diagnosis using signal processing is Time Domain Reflectometry (TDR). This method was initially developed to locate faults in high voltage transmission lines. Subsequently, this approach was adopted towards PV inverter fault detection[40]. A high frequency wave generator sends a wave into the inverter and the reflected wave is compared to the initial signal. This extent of correlation between the incident and reflected signal helps in detecting and localizing the faults. Despite the fact that TDR can detect, localize, and diagnose faults there are two major drawbacks [12]

- There is a need for sophisticated tools to introduce the input signal and analyze the reflected signal hence it leads to higher costs.

- The possibility of early detection is diminished as the comparison between the incident and reflected signal is almost instantaneous.

Several other sophisticated signal processing based methods such as discrete wavelet transform have been discussed in Alam et al. [2], but all of them suffer from similar disadvantages as mentioned above.

2.3.3. Process history based methods

Unlike the signal processing based methods, process history based methods do not require any special equipment. Process history based methods are composed essentially of two main approaches - Statistical and Machine Learning approaches [35],[46],[3].

Statistical approach

The statistical approach can apply both descriptive and inferential statistics to detect fault occurrence. Additional information on both statistical methods can be found in Stapor [44]. There is extensive literature available for fault detection, which uses statistical quality control methodologies using the process data. Garoudja et al. [15] describes the procedure on how statistical quality control is implemented in PV plants. The EWMA control chart has been implemented by Garoudja et al. [15]. The working principle of EWMA can be found in [20]. Assuming that $x1, x2..x_n$ are the monitored process logged observations, then the statistics of EWMA control chart is computed as shown below :

$$z_t = x_t + (1-\lambda)Z_t - 1t > 0z_0 = \mu_0 \tag{2.2}$$

μ_0 : Process mean at normal conditions
μ_0, λ are the parameters of fault free data (normal operation)
x_t the process value at t
z_t is the EWMA output
The higher the value of λ the lower is the weight of the historic values. A small value of λ is suitable to detect relatively small changes in the process mean, while a large value detects relatively large changes in the process mean. Using the control chart, the monitored process will be declared out of control at time t if its monitored statistic is outside the control limit boundaries, which are generally defined as :

$$LCL = \mu - L\sigma_{z_t} \tag{2.3}$$

$$UCL = \mu + L\sigma_{z_t} \tag{2.4}$$

$$\sigma_{z_t} = \sigma_0 \tag{2.5}$$

$$\sqrt{\frac{\lambda(1-(1-\lambda^2 t))}{2-\lambda}} \tag{2.6}$$

LCL : Lower control limit
UCL : Upper control limit
L : confidence limit
σ_0 : Standard deviation of fault free process
Based on the equations above z_t the monitored process variable is deemed to be faulty if it crosses the LCL or the UCL. Since the statistic is compared only to one timestamp prior to the current value, this method cannot be used to detect faults with a larger lead time.

Machine learning based methods

The Machine Learning method can predict the fault based on the time window that is controlled by the user, unlike statistical methods where the fault detection is almost instantaneous. In recent times, due to improvement in computing, the ML approach has gained traction for fault detection. If we are able to leverage the data available from past failures, then by using predictive maintenance we can improve the performance and reliability of the plants. In the renewable energy domain, wind energy has been a pioneer in implementing data driven predictive maintenance and the solar PV industry is catching up[9]. The first step to build any predictive model is the identification of failures of components[24]. Which in this case would be the failure of the inverter recorded by the data acquisition system. It cannot be modelled as a stand alone classification problem as we want to track the pattern of the signal, with a significant lead time. Henceforth, the expected power produced by the inverter needs to be modelled as a continuous time series. The difference between the predicted and actual value is then further used to classify the possibility of early fault detection.

2.3.4. Comparative study of predictive maintenance methods for inverters

Type		Description
Process physics based		Can accurately model the inverter but the quality of the output depends on the assumptions of the analytical equations. Additionally many variables may not be recorded would be required to have continuous predictions.
Signal processing based		These methods can accurately predict the location of fault but require sophisticated equipment which may be expensive. The lead time of detection very low making it unsuitable for predictive maintenance.
Process history based	Statistics based	Simple method with simple equations, however the fault can only be detected one time stamp prior to actual occurrence making it less suitable for predictive maintenance.
	ML based	These methods unlike analytical methods rely on available variables, and can also identify a pattern in the data way before the fault with the assumption that there is an inherent pattern to prior to the fault.

Table 2.1: Comparison of predictive maintenance methodologies for inverters

The comparative study in table 2.1 shows that the machine learning methods could be a suitable solution to develop predictive maintenance algorithms for inverters. Additionally the exponential rise in computational power in the last few years and the availability of large datasets make a compelling case in favor of machine learning based predictive maintenance.

2.4. Machine learning based predictive maintenance

Machine learning based methods rely on the availability of the data. They work based on the principle that whenever there is a deviation between the predicted and the actual value of the target variable then this indicates a possible fault. Since the first step is creating a suitable forecast based on available features, the subsection 2.4.1 discusses the types of forecasting techniques. That is followed by a previous research that has already been done in the field of predictive maintainence for inverters.

2.4.1. Forecasting techniques

The availability of an accurate forecast ensures the applicability of the forecasted values towards an early detection of faults by tracking the patterns prior to the occurrence of the fault. An optimal design of solar PV plant can be created with the help of accurate forecasting techniques. There have been several recent advancements in the field of solar PV forecasting. These methods can broadly be classified into five methods : time-series, physical, regression, ensemble and neural network methods.

- Time series methods : It can be defined as a sequence of observations on a parameter measured at successive points in time [48]. The method is used to reconstruct the relations between the past and current parameters. The model does not require the knowledge of system physics for modelling [49]. Generic time series modelling includes methods like Auto Regression (AR), Auto Regression Moving Average (ARMA), Auto Regression Integrated Moving Average (ARIMA). A simplistic flow chart explaining the time series modelling is shown in figure 2.6. Supplementary information regarding these methods can be found in [6, 26, 39]

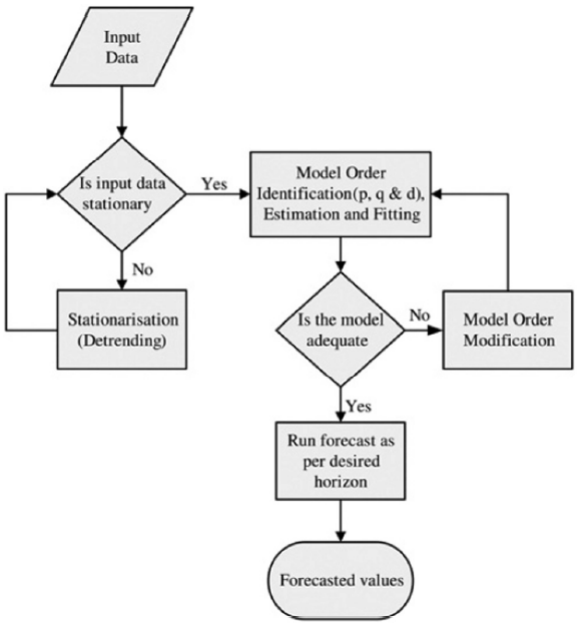

Figure 2.6: Flow chart for time series modelling [36]

- Physical methods : In this method meteorological data is used, which removes the necessity of sensor data at the surface. However, these models require a sophisticated device to convert the meteorological data into corresponding usable data. The applications of physical methods vary from very short to long horizons based on a underlying set of mathematical equations that describe the physical state and dynamic motion of the atmosphere [29]. These methods are dependent on Numerical Weather Prediction (NWP), sky imagery and satellite imaging [31]. They can be broadly classified as global and mesoscale methods, depending on the fraction of the simulated atmosphere which can either be worldwide or localised[29] . As the performance of the model is highly dependent on the weather conditions, the model performs sub optimally for extreme weather conditions. Furthermore another shortcoming of the method is that a resolution of only up to 16–50 km can be attained[4].

- Regression methods : The current research on using regression techniques to estimate the energy yield by PV systems focuses on short term intervals .The regression aims to model a relationship between local weather conditions such as time of the day, sky-cover, pressure, and wind-speed and the yield. This can be done either using linear regression or polynomial regression based on the relationship between parameters and decision variable to achieve accurate forecasts. In the context of this research the regression model was built on the available features of Global Plane of Array irradiance (GPOA), module temperature and active power of the inverter. The details of regression techniques have been explained in chapter 3.

- Ensemble methods : Ensemble methods assist in solving the weakness of individual methods which enhances their accuracy [4]. These methods also have a multitude of loss functions which are more suitable to solar forecasting [43]. Therefore, these methods were utilised in this research. Similar to the regression techniques the forecast model was created based on available features and targets. Further explanation of ensemble methods can be found in chapter 3.

- Neural network : These methods are very accurate in their predictions, although to use a neural network the critical mass of data that is required is very high [38]. A neural net consists of three parts : input, hidden and output layers. The input layer consists of the features, the hidden layers are optimised depending on the complexity of the problem. Consequently the output layer registers the result of the model. The schematic of a simple neural net is shown in figure 2.7, which showcased the input,hidden and the output layers. Since the data available in this research is limited and did not exceed the critical mass for a neural net, this approach was avoided.

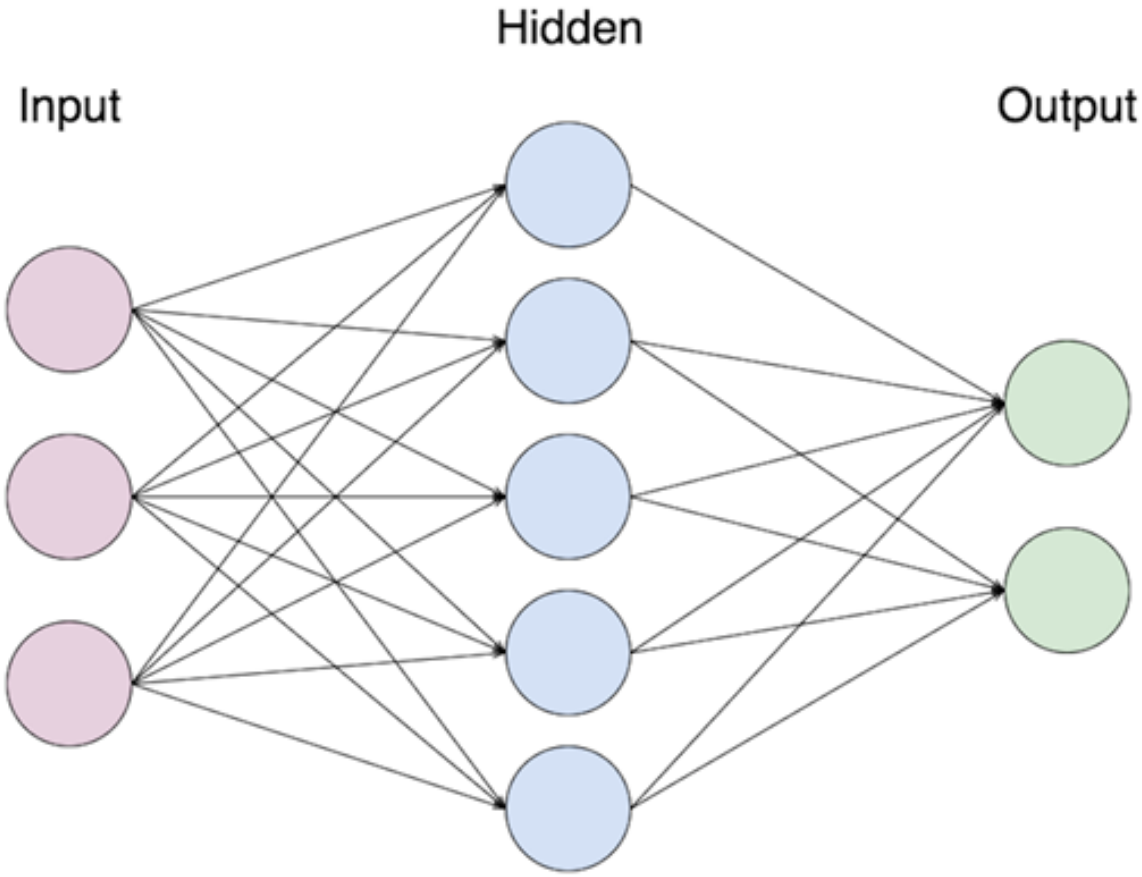

Figure 2.7: Schematic of a neural network ANN

2.4.2. Previous research on predictive maintenance

Although predictive maintenance is a key research topic for improving the operational efficiencies of PV farms, very limited research has been conducted in this domain. Only a handful of researchers have tried to develop predictive maintenance algorithms for inverters. Supervised and unsupervised machine learning approaches have been used by the researchers. Further information regarding these ML methods can be found in the chapter 3.1. The methodology and findings of two researchers have been discussed below.

Supervised learning approach

Research by De Benedetti et al. [11] claims that their model, which was built on two years of data, from a 6.2 MWp plant in Greece, can predict the occurrence of an inverter failure 7 days in advance with an accuracy of 90%. The data resolution was five minutes. Their approach is shown in figure 2.8.

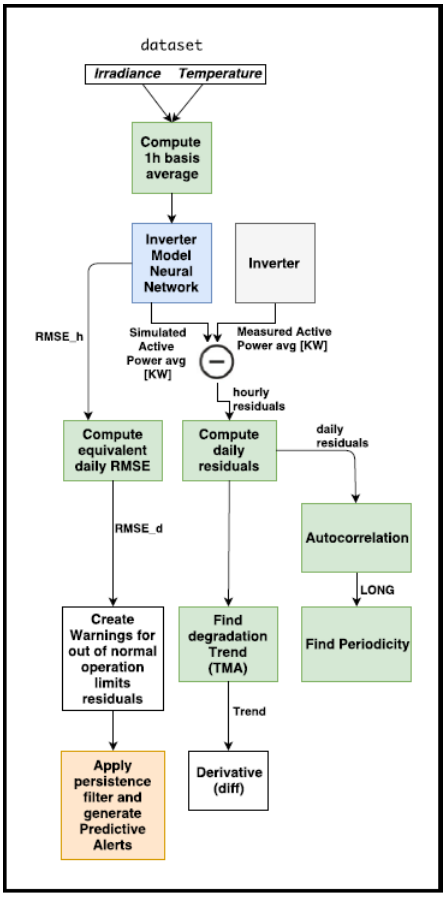

Figure 2.8: Flow chart for early inverter fault detection based on two years of operational data from a 6.2 MWp plant in Greece [11]

The five minute data was aggregated to an hour so as to create the model. Using irradiance and module temperature as features and active power of the inverter as a target variable, an Artificial neural network (ANN) was trained. The figure 2.9 shows how the neural network was modelled to predict the power produced by the inverter. Subsequent to the prediction, as shown in 2.8 difference between the predicted and actual power generation was calculated and was called residuals. Since the identification of fault required a distinction between modelling error and an actual possibility of fault, a safe operating window for the inverters was defined. This window basically tried to include all the possible modelling errors. If the inverter residuals were within the defined limit then no fault warning was triggered. The operating window was obtained by the validation error of the neural network. It consisted of an upper and a lower limit. The hourly lower limit was $3\sigma_h$, where σ_h was the validation error of the ANN and the hourly upper limit was $5\sigma_h$. These hourly residuals with their safe operating window are shown in figure 2.10.

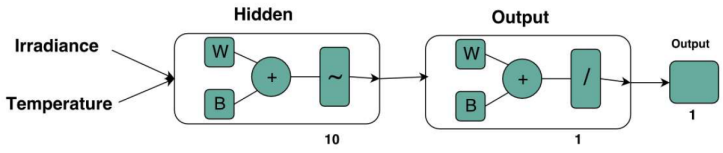

Figure 2.9: Schematic of an ANN developed for early inverter fault detection using operational data from a 6.2 MWp plant in Greece [11]

Since the occurrence of the faults was very sparse in time, the long term degradation patterns were required to be tracked . In order to establish the long term trends, the hourly residuals were converted to daily residuals and then daily limits defined the normal operation of the inverter. The normal operating window was again based on the aggregation of the validation error of the ANN.
Daily Lower Limit (DLL)= $3k\sigma_d$

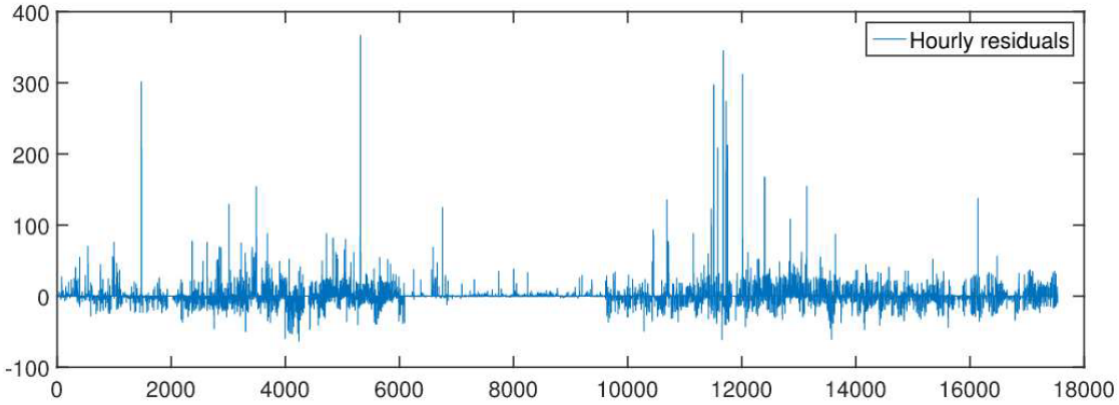

Figure 2.10: Hourly residuals generated based on the difference between actual
and predicted values
over a period of two years
from a 6.2 MWp plant in Greece [11]

Daily Upper Limit (DUL)= $5k\sigma_d$
Where k was the daily equivalent standard deviation index and k=4 was the optimal value in the study their study. The figure 2.11 shows the variation of daily residuals with their corresponding limits. Every instance when the residual crossed the DUL an alert was raised and depending on the discretion of the operator an action was planned.

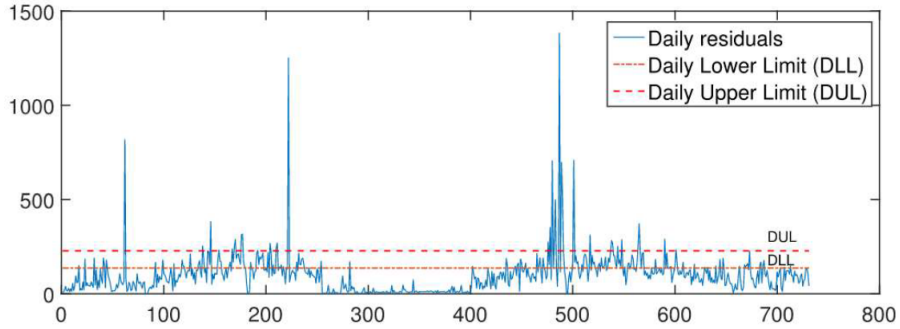

Figure 2.11: Daily residuals upper and lower limits to determine the early fault detection of inverters using two years of operational data from a 6.2 MWp plant in Greece [11]

The figure 2.12 shows the actual registered faults and the predicted alerts generated for plant. The valid predictive alerts are the successful predictions, i.e the algorithm correctly predicted the occurrence of the fault, whereas the false positives are the unsuccessful predictions.

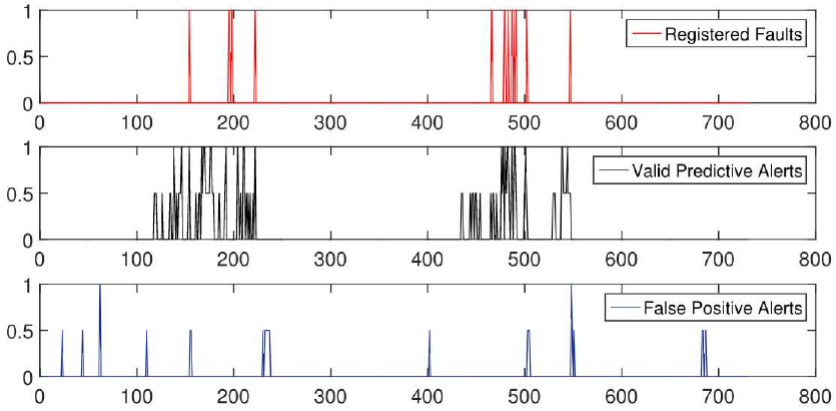

Figure 2.12: Registered faults and predictive alerts from a 6.2 MWp plant in Greece [11]

The method developed by De Benedetti et al. [11] was to create an hourly prediction model for the active power of the inverter using the irradiance and module temperature. Furthermore, it was compared to the actual generation to obtain residuals. The subsequent step was to determine the safe operating limits for the inverters. If the residuals were beyond the control limit then a predictive alert was raised. Although this method seems robust to create predictive alerts, it did not answer a very fundamental question of *what the cause for the fault was.* Additionally De Benedetti et al. [11] does not mention any information regrading the type of registered faults. This information is important to be stated because if the fault was caused by a disturbance in the grid then a prediction based on the weather conditions would be insignificant. This is due to the fact that a grid related fault needs to be dealt by looking at all the generators and loads. Some of the above mentioned shortcomings of De Benedetti et al. [11] were overcome by [7] by using a different method. Their research is discussed in the next subsection.

Unsupervised learning method

In contrast to the De Benedetti et al. [11] method which depended on supervised learning to predict faults, Betti and Lo Trovato [7] used an unsupervised learning approach. The data for the study was sourced from six different PV plants spread across Romania and Greece. The algorithm is trained to predict generic faults and specific faults based on the class of available faults. The methodology of building the predictive model is shown in figure 2.13.

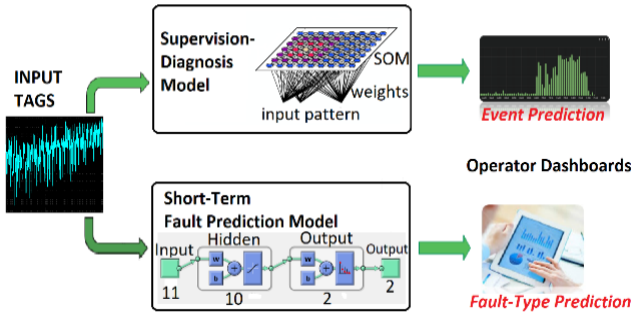

Figure 2.13: Unsupervised learning methodology towards early inverter fault detection adopted by Betti and Lo Trovato [7]

The Supervisory Diagnostic Model (SDM) was built using an unsupervised clustering approach based on a Self Organising Map (SOM). The mathematical term for SOM is kohnen maps. These reduce the dimensionality of the input features. In their research the eleven input features of various meteorological parameters were reduced to simple two dimensional data. A 20X20 SOM was used to detect the type of fault, a Key Performance Indicator (KPI) was defined according to the equation shown in equation 2.7.

$$KPI = \sum_{i,j} {P_{i,j}}^T EST \frac{1 - |\,{P_{i,j}}^T RAIN - {P_{i,j}}^T EST\,|}{1 - |\,{P_{i,j}}^T RAIN - {P_{i,j}}^T EST\,|} \tag{2.7}$$

Depending upon the KPI of the inverter the class and severity of the fault were decided. The figure 2.14 shows the levels of warning based on the KPI. For instance if the derivative of KPI crossing the threshold persisted for a day then it would be classified as warning of level 1. The higher the persistence of the derivative of KPI the more significant is the warning. The figure 2.15 shows the single inverter faults and the corresponding sensitivity of prediction.

Warning Level (w)	Crossing of Threshold	KPI Derivative	Persistence
1	$3\sigma_{KPI^{TRAIN}}$	< 0	1 d
2	$3\sigma_{KPI^{TRAIN}}$	< 0	2 d of w 1
3	$5\sigma_{KPI^{TRAIN}}$	< 0	1 day
4	$5\sigma_{KPI^{TRAIN}}$	< 0	2 d of w 3

Figure 2.14: KPI levels defined by Betti and Lo Trovato [7] for early fault detection of inverter

In figure 2.16 global performance overview on the PV park inverters is shown for classes AC Switch Open, DC Ground Fault and Thermal Fault. Additionally the accuracy was shown by the green bars and it was found to be around 80%. When few fault instances were available, sensitivity was satisfactory and generally on very short time horizons. Although in some cases larger time horizons were possible due to the strong correlations of features and failure data.

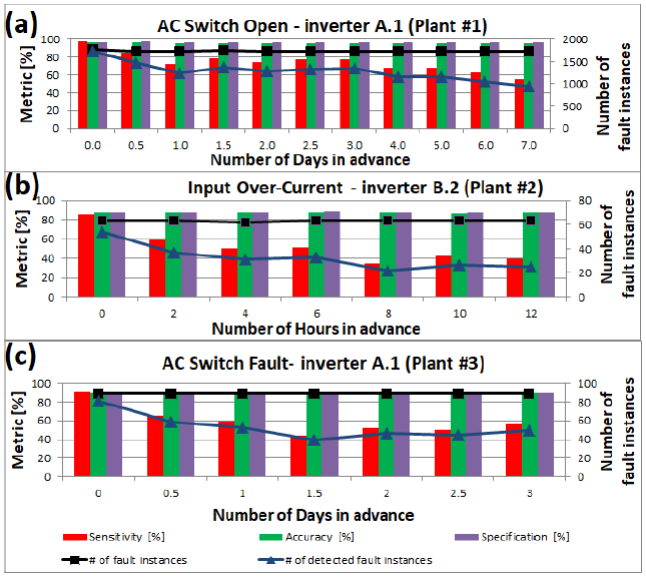

Figure 2.15: Statistics of fault prediction for a single inverter Betti and Lo Trovato [7]

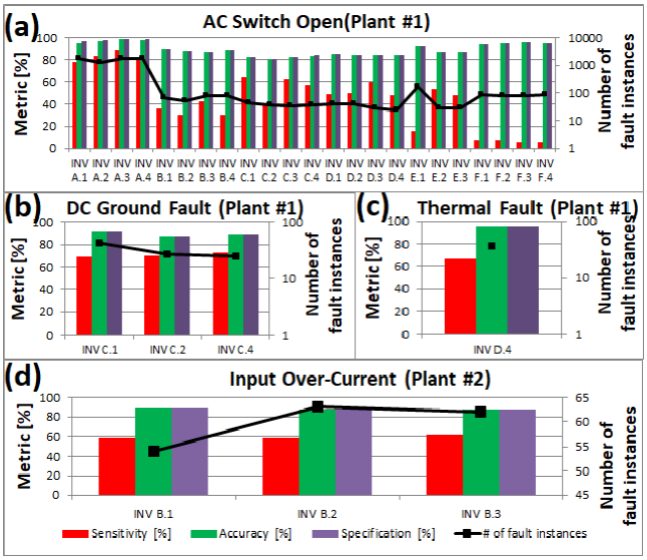

Figure 2.16: Statistics of fault prediction for all inverters in the plant Betti and Lo Trovato [7]

The results from Betti and Lo Trovato [7] indicate that their proposed method was effective in predicting generic faults up to 7 days in advance with sensitivity up to 95% . Whereas the specific fault classes with time predictions ranged from few hours up to 7 days. They claim their model is easily deployable for on-line monitoring of anomalies of new PV plants, requiring only the availability of historical SCADA, fault data, and inverter datasheet.

To summarise, in-spite of many components in PV parks the biggest contributor towards the failure are the inverters [22]. The inverter fails majorly due to power electronics and software issues. The current maintenance strategies of reactive and preventive maintenance cannot aid in early diagnosis of their faults. Therefore, predictive maintenance becomes an important aspect to improve the reliability of the PV plants. ML based methods for predictive maintenance of PV parks have immense potential. According to the literature reviewed very few researchers have been able to build working algorithms, and the major bottleneck for this is the availability of quality data sets from actual PV parks.

3

Machine learning

This chapter aims to explain the underlying principles of supervised machine learning algorithms which were utilised in this research. The section 3.1 describes the basic definition of ML. This is followed by an introduction to data pre-processing in section 3.2. Subsequently section 3.3 explains two major machine learning methods : regression and decision trees. The section 3.4 describes the methods to evaluate the ML models. Furthermore the section 3.5 describes the working principle of statistical tests which were used for the grouping of inverters into clusters.

3.1. Basics of Machine learning

Machine Learning (ML) can be defined as the ability of computers to learn from data rather than being explicitly programmed. The entire dataset used in the machine learning process is usually split into two groups; the training and the testing dataset. This is done to ensure the accuracy of the results and to increase the overall usability of the model. The subset of the data from which the algorithm learns is referred to as training data and the remaining fraction of the data which is used to evaluate the performance of the algorithm is known as testing data. The model is trained for a specified application. The performance evaluation of the trained model is done based on pre-defined metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and the models are reconfigured until an optimal minima is achieved. In general, machine learning approaches are classified into three major categories:

- Supervised learning : The input data and the output data are explicitly labelled. The ML models are trained using the training data, where each example is a pair of an input data and the corresponding output data. Therefore the derived function from a set of input examples can be used to make predictions on unseen data. This approach is named supervised learning, because the relationship between inputs and the resulting outputs is predefined.

- Unsupervised learning: In contrast to supervised learning this method has no labels to the data. The method tries to understand inherent patterns in the data in order to draw meaningful information.

- Reinforcement learning: It strives for an automated process for learning the best approach in any given context. It is achieved by performing a recursive action, where a positive outcome leads to a reward and a negative outcome is penalised. The model will change its parameters based on the feedback and retry the action with a new configuration until optimal performance is reached.

Building a supervised machine learning model involves several steps, including the gathering of the labelled data. For ease of understanding this report limits it to three major steps :

- Data Pre-processing

- Modelling

- Evaluation of the models

The steps of data modelling and tuning are recursive and are performed till we reached the maximum desired accuracy of the model. The accuracy of the model is calculated based on metrics such as MAE and RMSE which are discussed section 3.4. Each of the steps are mentioned in the list above are discussed in detail in the following sections.

3.2. Data pre-processing

The building of any ML model depends on the quality of the input data. Consequently, the first step before modelling is to prepare the data. The input data can suffer from numerous data quality issues depending on the mode of data acquisition. Additionally the outliers which are significantly different as compared to the majority values also need to be analysed. The outlier detection is specific to the problem statement. The input data for this research was acquired from sensors, this predicament is different as compared to an alternative source of data which is already processed. The complications related to sensor data are broadly classified as **system faults** and **data faults**[32]. The system faults could be a result of a calibration error, low battery, or an out of range situation. On the other hand, data faults are further classified into continuous and discontinuous [42]. The continuous faults constantly log inaccurate readings, and it is possible to observe a pattern in the form of a function. Whereas the discontinuous faults occur intermittently and their occurrence is discrete. The figures 3.2 and 3.1 represent the faults explained above.

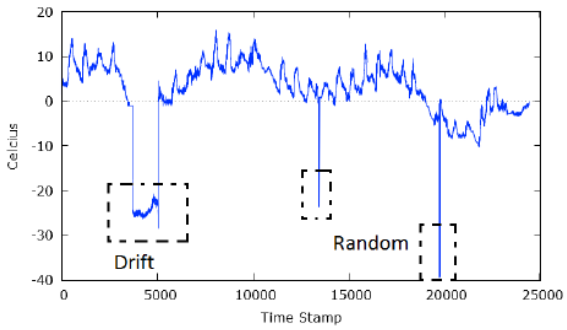

Figure 3.1: Drift fault [42]

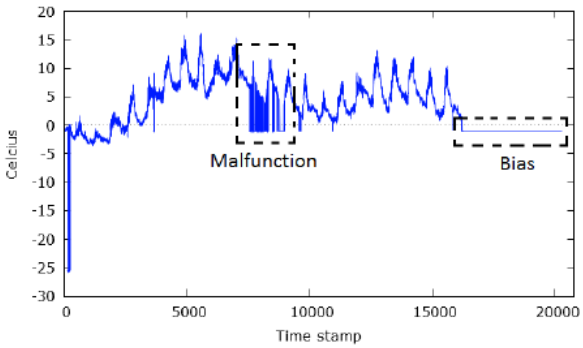

Figure 3.2: Malfunction fault [42]

3.3. Modelling

This section is divided into two subsections; 3.3.1 explains the principle and types of regression methods that were developed for this research. Whereas, subsection 3.3.2 introduces a more sophisticated type of ML algorithm known as Gradient Boosting Machine (GBM) and it also discusses its application using a quantile loss function.

To ensure models' reliability, they need to be tested on a hold-out data set which was not used during their training. Therefore, the first step of the modelling process is to split the data into training and testing sets. Training set is used for model fitting and based on the performance of the model on testing set, the model is either accepted or tuned further. The rule of thumb is to have 80% as train and 20% as test data, under the assumption that the training set is significant enough to generate a representative model.

3.3.1. Regression methods

The simplest regression method is known as linear regression. In principle, it models the relationship between variables by fitting a linear equation to observed data. Linear regression has been widely used to tackle many practical problems. A relationship between a scalar dependent variable y known as **target** and a scalar independent variable x known as **feature** is estimated in the linear regression modelling approach. This is illustrated in the fig 3.3 where the actual data is represented by the scatter plot, whereas the fit is represented by a line. A generic mathematical formulation of two variable linear regression is shown in equation 3.1. This approach can be extended to multiple independent variables, which is known as multiple or multivariate linear regression. The equation 3.2 represents a multivariate regression.

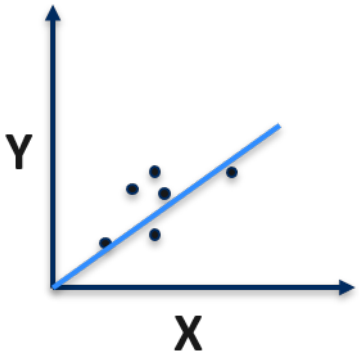

Figure 3.3: Simple linear regression

$$Y = mX + C \tag{3.1}$$

Y : Dependent variable
X : Independent variable
m : Slope
C : Intercept

$$Y_p = \alpha_0 + X_1\beta_1 + X_2\beta_2 .... + X_n\beta_n \tag{3.2}$$

y_p : Dependent variable (predicted value)
$X_1..X_n$: Independent variables
α_0 : Intercept
$\beta_1...\beta_n$: Coefficients of independent variables
The simplified vector representation of 3.2 is be given by equation 3.3:

$$Y_p = \Theta^T X \tag{3.3}$$

Θ : Vector consisting of all coefficients $\beta_1...\beta_n$ X : Matrix consisting of all independent variable vectors $X_1...X_n$
The coefficients are $\beta_1..\beta_n$ are obtained by minimising the loss function which is defined as Residual Sum of least Squares (RSS) and is represented in equation 3.4. The squared difference between every predicted value and the actual values is added to get the RSS. The process iteratively updates the coefficients θ based on the gradient descent method shown in equation 3.5. The general principle is that the chosen coefficients are the ones that lead to the lowest prediction error, depending on the loss function.

$$RSS = \sum_{i=1}^{n} (y_i - (y_i)^p)^2 \tag{3.4}$$

RSS : Residual sum of squares
y_i : Actual value
$(y_i)^p$: Predicted value

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \tag{3.5}$$

α : Learning rate
α : Cost function RSS

Although a simple linear regression may be computationally fast , it is prone to inaccuracies such as over-fitting and under-fitting. In order to overcome these inaccuracies methods such as elastic net can be employed. Before understanding the elastic net regression, it is important to introduce a few relevant terminologies.

Over-fitting and Under-fitting

Overfitting in a model can occur if we minimise the error in such a way that it becomes tailored only for the specific training dataset. As a result when the model is applied to the testing dataset, it performs poorly. This means that the model cannot be generalized. The figure 3.4 shows that over fitting happens when the fit passes through all possible training data points, in which case the model may not perform well on a previously unseen data. This phenomenon of deviation is known as **variance**.

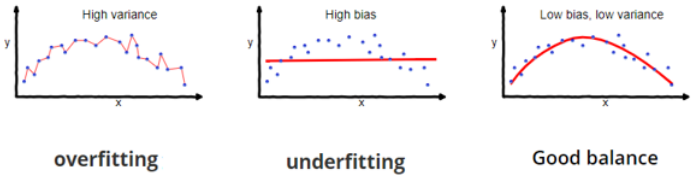

Figure 3.4: Overfit and underfit

A fit is classified as under-fit if it is not representative of the data. Under-fit can occur if the relationship between the variables is poorly represented by the model or the model was tuned to achieve a good fit. An under-fit suffers from a high bias . In order to diminish this behaviour a good fit requires a balance of bias and variance,otherwise it may not perform well on unseen data. **Regularisation** is a technique which can help achieve this balance of bias and variance by introducing a bias to the loss function mentioned in equation 3.4. Elastic Net is a commonly used regularisation algorithm.

Elastic Net

In order to explain the elastic net algorithm, its components LASSO and Ridge regression need to be explained.

- **Ridge regression**:In order to perform better on unseen data, the ridge algorithm introduces a bias which is mathematically known as the L2 norm. The equation 3.6 shows that the loss function is modified by a bias, this bias basically desensitizes the algorithm by penalising it. When the value of λ which is known as the hyper-parameter is zero, there is no difference between a normal regression and ridge. However as λ approaches the loss function becomes less sensitive to the input data (independent variable). The hyper parameter λ needs to be tuned accordingly to minimise the loss function.

$$RidgeLoss = RSS(3.4) + \lambda \sum_{j=1}^{p} {\beta_j}^2 \tag{3.6}$$

- **Least Absolute Shrinkage and Selection Operator (LASSO)**: Although ridge regression helps in penalising the features to a certain extent, it can never make the dependency of any independent variable to zero. This is because of the squared L2 norm($(\beta_j)^2$). To solve this inherent issue a LASSO can be utilised. Unlike the Ridge the LASSO penalises the $|\beta_j|$ as shown in equation 3.7.

$$LASSOLoss = RSS(3.4) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.7}$$

This advantage of LASSO can help in reducing features which do not aid in the prediction and hence improving the quality of the algorithm.Similar to the ridge, the hyper-parameter λ needs to be tuned. The elastic net combines the power of both ridge and LASSO as shown in equation 3.8. The hyper parameters λ_1, λ_2 decide the quality of the regularised regression.

$$EnetLoss = RSS(3.4) + \lambda_1 \sum_{j=1}^{p} {\beta_j}^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \tag{3.8}$$

The hyper-parameter tuning is done through the method of cross validation, which is shown in figure 3.5. During every fold, a predetermined section of the data is left out and the performance of the algorithm is

evaluated. In addition to this for hyper-parameter tuning we need to specify the range of values for L1 and L2 norms. The way that this is carried out is by specifying the ratio of L1:L2. Further information regarding hyperparameter tuning can be found in the Appendix A.

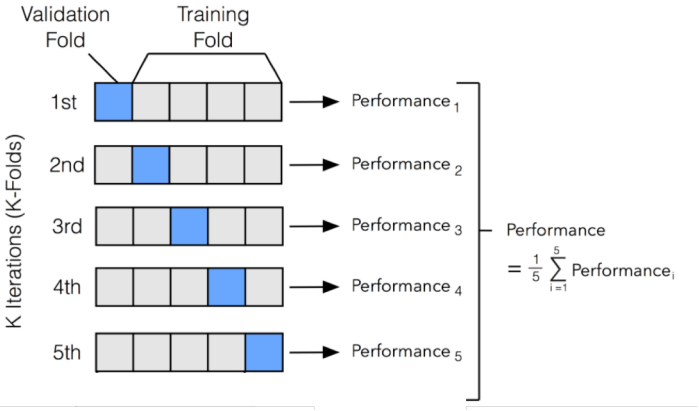

Figure 3.5: k-fold cross validation [17]

3.3.2. Gradient Boosting

Gradient boosting falls under the family of decision tree models in machine learning. It combines the power of many weak models in order to form a strong model known as an ensemble or a decision tree.
The basic structure and terminology is shown in figure 3.6 .
Root node: This is the base of the tree , from which all other branches arise through splitting of the data.
Internal node : There are arrows pointing to and away from these nodes, usually also known as the ineterim nodes.
Leaves : These nodes are the final decision points and generally vary from 8-32 depending on the precision required. The detailed explanation can be found in [47] [19].

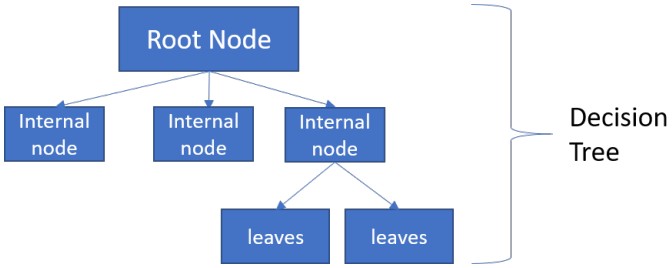

Figure 3.6: Decision trees

The loss function is also a squared error just like in linear regression. This loss function is chosen because the initialisation of values is the mean of the observed values. It is due to the fact that the differential of the loss function with respect to predicted value yields mean.

$$loss = \sum (y_i - (y_i)^p)^2 \tag{3.9}$$

$$(y_i)^p = (y_i)^p + \alpha * \delta \sum \frac{\partial (y_i - (y_i)^p)^2}{\partial (y_i)^p} \tag{3.10}$$

y_i : Actual values $(y_i)^p$: Predicted values There are additional loss functions which are suitable to features. The main feature in this research is irradiance , and it's temporal variation throughout the day is very high. Hence if squared error loss can lead to structural error in the predictions. Therefore quantile loss function is an alternative to the squared loss function. Instead of having a single value of prediction, a range of predictions can be created depending on the number of quantiles that are chosen.

$$quantileloss = \sum_{i=y_i<y_i^p} (\gamma - 1) \mid y_i - y_i^p \mid + \sum_{i=y_i>y_i^p} (\gamma) \mid y_i - y_i^p \mid \tag{3.11}$$

The advantage of quantile regression as previously mentioned is that the data can be split into range of percentiles. The figure 3.7 illustrates the implementation of quantile regression to a simple sinusoidal function. It shows that unlike squared loss function where we could only have a single prediction, quantile regression can accommodate a range of predictions. The accuracy of the quantile regression is dependant on hyperparameter tuning and the methodology is explained in Appendix.

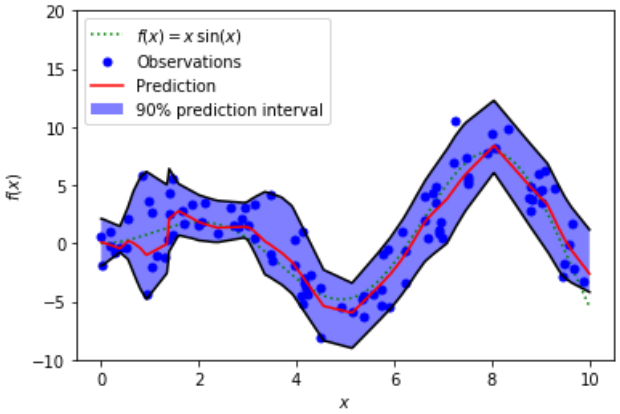

Figure 3.7: Quantile regression

3.4. Evaluation of models

The evaluation of models based of metrics is an important process of modelling. It determines the relative performance of different models which helps in choosing an optimal model. In addition to choosing the model, metrics also help in choosing the best hyper-parameters for the model. Therefore it is op prime importance to decide the most relevant metric for the problem statement. This section focuses.

3.4.1. Coefficient of determination

The coefficient of determination is generally represented as R^2 and can be calculated by the equation 3.12 . In regression it a commonly used metric. The higher the value better is the quality of the fit. Although R^2 is primarily important during the initial stages of model building, it does not contribute at a later stage. This is due to the fact that the initial models tend to have very low values whereas as the model quality improves the marginal change in R^2 is very low. The drawback of this metric is that it cannot capture *overfitting*.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - (y_i)^p)^2}{\sum_{i=1}^{n}(y_i - (\overline{y})^2} \tag{3.12}$$

3.4.2. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is calculated using equation . It is defined as the mean of the absolute difference between observed and predicted values. The MAE does not penalise higher errors differently unlike the mean square error, which scales the higher errors more exponentially. Hence when we want to avoid higher values of error MSE is a better alternative.

$$MAE = \sum_{i=1}^{n} \frac{\mid y_i - {y_i}^p \mid}{n} \tag{3.13}$$

3.4.3. Mean Square Error (MSE)

The Mean Square Error (MSE) aims to calculate the squared difference between observed and predicted values, the equation

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - (y_i)^p)^2 \tag{3.14}$$

3.4.4. Mean Absolute Percentage Error (MAPE)

The MAPE is useful because it is independent of the size of the system. It must be used alongside with other metrics as standalone MAPE can be very large indicating low performance, which may not be the case.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \mid \frac{y_i - (y_i)^p}{y_i} \mid \tag{3.15}$$

3.5. Statistical tests

Statistical methods are very important in determination of similarity between datasets. While comparing multiple data-sets it is important to keep in mind that it is not possible to prove that two data sets are alike, only a likelihood can be established [16]. The methods such as t-test , ANNOVA are widely accepted to prove the likelihood of the data. t- Test has been implemented in this research and hence discussed elaborately. Further information regarding t- Test and ANNOVA can be found in [47] [5] .

3.5.1. t-Test

There are multiple types of T-tests such as One-sample, independent two sample, Paired test. Since paired t-Test is relevant to the research that is discussed.

Independent two sample T-test:
In a paired t test the used we compare the means between two related groups. To perform any statistical there are three major requirements:

- Data availability
- Defining the null hypothesis H_0
- In case the data has more than 2 groups, an alternative hypothesis H_1 needs to be defined.

The null hypothesis can be rejected or failed to be rejected depending on the threshold of the P-value proposed. Consider two samples S_1 and S_2 , with sizes n_1 and n_2 , and μ_1 and μ_2 as their respective means . The t -statistic is calculated as

$$t = \frac{\mu_1 - \mu_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.16}$$

$$s_p = \sqrt{\frac{(n_1 - 1)(\sigma_1)^2 + (n_2 - 1)(\sigma_2)^2}{n_1 + n_2 - 2}} \tag{3.17}$$

s_p : Pooled standard deviation ($n_1 \neq n_2$)
σ_1 : Standard deviation of S_1
σ_2: Standard deviation of S_2

The t statistic that is obtained from 3.16 would correspond to a p- value , which is obtained from t-tables . P- value is further used to check if the null hypothesis can be rejected or fail to be rejected.
An example of t - Test is shown below.
H_0 : The samples S_1, S_2 are unequally distributed.
$P < 0.05$, specifies if the difference between the samples is less than 5% then we reject the null hypothesis else we fail to reject it.

4

Baseline data

4.1. Moerdijk

The data for this research is acquired from a solar plant located at Moerdijk which is a part of Shell's new energy assets.This plant was commissioned on 1st February 2019 and has been functional since then. The rated capacity of the plant is 26.6 MWp and is spread over an area of 0.3 km^2. The plot is divided in fourteen geographical sections as shown in figure4.1

Figure 4.1: Moerdijk Solar Park outline

The PV system has an azimuth of 176^o and fixed tilt of 20^o. The technical details of the relevant instruments and equipment's are explained in the following subsections.

4.1.1. Electrical architecture

The electrical architecture is comprises of the interconnections between the power producing PV panels and the grid, through the inverters and substations. The geographical plot as shown in figure 4.1 has 10 substations numbered from 100TSR to 109TSR. Each substation has 34 inverters connected to it. Every inverter has 8 strings (2 per MPPT) . This architecture is illustrated in 4.2

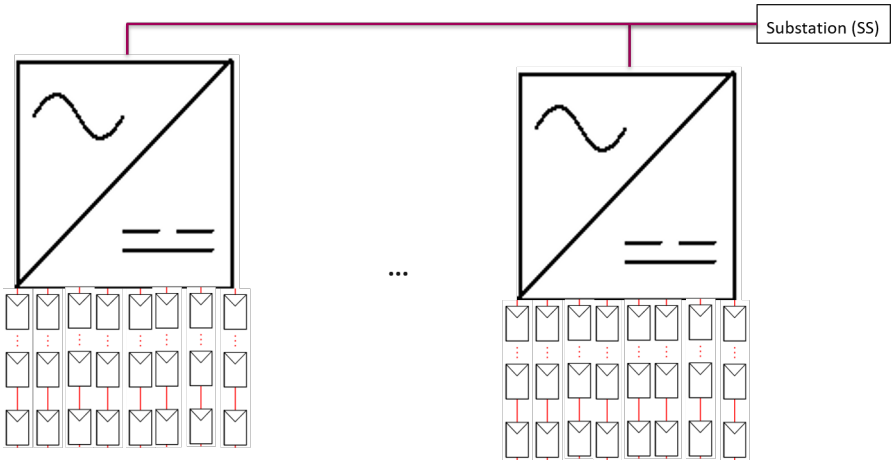

Figure 4.2: Schematic of electrical architecture

The inverter can be located based on the tag name.

4.1.2. PV panel

There are approximately 76,000 high efficiency mono crystalline half cell panels.The major technical specifications of the panels are specified [45]. The PV panels shown in 4.3

Figure 4.3: PV panels

- Original Equipment Manufacturer (OEM) : Suntech
- Model : STP 340/345/350 S
- Power : 340 / 345 / 350 Wp
- Voc : 42.6 / 42.7 / 42.9 V

- Isc : 7.65 / 7.72 / 7.81 A

4.1.3. Inverter

To convert the DC power produced by the solar panels, there are 340 grid tied inverters which further deliver the power to the substations. The electrical architecture is explained in the 4.1.1.

- OEM : Huawei
- Model: SUN 2000- TKL series
- Power: 60 kVA
- Voltage: 1500 V

The architecture of the inverter is shown in 4.4, which depicts that each inverter has 4 MPPT 's with a tranformerless configuration.

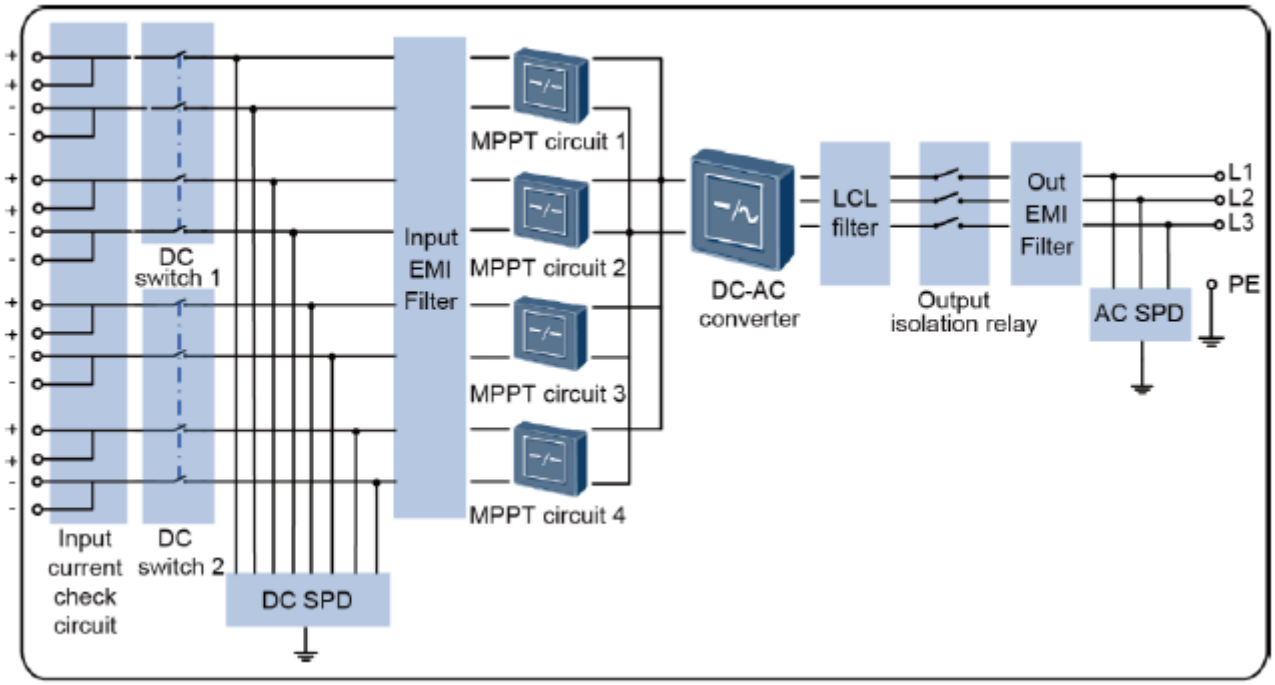

Figure 4.4: Huawei inverter architecture

4.1.4. Measuring Instruments

The list of instruments and their positions in the plant are marked in figure 4.12 .

- Pyranometer : 6
- Module temperature : 5
- Meteo station : 1

Pyranometer

Pyranometer is a device that measures the incident irradiance on a plane. There are multiple types of pyranometer, the one used in this project is of thermopile type. If the pyranometer is mounted on a flat plane it measures Global Horizontal Irradiance (GHI), but if it is inclined at the tilt of the modules it measures GPOA. This plant has five GPOA and one GHI sensor.The typical installation of pyranometer is visualised in 4.5.The other technical details pertaining can be found in the data-sheet attached in the appendix.

- Make : Hukseflux

- Model: SR D-20
- Class: Second class

Figure 4.5: Pyranometer mount

Module temperature sensor

Module temperature is an essential property to determine the deviation of expected to the actual power. The module temperature in the plant is measured using mounted pt100 resistive element. As shown in figure 4.6 it has 4 wires which sense the change in resistance of the pt100 element and hence measures the temperature.

- Make: Lambrecht GmbH
- Model:
- Operating range : -50 to 150 deg$C$

Figure 4.6: Pt100 module temperature sensor

4.1.5. SCADA system

A typical window of the SCADA is shown in 4.7 . The download option is available in the export data tab, which has a drop down to select the device and the substation to be selected. It is shown in figures 4.8 and 4.9. The data resolution is 5 minutes, and it can be downloaded as .csv files for further analysis.

Figure 4.7: SCADA

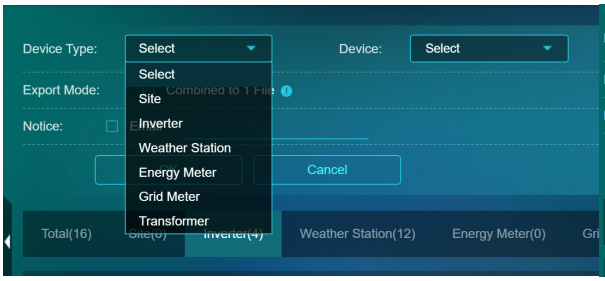

Figure 4.8: Device drop-down

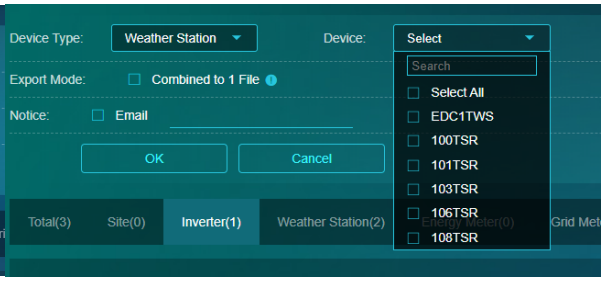

Figure 4.9: Substation drop-down

In addition to the process data, there is also a possibility to download the fault occurrences as a .csv file which contains, the type of fault , start time, end time. This is illustrated in figure 4.10.

	Farm	Device	Device Type	Severity	Category	Description	Start Time	Duration	End Time
2	Moerdijk Solar	106EIN-0002-0003	Inverter	Fault	DC ground fault	Low Insula	27-12-19 9:14	0 00:20:36	27-12-19 9:35
3	Moerdijk Solar	106EIN-0002-0003	Inverter	Fault	DC ground fault	Low Insula	26-12-19 9:03	0 00:20:18	26-12-19 9:23
4	Moerdijk Solar	106EIN-0002-0003	Inverter	Fault	DC ground fault	Low Insula	24-12-19 9:29	0 00:20:32	24-12-19 9:49
5	Moerdijk Solar	106EIN-0003-0002	Inverter	Fault	DC ground fault	Low Insula	22-12-19 9:21	0 00:20:13	22-12-19 9:41
6	Moerdijk Solar	106EIN-0002-0003	Inverter	Fault	DC ground fault	Low Insula	22-12-19 9:20	0 00:45:33	22-12-19 10:05
7	Moerdijk Solar	106EIN-0002-0003	Inverter	Fault	DC ground fault	Low Insula	17-12-19 9:08	0 00:20:26	17-12-19 9:29
8	Moerdijk Solar	106EIN-0002-0003	Inverter	Fault	DC ground fault	Low Insula	12-12-19 8:56	0 00:20:31	12-12-19 9:17

Figure 4.10: Fault log snip

The faults are divided into 7 sub categories, and the fault count per sub category is visualised in figure 4.11. Each of these faults can be mapped to a fault code in the operating manual provided by the OEM. Given that the faults can occur either due to internal circuit problem, or external disturbances from grid (over-voltage, frequency imbalance). The table 4.1 clearly enlists the cause for the fault and the corresponding alarm ID in the OEM manual. The focus of this research project is to predict the occurrence of equipment failure as this sub-category corresponds to failures because of internal reason unlike the others. The terminology of failure can be misleading, as it suggests that the component is replaced after every failure. Although this subcategory is tagged to a scenario of zero power and not necessarily a complete breakdown of the inverter.

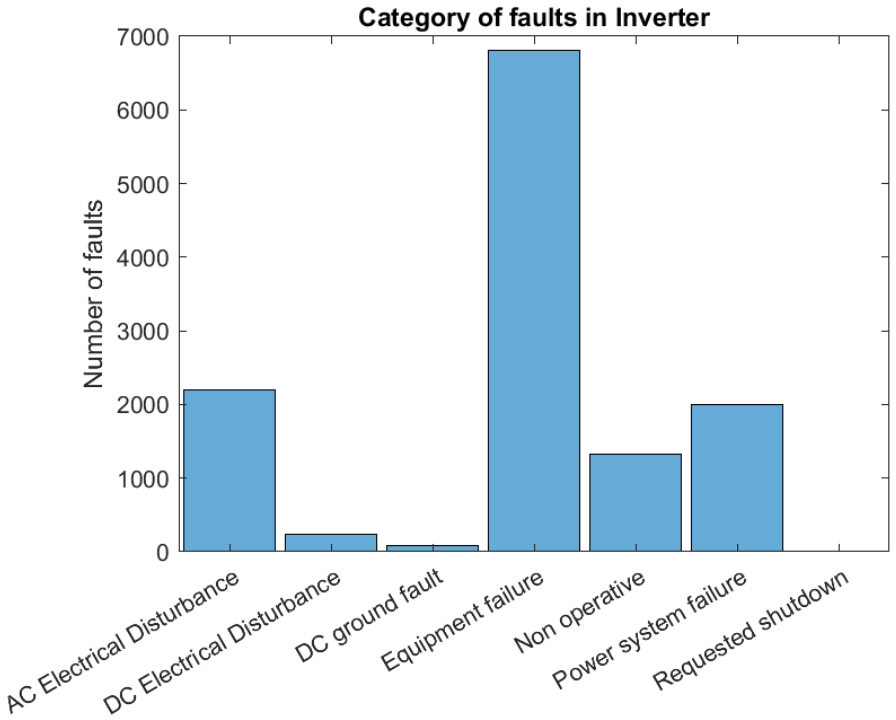

Figure 4.11: Count of faults in 2019

Type of failure	Corresponding Alarm ID	Cause
AC Side failure	Alarm id-301 cause id 29	External
DC Electrical disturbance	Alarm id-413 caseid-3,4,alarm id-127 case id-2	External
DC Ground fault	Alarm id-313 cause id 1	External
Equipment failure	Alarm id-400 cause id 23, shutdown fault	Internal
Non operative	DC switch off, Power limit	External
Requested shutdown		
Power system failure	Alarm id-200,202,318	External

Table 4.1: Fault taxonomy

Actual tag	Modified tag
100 TSR / SS100	S1
101 TSR/ SS101	S2
103 TSR/ SS103	S3
106 TSR/ SS106	S4
108 TSR/ SS108	S5

Table 4.2: Modified naming convention for substations

4.1.6. Summary of Equipment's

In summary ,there are 10 substations where the power from 340 inverters is aggregated , out of the 10 substations only 5 have monitoring stations which measure GPOA irradiance and module temperature. As the naming convention for the substation is considerably long, a simplistic convention was followed according to the table. Furthermore the figure 4.12 visualises the schematic of the plant with associated locations of the sensors and the substations.

Figure 4.12: Moerdijk schematic

5

Economic analysis

This chapter presents the economic feasibility of predictive maintenance for the Moerdijk plant. Based on the current financial data and assumptions, it was found that by implementing predictive maintenance at Moerdijk a potential saving 10% of Operational Expenses (OPeX) can be possible. This chapter is divided into three sections : section 5.1 discusses the scope of current operations and the costs related to it, section 5.2 establishes the economic feasibility of predictive maintenance based on the data of Moerdijk . Section 5.3 provides a hypothetical model for dispatch of personnel based on predictive alerts generated by an algorithm.

5.1. Current operational expenses

The OPeX for Moerdijk is dominated by three main items : Land lease, maintenance contractor, asset management. The table 5.1 enlists the line items including with their current annual costs. The figure 5.1 shows that the O&M service provider fee is the highest amongst the OPeX. The land lease expenses cannot be mitigated as they are decided according to the local land prices. Asset management includes the cost of a Shell representative handling the contracts and negotiations for the plant. All the other miscellaneous need to be incurred due to the nature of the activities. For example the SCADA system needs to be in place if there was no production of energy.

	Current costs (kEuros/year)
O&M service provider	250
Land lease	166
Asset management	64
SCADA	34
Utilities	21
Property tax	71
Total	**631.6**

Table 5.1: Current OPeX costs

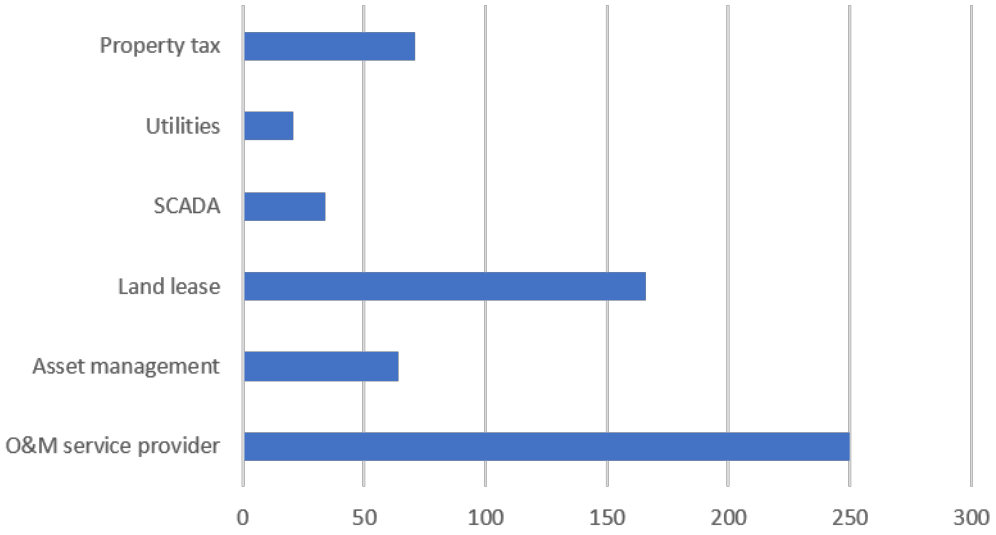

Figure 5.1: Current annual OpeX cost comparison

5.2. Economic feasibility of predictive maintenance

The current O&M service provider carries out scheduled & reactive maintenance activities for the plant and therefore their activities were analysed to identify cost optimization potential using predictive maintenance. The table 5.2 below shows a break down of the O&M service providers activities and their corresponding costs. It was found that 50% of the costs related to their fee are fixed as they include activities such as vegetation management,meter maintenance and security. The remaining 50% of the costs are linked to the bi-yearly scheduled maintenance & reactive maintenance. As a result, the scope of introducing predictive maintenance lies in this remaining 50%. To obtain a better understanding of what is included in the contract was studied. Currently the scheduled maintenance occurs every 6 months, for which the O&M service provider arranges for inspection and changes any damaged components. Therefore under performing assets are not detected before the periodicity, which can lead to a significant reduction of production and in turn the revenue of the plant. Although if the equipment fails. The current clause for reactive maintenance response times linked to reactive maintenance were studied. The clause in the contract from the OM service provider can be seen in figure 5.2. It shows that the contractor has upto 10 days in case of a fault which leads to a loss of 1% of the capacity . This leads to a significant loss in the production. The loss of production leads to a revenue loss of 107 eurosMWh in addition to the fee that is being paid to the OM service provider. Before establishing the superiority of predictive maintenance over the current techniques, an assumption needs to be stated:

- The predictive maintenance algorithm can detect the faults with a reasonable lead time and accuracy.

The pie chart shown in figure represents the cost distributions, 25% of the total OPex is spent on the contract with OandM service provider. The variable costs account to 12.5% of OPeX. Therefore in the case of predictive maintenance instead of service provider contacting the subcontractors to avail technical help, it could be done in house with the help of predictive alerts. This will not only save the service provider fee, but reduce the notional losses due to production as the predictive alerts would have helped in avoiding the downtime altogether. The additional costs associated with the building and maintenance of the predictive maintenance algorithms were calculated to be around 2.5% of the total OPeX. **Therefore a potential saving of upto 10% can be achieved in Moerdijk if predictive maintenance is implemented**.

O&M service provider-activities in the scope	**Cost kEuros/year**
Gate Maintenance	125
Grounds Maintenance	
O&M Building Upkeep	
Parts Inventory	
Snow Clearing	
Waste Removal	
Janitorial work	
Communications (phone, internet)	
Met Tower Maintenance	
Lighting Alerting	
ISNetworld HSSE Management	
Vegetation Management	
Meter maintenance	
Road Maintenance and repairs	
Schedule & corrective Maintainence	125

Table 5.2: Cost break down of O&M service provider contract

ANNEX C – CORRECTIVE MAINTENANCE RESPONSE TIMES

CONTRACTOR will ensure trained maintenance personnel respond within the following time frames following its receipt of a NOTICE from OWNER:

Percent of FACILITY Capacity impacted	**Response time for Claim Acknowledgement upon receipt of OWNER's NOTICE**	**Response time for initiating investigation Upon Acknowledgement by CONTRACTOR**
100%	8 hours	5 hours
>25%	8 hours	12 hours
>10%	24 hours	24 hours
>5%	24 hours	24 hours
>3%	24 hours	6 days
>1%	24 hours	8 days
<1%	5 days	10 days

Figure 5.2: Snip from O&M service provider contract

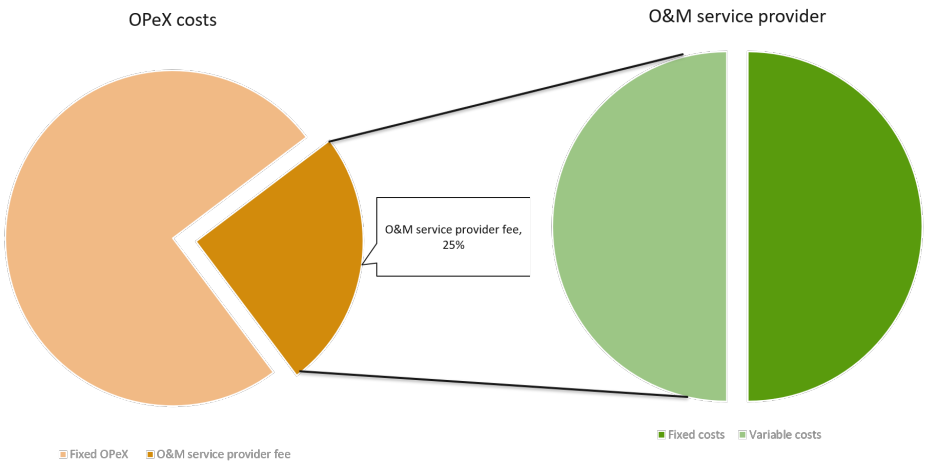

Figure 5.3: OpeX Cost distribution

5.3. Dispatch of personnel based on predictive alerts

In addition to the assumptions stated previously, it is also assumed that the algorithm can predict the extent of damage, that is further translated to duration of downtime. The list of other assumptions is given below.

- Based on the contract ,faults were divided into three classes : Class A, Class B, Class C
- The notional loss because of loss of production was valued at 107.5 euros.
- The cost of Electrician(E), Technician(T), Labourer(L) as taken as per the table 5.4.

Table 5.3: Add caption

Cost of personnel	Hourly cost (euros)
E+T+L	154
E+T	100
E	55

Table 5.4: Cost of deploying personnel

- Then fixing duration was divided into three classes : 8 ,16 24 hours.
- The travel costs associated with movement of personnel have not been considered.

Based on the assumption a simple model was developed to decide how many duration and the number of personnel to be dispatched. The severity of fault decided the minimum number of personnel, the table shows that the most sever class of faults : class A, three people have to be sent for the repair unlike class C, where 1 person might be sufficient. Following this a MWh matrix was formed on the basis of the equation 5.1. The two matrices are then multiplied to generate the energy loss. This resulted in the minimum amount of loss of energy to be predicted by the algorithm, in order for it to be feasible to dispatch the personnel.

	MWh matrix		
Full breakdown dispatch	3 people (euros)	2 people	1 person
8	11.43387	7.424594	4.083527
16	22.86775	14.84919	8.167053
24	34.30162	12.25058	12.25058
	Class A	Class B	Class C

Table 5.5: Dispatch Activation Matrix

	Class A	Class B	Class C
E+T+L	1	1	1
E+T	0	1	1
E	0	0	1

Table 5.6: MWh loss matrix

$$MWh_{loss} = \frac{deputationcost(euro)}{notionalloss(euro/Mwh)} \tag{5.1}$$

6
Results

This chapter presents the two methods developed for early fault detection of inverter faults in order to generate predictive maintenance alerts. The first method was developed using an Elastic Net, based on which monthly limits were defined for early fault detection. Whereas the second method used a GBM approach with quantile loss function based on which continuously varying limits were defined for early fault detection. Since the developed algorithms did not detect any early signal of fault occurrence, further analysis into the fault data was done. It presented the data quality issues in the SCADA fault tags which resulted in the discovery that the mapping of the fault was erroneous. Only seven cases of actual invereter fault were identified and they were insufficient for developing a representative algorithm. The chapter is divided into four sections : data pre-processing, baseline model, advanced modelling results and fault detection, advanced fault analysis. Section 6.1 explains the procedure that was followed to clean and process the sensor and inverter data to create the final dataset used in modeling. Prior to deciding the final models, a baseline model was developed and the results are discussed in section 6.2. It is followed by section 6.3 where advanced modelling based on the final dataset is explained. Additionally, it also presents the method of early fault detection in inverters. The algorithms were developed based on the data from the Moerdijk plant and the details have been described in chapter 4.

6.1. Data Pre-processing

The quality of any algorithm primarily depends on the quality of the input data and therefore it was prepared before building the model. Data pre-processing was performed on the features as well as the target variables. The sampling rate for features and target is 5 mintues. The *features* that were considered are GPOA and module temperature, while parameters like wind speed and local humidity could not be considered because of unavailability of the data. The *target* variable can be either voltage, current or active power of the inverter. Active power was selected because it includes the inverter performance as it is measured after the DC-AC conversion. Additionally it was chosen as the target variable based on the recommendation from literature [11].

6.1.1. Feature pre-processing

The two features GPOA and module temperature were recorded by five different sensors on the site. The locations and characteristics of the sensors are mentioned in chapter 4. The sensor data has quality issues such as unavailability, faulty values and outliers which are significantly different from normal operating values. The concerns regarding the quality of the sensor data have already been addressed in **??**.
The figures 6.1 and 6.2 show the raw data from the sensors. The illustration shows that there are instances of very low values (Module Temperature $< -5^{o}C$) of module temperature which were possibly caused due to a malfunction of the sensor. Similarly, the negative values of GPOA($< 0\ W/m^2$) were due to data acquisition+ at night when GPOA is 0 Wm^2. Additionally, the discontinuities were possibly caused because of a sensor malfunction. Moreover, as the data is captured through out the day, the night values need to be dropped as they correspond to a zero power scenario. The sensors are connected to a data acquisition system and can suffer from either a network outage or a sensor failure; both of which can lead to unavailable values. In addition to unavailable values, sensor malfunction can lead to local spikes or troughs that need to be filtered. The list

below shows how the described data quality issues and outliers were handled.

- Night values : Since the plant is based in the Netherlands , the sun hours expected are very low. According to the meteorological data from [1] the longest day was recorded to have sunlight from 5:00 AM-10:30 PM. Hence, all the values between 10:30 PM and 5 AM are dropped. During the winter months the values in the evening had very low irradiance, so these data points were dropped.

- Missing values : The data points which correspond to not available values were dropped from the dataset because no values were available during those timestamps.

- Data Spikes : Spikes can occur either due to a local weather fluctuation or due to a sensor failure. A simple moving average filter with a window of 15 minutes (3 intervals, as 1 interval was 5 minutes) was applied to detect unusual spikes. On an average 3 % of the GPOA values recorded were classified under spikes. However, not all of the values were discarded, due to the probability of momentary weather fluctuations. These values were examined with their corresponding power values to identify if they were actual spikes.

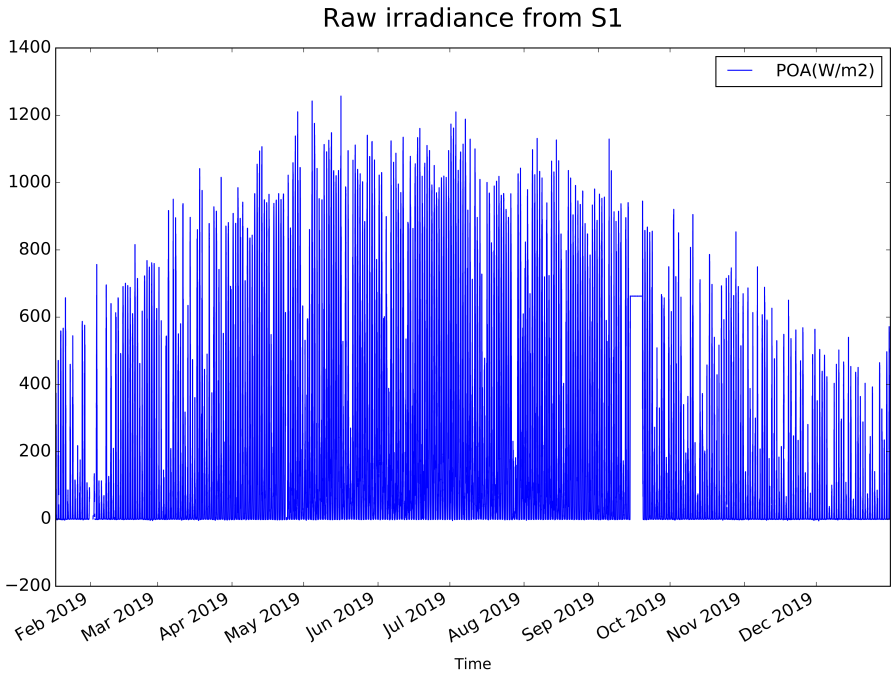

Figure 6.1: Raw GPOA sensor data from a sensor located at 100TSR(s1)

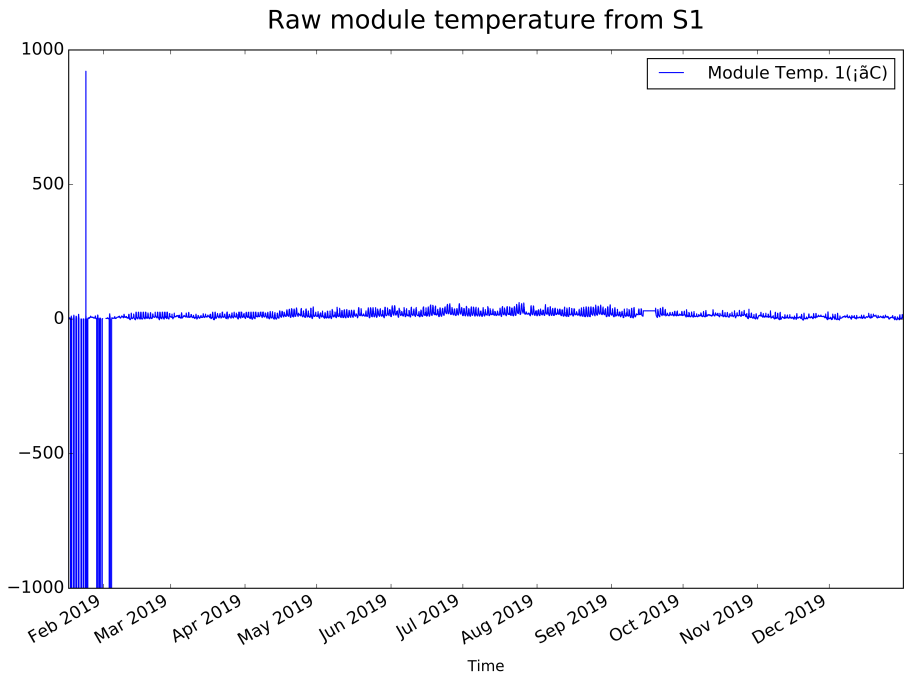

Figure 6.2: Raw module temperature sensor data from a sensor located at 100TSR(s1)

	S1		**S2**		**S3**		**S4**		**S5**	
	GPOA	Module temp	GPOA	Module temp	GPOA	Module temp	GPOA	Module temp	GPOA	Module temp
Night(%)	28	28	27	27	28	28	28	28	28	28
Gaps(%)	0.5	0.8	0.5	0.9	0.5	0.8	0.5	0.9	0.5	0.8
Peaks (%)	3	8	5	10	4	8	5	6	5	8

Table 6.1: % of outliers

The table 6.1 shows the % distribution of the outliers from all the sensors. Subsequently the data that was further available for modelling was about 70 %.

After the processing specified in the list above, the features were plotted and are shown in figure 6.3 and 6.4. The plots indicate that the values between the sensors are in similar range, which means that the most extreme outliers were dropped.

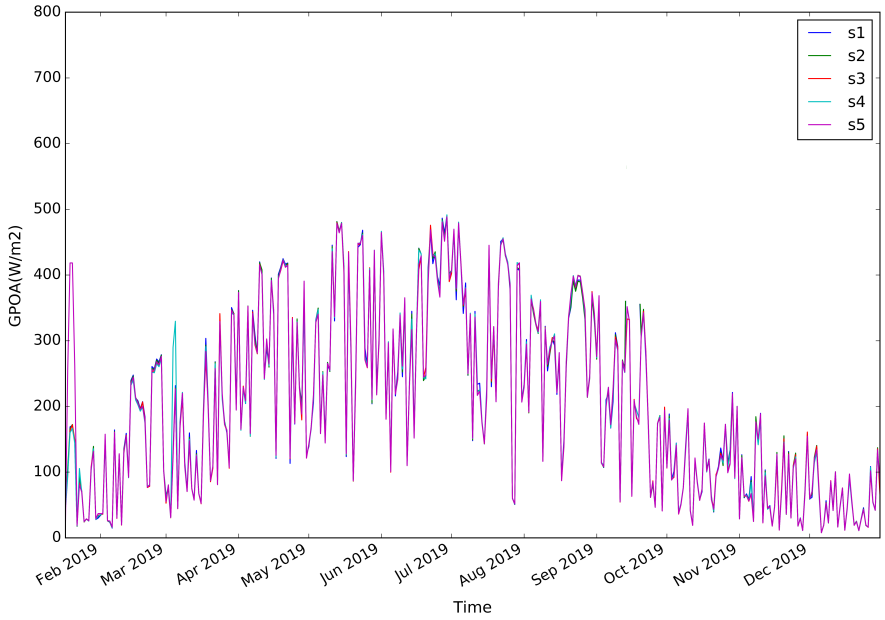

Figure 6.3: Processed GPOA sensor data from all 5 GPOA sensors

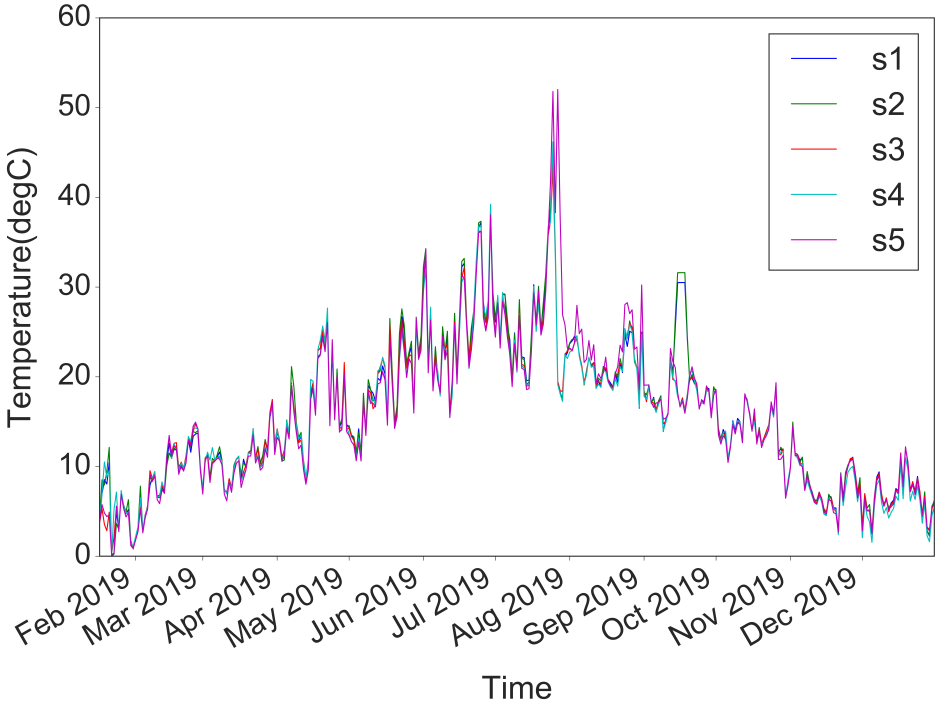

Figure 6.4: Processed module temperature sensor data from all module temperature sensors

6.1.2. Target pre-processing

Fault analysis
The step prior to processing the target variable is to analyse the type of faults that are recorded in the SCADA system. The figure 6.5 shows the major reasons for inverter downtime in the plant: grid disturbance, cable insulation, internal circuit. The figure 6.6 shows the types of the faults with their count that were recorded in

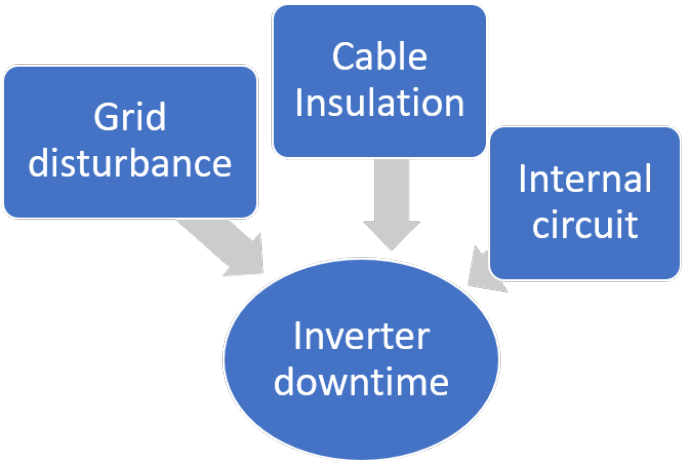

Figure 6.5: Major contributors for inverter downtime

the year 2019 at the Moerdijk plant. The faults are classified into two types: external and internal. Internal faults refer to the failure of internal circuit and associated components; whereas external refers to faults which were triggered by an external source such as low/ high frequency of the grid. The table 6.2 shows the reason for the fault as per the fault mapping.

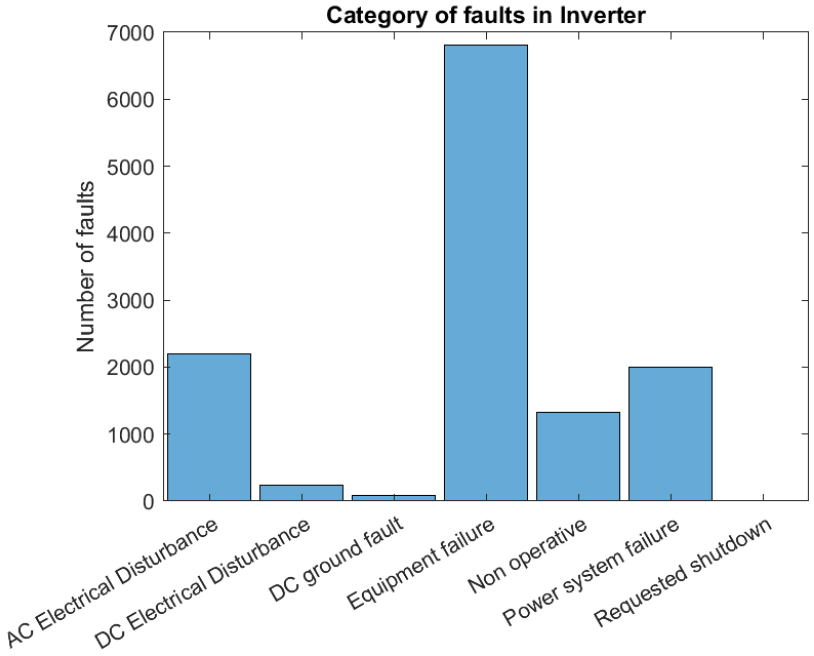

Figure 6.6: Inverter fault taxonomy from Moerdijk
with fault count in 2019

Since the internal failure is represented by *equipment failure*, the occurrence of *equipment failure* per inverter throughout the year 2019 was plotted and is shown in figure 6.7. The plot clearly shows that three inverters were responsible for a high count of *equipment failure* as shown in figure 4.11. The tags of the three inverters were *103EIN-0002-0001, 107EIN-0005-0004, 105EIN-0007-0004* and the reason for such a high count was due to a manufacturing defect which led to their replacement. These inverters were dropped

Fault name	Cause of failure	
AC Side failure	This failure was recorded when the AC circuit breaker opened due to either a frequency imbalance in the grid.	External
DC Electrical disturbance	Any disturbance at the PV side such as low output voltage triggered this failure.	External
DC Ground fault	Due to constant exposure to humidity , the insulation resistance with respect to the ground deceases which causes an excessively high residual current causing a ground fault.	External
Equipment failure	This failure was registered when there was a problem with the internal circuit / the associated components which are responsible for the DC-AC conversion.	Internal
Requested shutdown	During a planned shutdown , it was registered in the SCADA for safety reasons.The components were in a lock out condition, till the requested shutdown was complete.	External
Power system failure	Whenever the AC side had imbalances in the three phases of voltage this failure was registered.	External

Table 6.2: Explanation of fault taxonomy of SCADA fault tags used in the Moerdijk plant

before modelling because it cannot model the normal operation of an inverter due to the high failure rate. Additionally it may lead to establishing of false positive patterns for early detection of faults, which is not desirable. Therefore, these inverters were dropped from the dataset, which reduced the count of equipment failure by 73%. The modified *equipment failure* count per inverter was re-plotted and is shown in figure 6.8. This plot shows an average failure rate of 10 per inverter, which is reasonably acceptable to model the normal operating condition. Finally 337 inverters out of 340 were considered for further modelling and analysis.

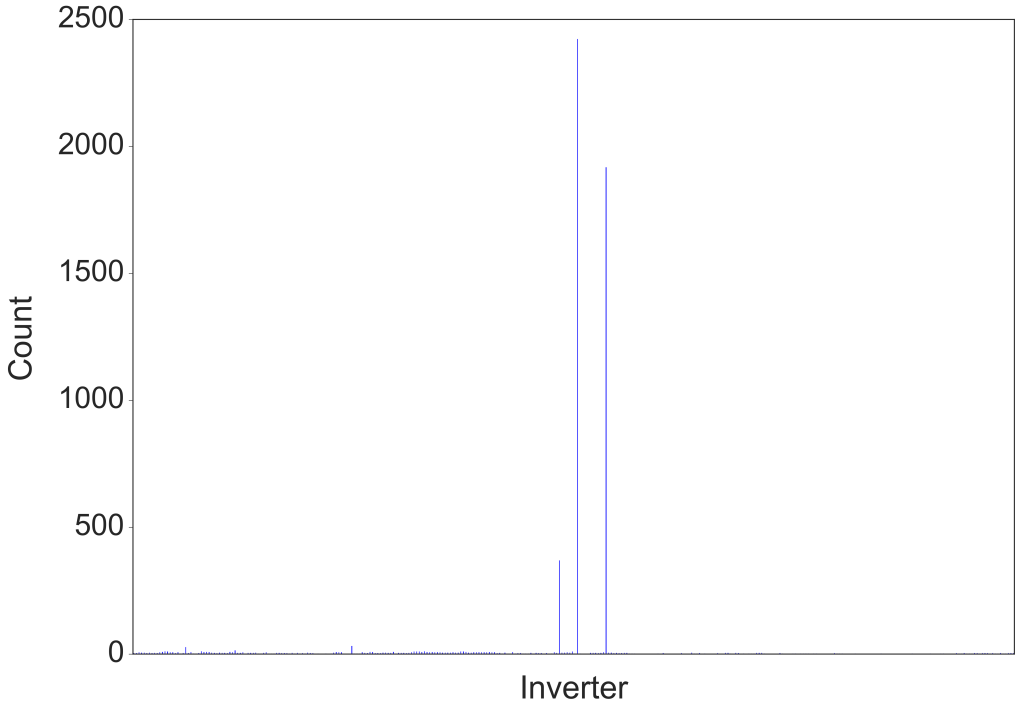

Figure 6.7: Count of *equipment failure* per inverter
from the Moerdijk plant for the year 2019

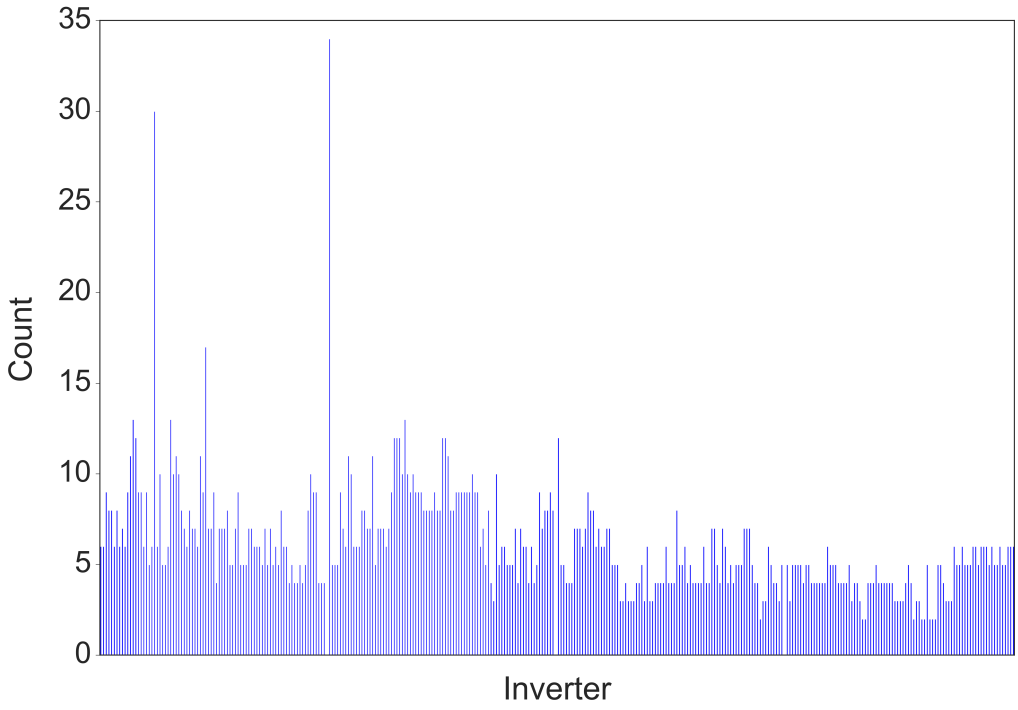

Figure 6.8: Modified Count of *equipment failure* per inverter from the Moerdijk plant for the year 2019

Since the inverter output could not be independently evaluated it was essential to assign every inverter

to a sensor. There were 5 substations which had sensors and 5 which did not, therefore both substations were analysed separately. In the case of substations with sensors, the possibility of having one representative model for all inverters in the substation was explored through statistical tests. Furthermore the inverters in substations without sensors were first allocated to a sensor block and the same test was repeated. The detailed method is explained below.

Substations with sensors

Out of the five substations that have GPOA and module temperature sensors, 103TSR has 33 inverters because one of the inverter(*103EIN-0002-0001*) was dropped from the dataset. The other four substations had 34 inverters each. In order to group all the inverters in a substation, a two independent sample t-test was performed. It yielded a numeric value which was indicative of how closely the active power profiles of each inverter were related 3.5 to be able to create one representative model. The null hypothesis H_0 for this t-test is mentioned below . The criterion for rejecting the null hypothesis was set at a P-value of 0.05. The criterion for accepting the null hypothesis was P greater than 0.05.

H_0 : The inverter power profiles are dissimilar and therefore cannot be grouped to create a single representative model.
if (P < 0.05) : reject H_0.
else : fail to reject H_0.

As a two independent sample t-test was used , a corresponding P value was generated between every pair of the inverter. The matrix 6.1.2 shows how the P-value matrix is formed, I_{1-1} represents the P value of I_1 with I_1, similarly I_{1-34} is between I_1 and I_{34}. Although in the case of substation 103TSR it was a 33X33 matrix, in comparison to 34X34 in the other substations. This was a symmetric matrix and hence only the upper triangle was evaluated. The mean of the mean of every column was calculated to get an overall P value. Since each sensor block has multiple inverters it was referred to as a cluster. The P values of all the clusters are tabulated in table 6.3. As all the P values were less than 0.05 the null hypothesis H_0 was rejected. *Therefore all the inverters power profiles are similar and therefore can be grouped to create a single representative model.*

$$\begin{Bmatrix} I_{1-1} & . & . & .I_{1-34} \\ . & & & \\ .. & & & \\ . & . & . & . \quad I_{34-34} \end{Bmatrix}$$

Cluster	P Value
C1	0.03
C2	0.024
C3	0.025
C4	0.04
C5	0.038

Table 6.3: Average P values obtained through two independent sample t-test for all inverter clusters

Substations without sensors

The substations which do not have a GPOA sensor were first assigned a sensor block and then the null hypothesis testing was repeated. In order to assign every inverter to a sensor, the first step was to count the number of timestamps when the maximum power produced coincided with the maximum GPOA of a sensor. The inverter was assigned to the cluster corresponding to the sensor from which the highest count was obtained. The snippet from the code is shown in figure 6.9. In the snippet the inverter *104EIN-0001-0001* had 157 counts of timestamps with the sensor 2, so it was assigned to cluster 2. The other inverters were allocated similarly.

```
                     C1   C2   C3   C5 cluster
104EIN-0001-0001    146  157  152  112      C2
104EIN-0001-0002    147  161  145  112      C2
104EIN-0001-0003    138  159  149  109      C2
104EIN-0001-0004    144  153  140  110      C2
104EIN-0002-0001    139  159  147  111      C2
104EIN-0002-0002    158  156  147  113      C1
```

Figure 6.9: Code snippet of allocation of inverters in substaions without sensors to a sensor cluster

The allocation of 337 inverters in the corresponding clusters is shown in table 6.4. It shows cluster 4 has highest number of inverters. Additionally, an updated t-test was performed to see if the inverters could still be grouped. The updated P values alongside the difference in the values are tabulated in table 6.4, the values were found to be less that 0.05 and the null hypothesis was still rejected.

Cluster	Count	Previous P value	Updated P value	Difference in P value
C1	35	0.03	0.0316	0.0016
C2	55	0.024	0.0246	0.006
C3	69	0.025	0.0271	0.0021
C4	100	0.04	0.048	0.008
C5	78	0.038	0.0412	0.0032

Table 6.4: Count per cluster and Change in P value

6.2. Baseline Model

Due to the significant variation of the features annually and the spatial distribution of sensor inverter pairs, two major questions were raised:

- Should a single annual model be built or will a monthly model be more suitable for the given data?
- Can one single global model be representative of the inverters across the plant or should there be a model built per cluster.

To answer these questions a simple linear regression model was built per cluster per month. The feature selected was only GPOA and the average active power of all the inverters was taken as the target variable. The slopes of the average fit were plotted against the months and are shown in figure 6.10. The plot shows that the average slope values during low irradiance months November-February are in the order of 20kW whereas the other months are in a significantly different order of 60-70 kW. This meant that using annual models was erroneous, in-spite of which, initially an annual model was built due to the ease of modelling. The value of slopes along different clusters was found similar as the the fit lines were almost overlapping. Hence a global model was the choice for a preliminary model. The initial choice of a global annual model was changed to global monthly models due to the high inaccuracy of the global annual model.

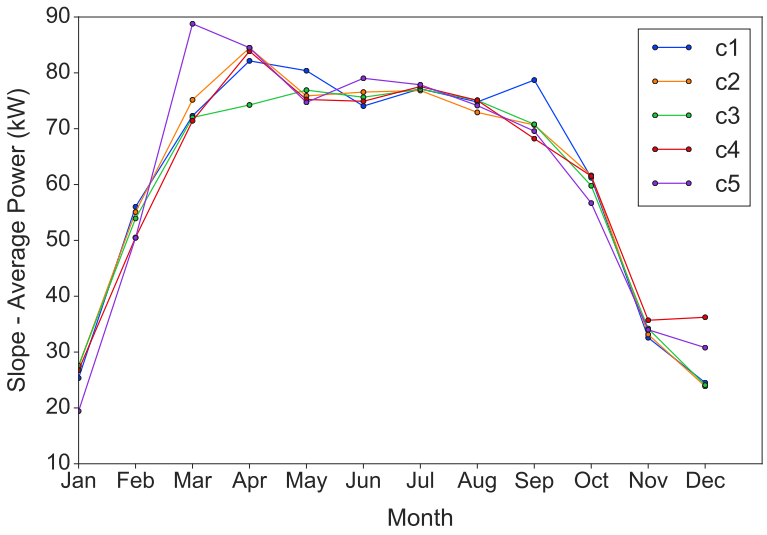

Figure 6.10: Average monthly fit across all clusters to identify the choice of global/local, annual/monthly models

6.2.1. Linear Regression model

Considering the global annual model, the number of inverters which would be required to represent normal operating conditions was decided to be 50. This was found by choosing inverters in the ascending order of their number of faults shown in figure 6.8. A simple linear regression model was built by adding one inverter at a time and the optimal value for inverters was chosen by minimising the MSE. The red line in the figure 6.11 shows that at 50 inverters the curve started saturating, the incremental decrease in the MSE error was below 3kW. The error would have further decreased if more inverters were added but the trade-off between decrease in error and increase in computational time was not justified.

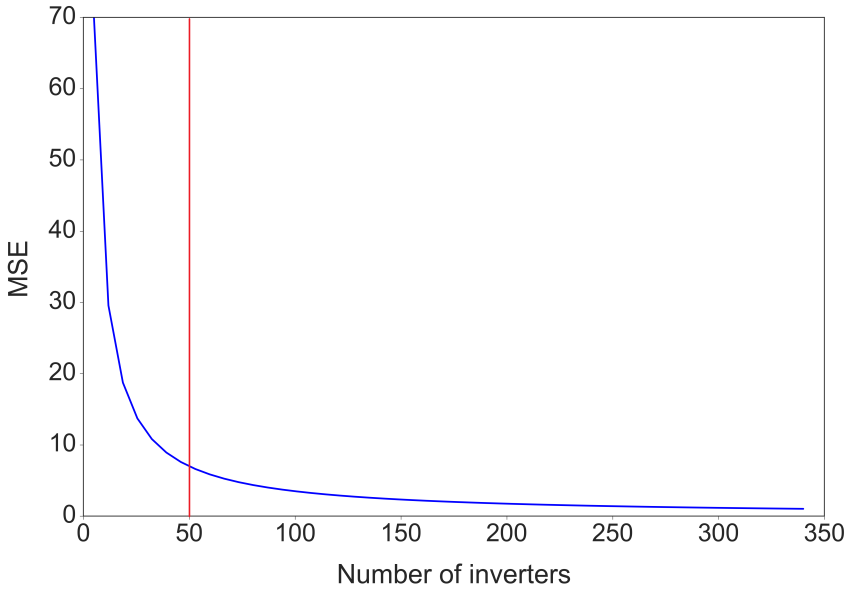

Figure 6.11: Selection of 50 inverters for modelling based on the MSE optimisation

Subsequently, with GPOA as feature and average active power of the 50 inverters as a target variable a simple linear regression model was built. The resulting model has been plotted in green in 6.12. The plot shows that the predicted values through linear regression (green line) are not representative of actual values

(blue line). The performance metric of the linear regression model is specified in table 6.5. A random day was plotted for one of the model inverters, which is shown in figure 6.13 .

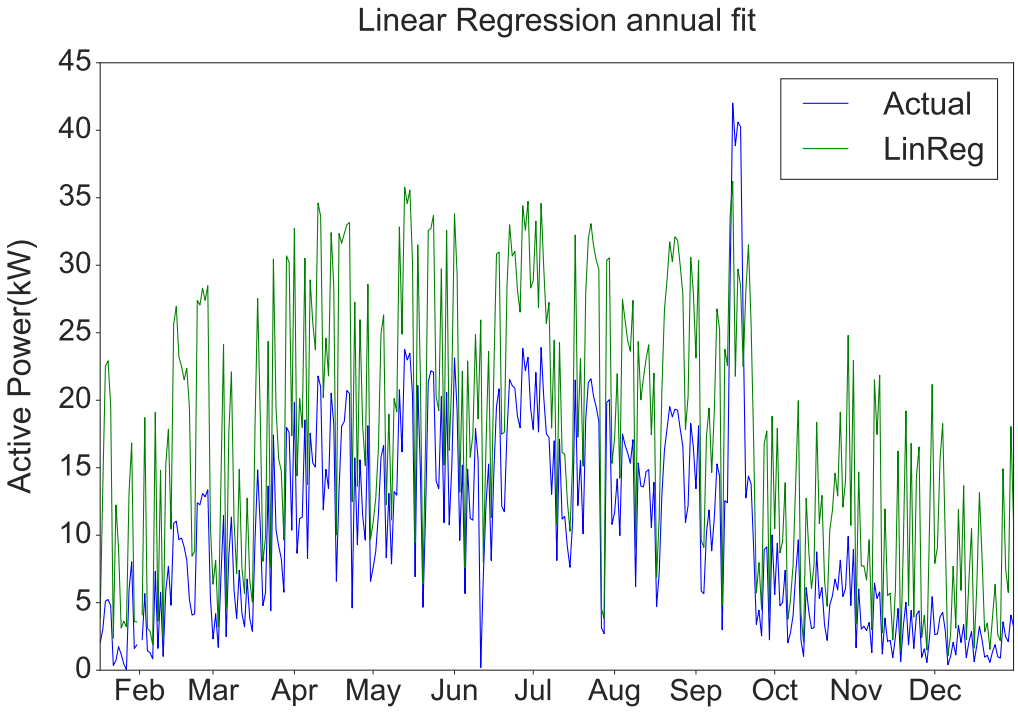

Figure 6.12: Yearly variation of regression fit vs actual values

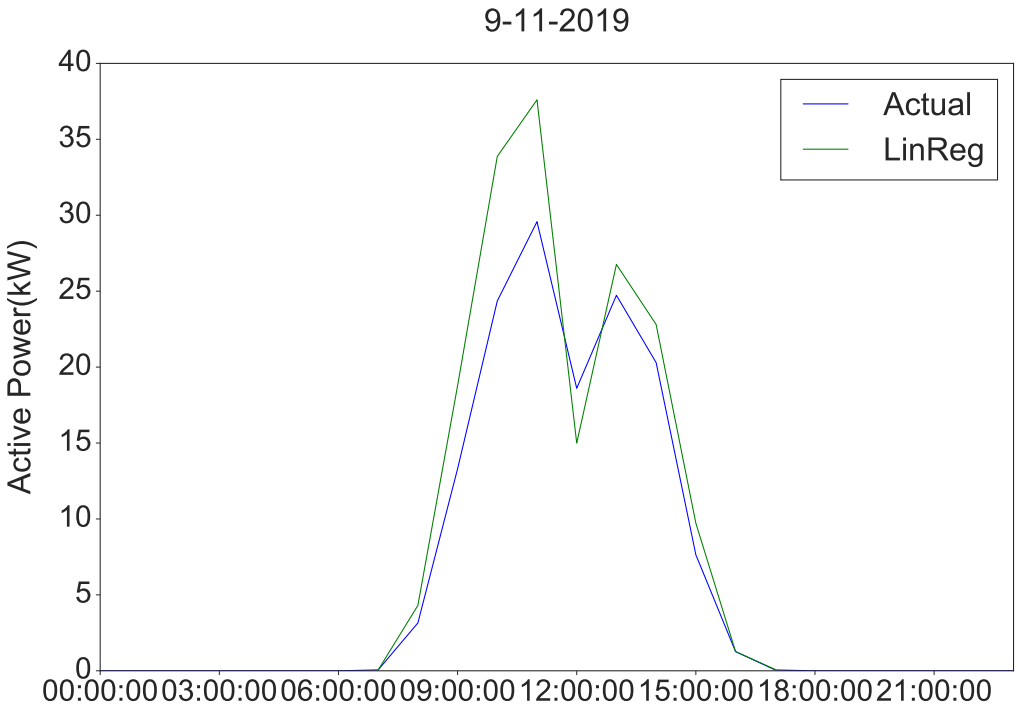

Figure 6.13: Predicted power values using linear regression vs actual power

In addition to this, the residuals, that is the difference between the actual and the predicted values are shown in 6.14. This shows the periodicity of errors which makes linear regression a sub optimal choice of prediction. This behaviour was expected due to the non linear time variance of the feature. In order to incorporate the time variance of the feature as an additional feature a time-lag was added in a more advanced algorithm.

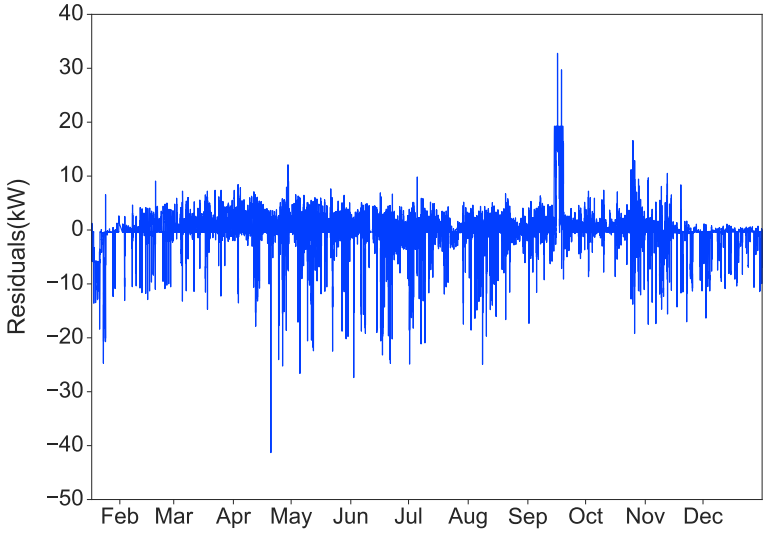

Figure 6.14: Simple Linear regression annual residual variation

	Lin Reg - Annual	**Elastic Net - Annual**	**Elastic Net - Monthly**	**GBM - 1% quantile**	**GBM - 99% quantile**
R^2	0.87	0.91	0.92	0.83	0.85
MAE(kW)	2.6	2.054	0.98	5.29	2.11
RMSE(kW)	5.5	3.39	1.8	6.68	4.8

Table 6.5: Performance metrics of linear regression, elastic net, GBM with quantile regression

6.3. Advanced modelling results

This section is divided into four subsections : Subsection 6.3.1 and 6.3.3 showcase the results of predictions of active power based, using the elastic net and GBM approach respectively. The results of using these predictions for early fault detection are shown in 6.3.2 and 6.3.3.

6.3.1. Elastic Net

Initially a global annual elastic net model was built. While building the elastic net model, module temperature was added as a feature. The features were time lagged to reduce their time dependence, the lags were further added as an additional feature. The feature importance based on their weights was plotted and is shown in figure 6.15. With increase in lag the feature importance decreases due to a decrease in the correlation. For GPOA 60 minutes was found to be the optimal timelag as any further lag would have a zero feature weight. Similarly a fifteen minutes lag was found to be enough for module temperature.

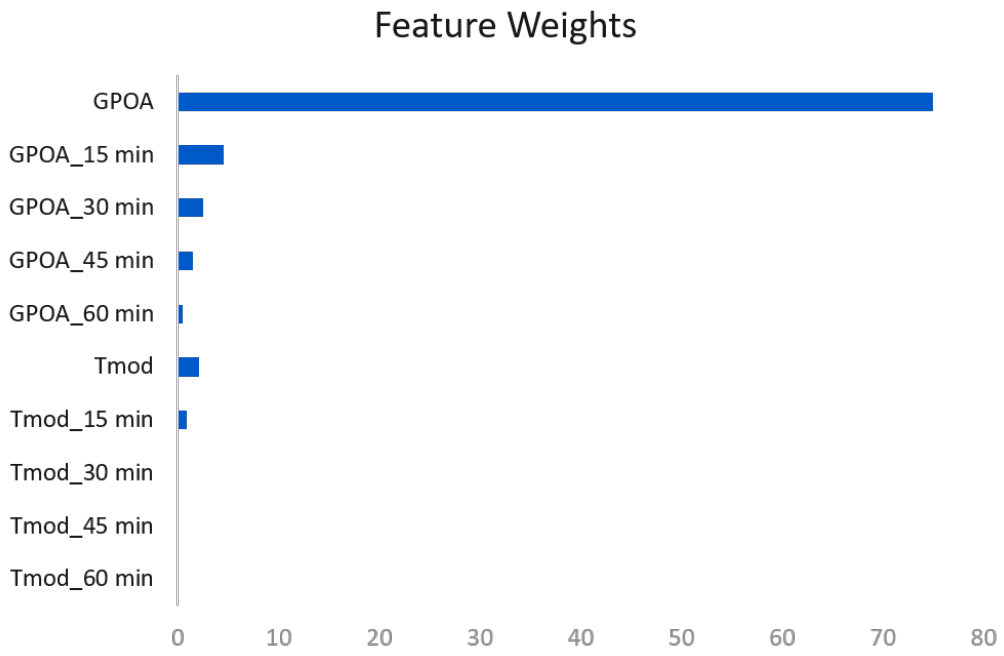

Figure 6.15: Feature weights in elastic net

The annual model had an unsatisfactory performance in comparison to the monthly models, the residuals of both models are shown in figure 6.16. The spread of the residuals of the annual residuals is broader in comparison to monthly models. This is because the annual model tries to generalise the behaviour of the values throughout the year, whereas the monthly model incorporates the monthly variation. The main reason for the lower performance of the annual models is the predictions during low irradiance months. This can be seen in figure 6.17 where the power variation during a low irradiance month is plotted. The annual model over predicts the power as compared to the monthly model.

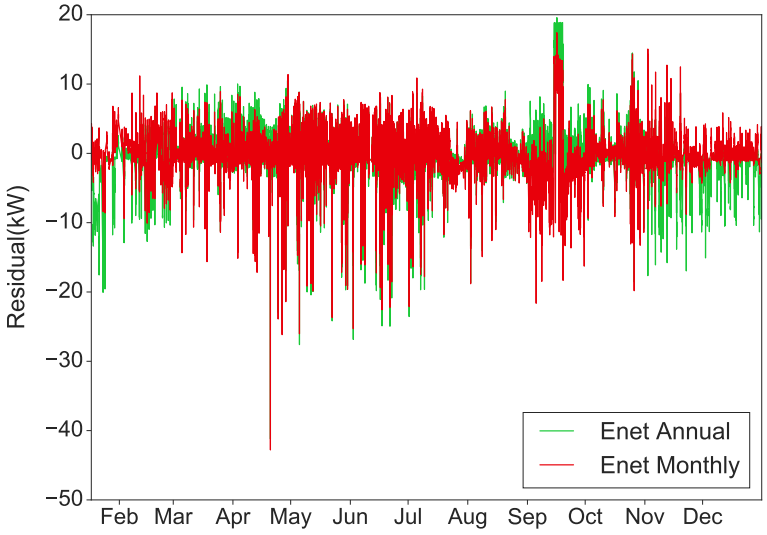

Figure 6.16: Elastic net residuals

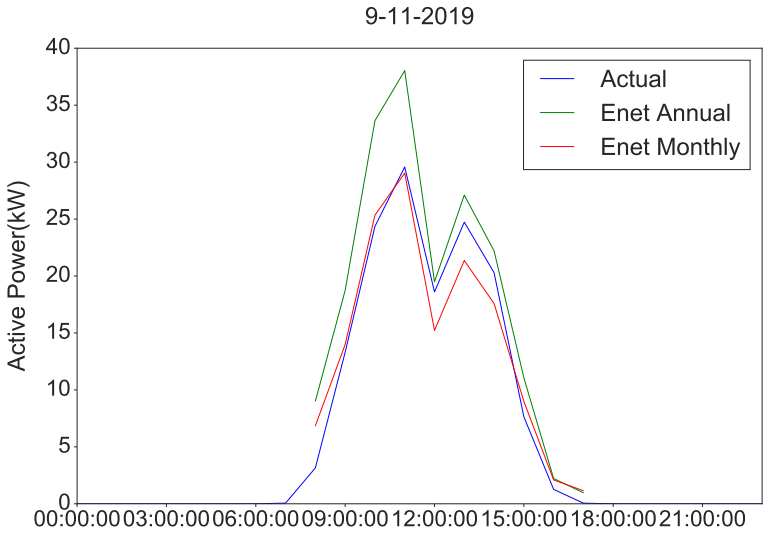

Figure 6.17: vVriation of actual, predicted power values during a day in a low irradiance month

To confirm the superiority of the monthly models a box plot was plotted and is shown in figure 6.18. The spread of residuals of annual model during the low irradiance months of Nov-Mar have a higher spread in comparison to monthly models. The performance based on RMSE of the monthly models was found to be 25% better than the annual model. This can be seen in metrics presented in table 6.5.

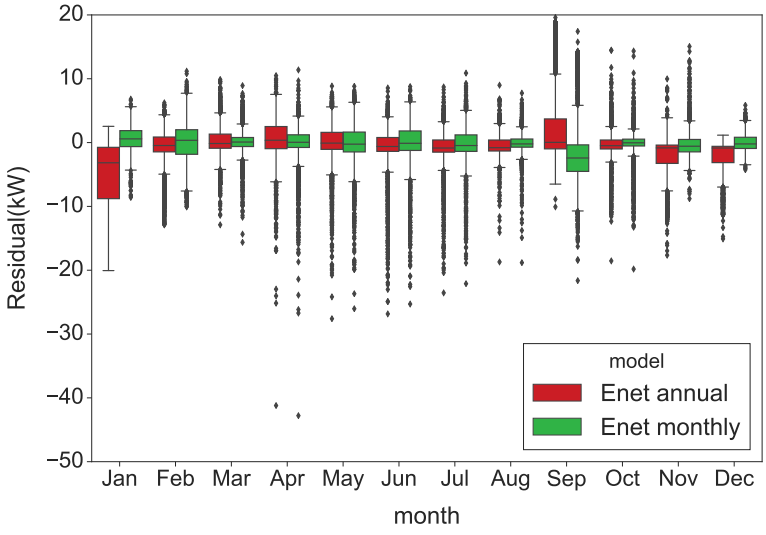

Figure 6.18: Comparing annual and monthly Elastic Net models

6.3.2. Early fault detection using Elastic Net

As the monthly predictions outperformed the annual model, monthly models were chosen to detect an early signal for failures. The first step was differentiating between modelling and actual error. In order to decide the extent of modelling error, the residuals were analysed. Due to the monthly variation of residuals it was found that $(-\sigma, 2\sigma)$ was representative of monthly modelling error and this represented the normal operating limits of the inverter.

The crossing of the residual from the defined control limits was a potential signal of an actual fault. The Lower control limit (LCL) was $-\sigma$ and Upper control limit (UCL) was 2σ. From the residual distribution of every month the LCL and UCL were extracted to generate the safe operating zone. This zone represented the allowable error, which means if the residual was within the limits then no signal could be detected. This conversion of residuals spread to safe operating zone is shown in figure 6.19. Close to 80% of the residuals were within the limits so for all other values there was a possibility of identification of the signal.

Once the safe operating limits were set, the recorded timestamps of *equipment failure* analysed for early signals. The testing was done on all *equipment failure* timestamps for all inverters. The initial result was promising as there was a visible signal seen 24 hours prior to the fault. This can be seen in the figures 6.20 , 6.21. The title of the figure mentions the exact timestamp of the occurrence, in the graph it is the ending of the curve which depicts the timestamp of the fault. When the residual signal was analysed 24 hours in advance it was found that it crossed the operating limits **thereby providing a basis for an early signal** . Although it did seem like a breakthrough this was actually a **false positive**. The spike in the residuals was the poor prediction of the model during very high and very low GPOA. This was found by plotting the residuals several days in advance of the fault which showed a **periodic structural error in the residuals**. It can be seen in figure 6.22, 6.23 . The residual can be seen crossing the limits whenever the GPOA was high (in the afternoon) and during highly fluctuating irradiance points. **Inspite of having different limits per month the approach still has fixed limits within the month and cannot mitigate the local spikes in the residuals**. Therefore, as an alternative approach to single prediction a range of predictions were adopted and are discussed in the next section.

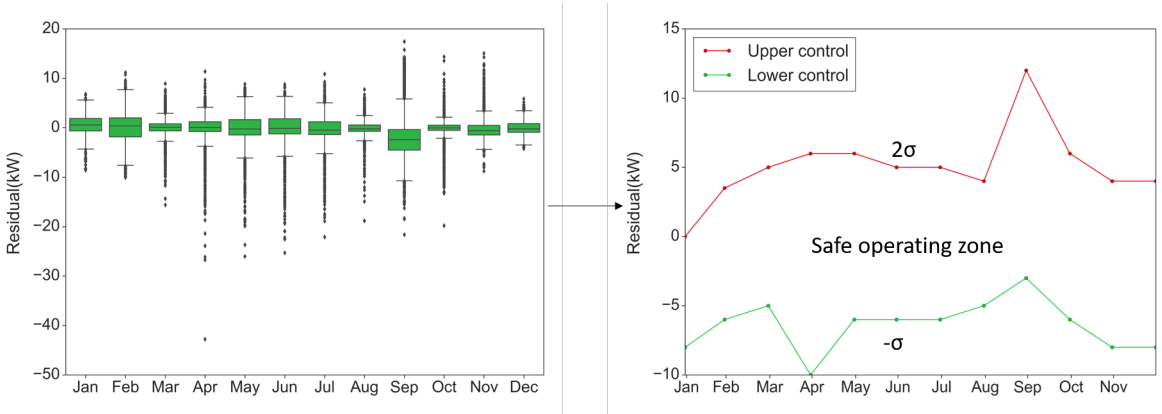

Figure 6.19: Safe operating zone for inverters

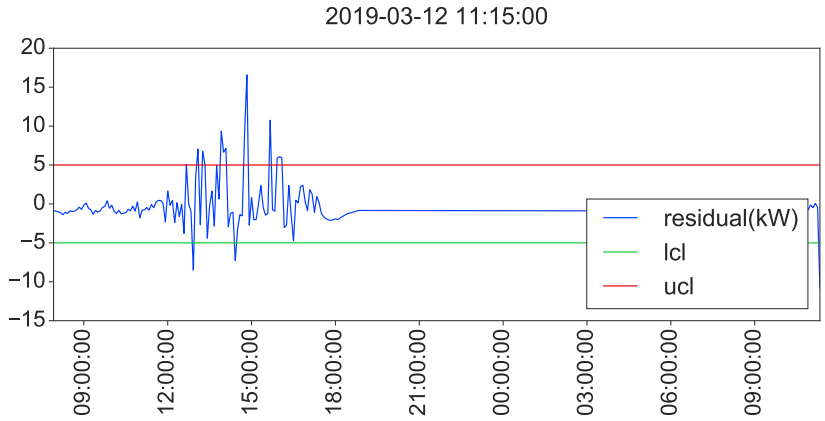

Figure 6.20: *100EIN-0001-0001* Equipment failure timestamp with 24 hours historical variation showing an early signal

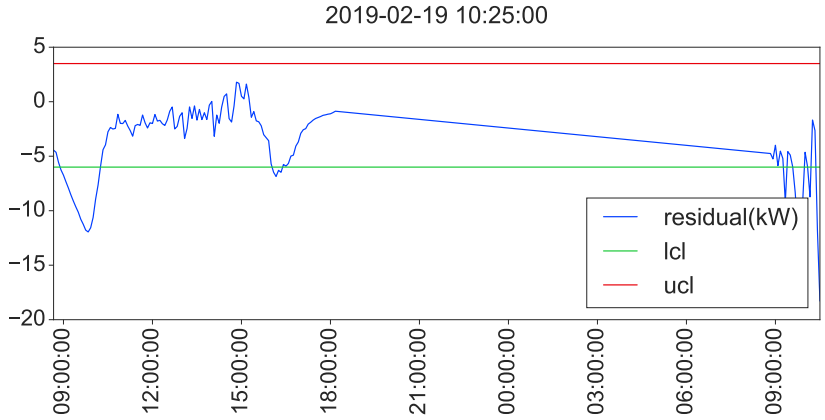

Figure 6.21: *102EIN-0003-0004* Equipment failure timestamp with 24 hours historical variation showing an early signal

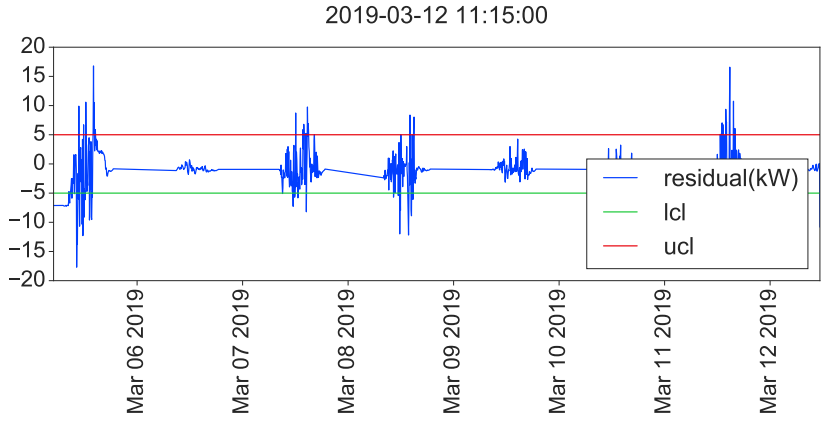

Figure 6.22: *100EIN-0001-0001* Equipment failure timestamp with 24 hours historical variation

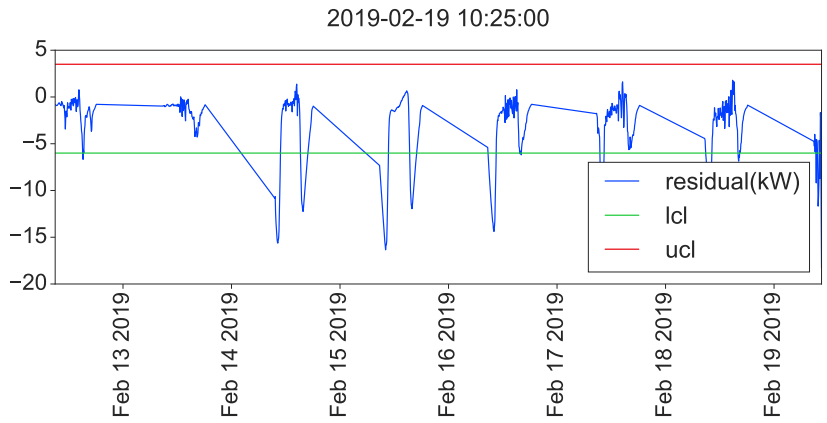

Figure 6.23: *102EIN-0003-0004* Equipment failure timestamp with 24 hours historical variation

6.3.3. GBM with quantile loss function

The GBM approach was aimed towards creating a range of predictions in contrast to elastic net where only a single prediction was made. In order to get a range of predictions a quantile loss functions was used instead of a squared loss. The quantiles selected were 0.01 and 0.99, representing the very low and very high irradiance respectively. This choice of percentiles was partly based on literature[8]. The theory behind GBM with quantile loss is explained in chapter 3. Separate monthly models models were built for lower quartile of 0.01 and the higher quantile of 0.99. The figure 6.24 shows how this method creates a range of predictions. The actual power is bounded by the upper and the lower quantile. The metrics for the both the quantiles is presented in the table 6.5.

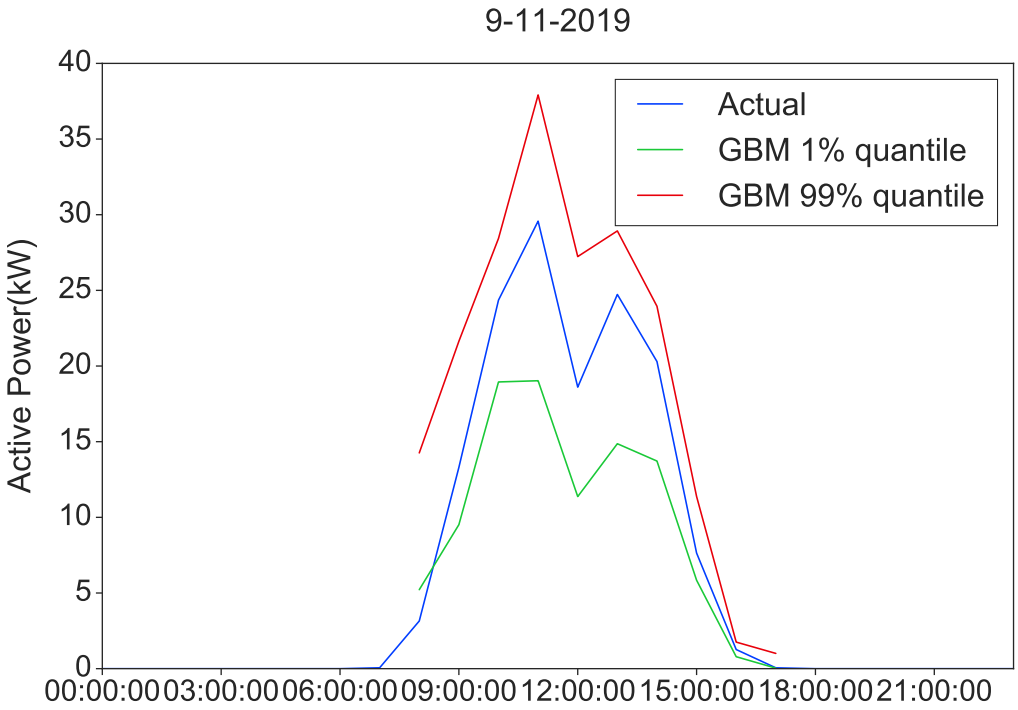

Figure 6.24: GBM predictions based on quantile regression

6.3.4. Early fault detection using GBM

Similar to the elastic net early detection, the timestamps at which the *equipment failure* faults occurred were analysed to find if any early signal was visible. Unlike the elastic net approach where the limits were fixed throughout the month, the limits in this approach were varying continuously, thereby reducing the possibility of false positive predictions due to structural errors. The figure 6.25 shows the historical power generation before the occurrence of a fault. The actual power generation does dips only at the point of time of fault, this is clearly visible in figure 6.26. Since the test across the inverters did not yield any early signal it was decided to probe the faults into more detail.

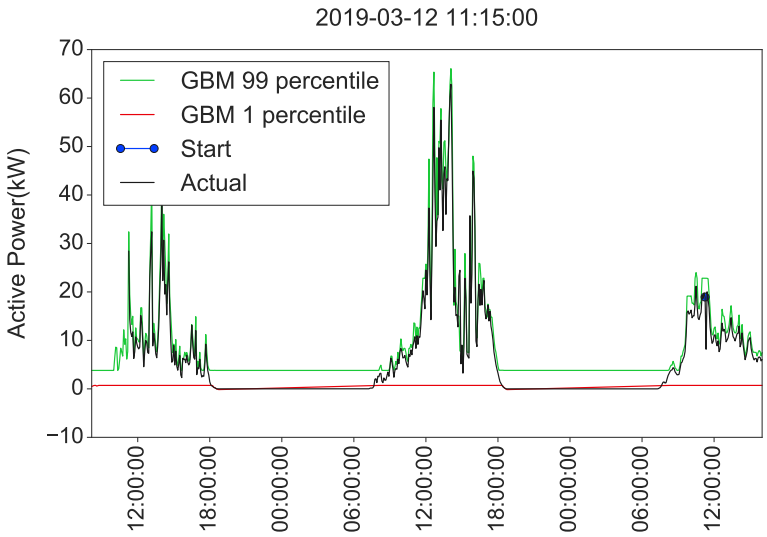

Figure 6.25: *100EIN-0001-0001* Equipment failure timestamp with 24 hours historical variation

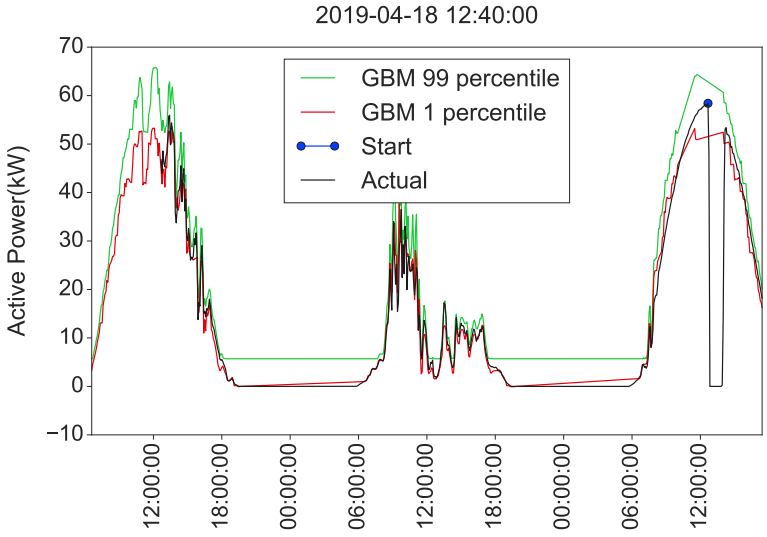

Figure 6.26: *100EIN-0001-0001* Equipment failure timestamp with 24 hours historical variation

6.4. Advanced fault analysis

As per the current architecture of fault tagging in the SCADA, *equipment failure* *AC disturbance* were supposed to be triggered based on actual inverter failure and external grid disturbances respectively. The figure 6.27 shows that in addition to the *AC disturbance failure* an *equipment failure* was also recorded whenever

there was an external grid disturbance. This led to the **erroneous registration** of *equipment failure*, whereas in reality it was actually an *AC disturbance* due to grid fluctuation. The monthly variation of count of *AC fault*s and *equipment failure*, was plotted and this is shown in 6.28. In this figure it can be seen that both the counts of faults almost overlap, which meant the frequency before the fault timestamps needed to be checked. The frequency was plotted alongside the power values in figure 6.29 and figure 6.30. In the figure 6.29 it can be seen that the dip in frequency (orange line) coincided with the start time of the fault depicted by a black dot. This timestamp was registered as an *equipment failure* but it was clear that there was no actual failure, as all the inverters recorded the same timestamp and had the same frequency characteristics as shown. Therefore it made no sense to predict a fault which was preceded by an external event as the external event (in this case grid) has its own uncertainties and cannot be quantified with dataset of one plant. Furthermore all such faults were dropped from the dataset, which left only 7 instances for establishing a signal. One of the seven cases has been shown in figure 6.30. In this figure at the starting of the fault shown by a bl dot, the frequency remains stable showing that this was an actual fault. No early signal could be found detected in these cases, creating a need for availability of more data.

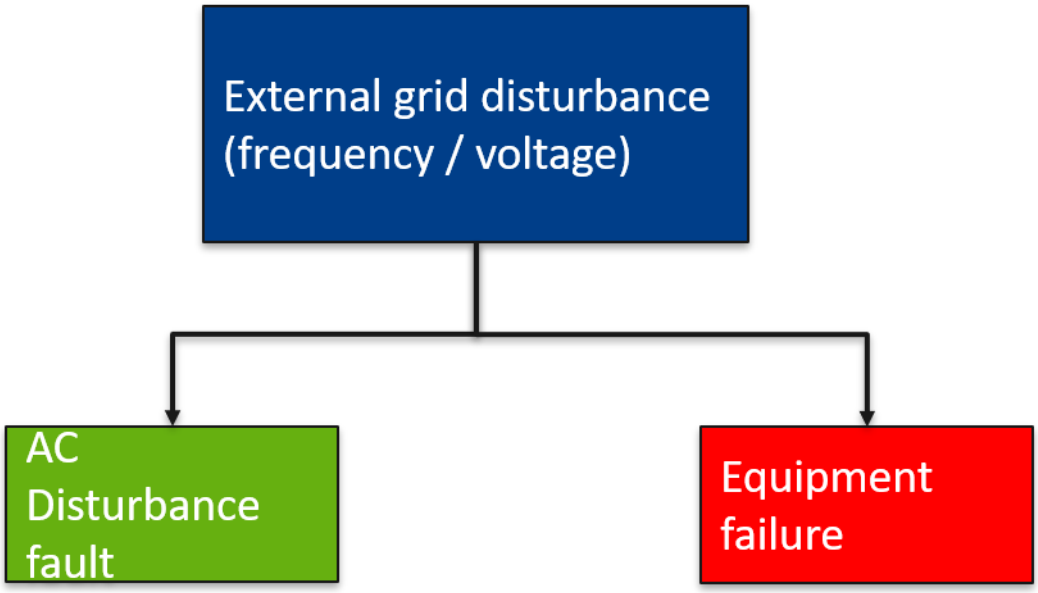

Figure 6.27: Erroneous mapping of faults in the SCADA

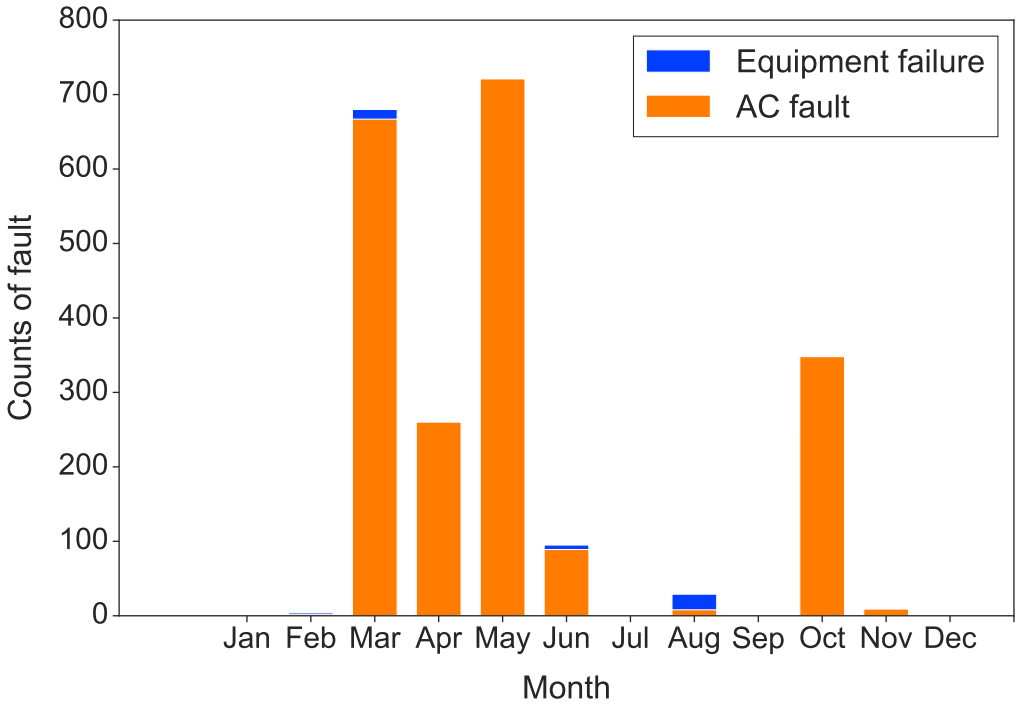

Figure 6.28: Monthly variation of counts of equipment failure and AC fault

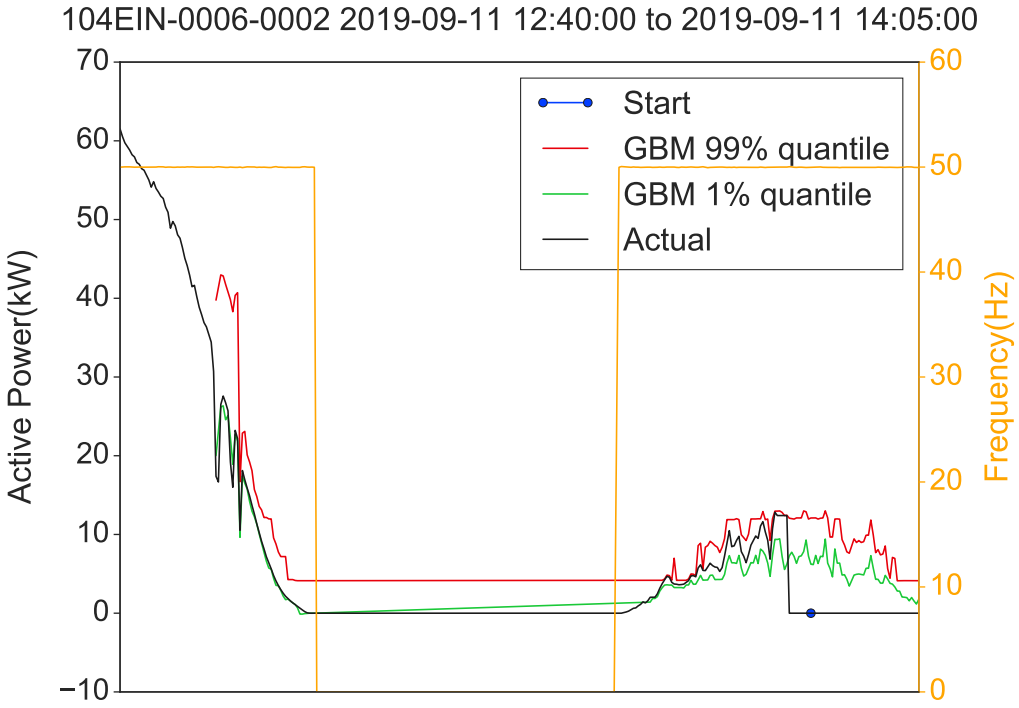

Figure 6.30: Actual equipment failure

In summary, an Elastic net and a GBM were used to develop predictions for active power produced by the inverters. The performance of monthly Elastic net models was found to be 25% better than annual models.

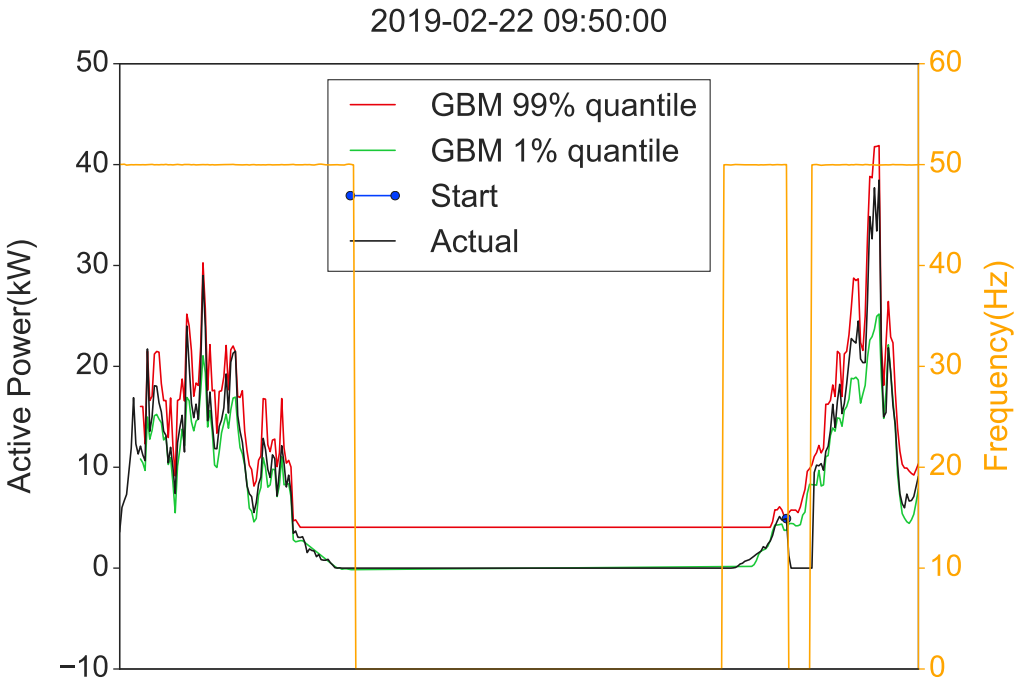

Figure 6.29: Equipment failure preceded by AC fault shown by the dip in the frequency

The difference between the predicted and the actual value referred as residuals was used to find an early signal. The visibility of early signal was dependant on differentiating between modelling and actual error, therefore a normal operating zone based on the distribution of residuals was set. $(-\sigma, +2\sigma)$ were decided as LCL and UCL respectively. The first test resulted in a possibility of an early signal, but was found to be a structural error in the predictions. To overcome that a GBM approach with quantile regression was built and tested. Unlike the elastic net this approach had a range of predictions which eliminated the possibility of momentary spikes leading to a structural error. Data quality issues were found in the SCADA fault tags, which limited the availability of fault data for early signal detection and establishing a pattern for generating predictive maintenance alerts.

7

Conclusion & Discussion

The main objective of this research was to build a machine learning algorithm which would detect an early signal for inverter faults in order to generate predictive maintenance alerts. Two methods for early fault detection were developed: Elastic Net & GBM with quantile loss.

- The results from both algorithms indicated no early signal for inverter fault, this was partly owing to the issues with data quality. This highlights the dependency of the algorithm on the quantity and quality of the data.

After presenting the major findings the answers to the research questions stated in chapter 1 are summarized below.

Q1 : For the given data should the model be global or local, additionally should it be annual or monthly?

This question was important to answer before starting the modelling because there were ten substations but only five had GPOA sensors. The baseline linear regression model was built per cluster, this can be seen in figure 6.10. From the variation of fits across clusters, it can be concluded that if the inverter power profiles are statistically similiar , then one global model is suitable to represent the normal operation of all inverters in the plant. The research also concludes that if the dataset is available only for a year then it is better to build monthly models over one single annual model. In this research monthly models outperformed the annual model by 25%. This is because of the high errors during the low irrdiance months where the loss rises exponentially.

Q2 : Which loss function is more suitable towards early detection inverter faults?

The principle behind fault detection using ML is to distinguish between modelling error and actual error and trace a signal before the occurrence of a recorded fault. In case of normal loss functions like squared error, only one value is predicted. Due to the variability of the main feature GPOA throughout the day, single predictions can turn out to be erroneous at extreme irrdiance case leading to a periodic structural error. This can be seen in the figure 6.23 where the periodicity of the spikes confirm the existence of structural error Therefore this research concludes that it is better to use quantile loss while building any algorithm based on GPOA as a feature. This is due to the fact that a range of predictions can be created for different quantiles of the irradiance. Additionally quantile loss helps in creating continuously varying safe operation limits in comparison to squared loss approach where safe operating limit is constant throughout a month. In this research the quantiles 0.01 & 0.99 were found suitable and they can be seen in figure 6.24.

Q3 : Is predictive maintenance economically feasible to implement in the Moerdijk case?

Based on the analysis of current OPeX it was found that 10% potential saving can be obtained if predictive maintenance solution is implemented in Moerdijk. This saving is due to the variable cost involved in the current O&M provider's contract which can be replaced with the help of predictive maintenance. The cost

distribution saving potential can be seen in figure 5.3.

Up to the knowledge of the author, currently only three prominent researches [7],[11],[9] have been published for the predictive maintenance of inveters in PV parks. [11] used a supervised learning approach. However they do not mention the type of faults that were predicted, which raised the question on the reliability of the results. As observed in this research that the quality of fault data can significantly affect the possibility of building an algorithm; the unavailability of fault tags means that even grid related faults were predicted. This may provide with a false accuracy of the algorithm as any external fault is very stochastic in nature and it cannot be predicted based on one plant alone. The research by [7] uses data from ten different plants to create an unsupervised learning approach. This research had specific fault class reducing the ambiguity of the results. Finally the research by [9] evaluates the predictive maintenance approaches qualitatively and does not give any actual results in terms of actual early fault detection.

This Master thesis research has created an inhouse understanding towards the development of predictive maintenance methods for inverters of utility scale PV parks.

8

Recommendations

This chapter presents the recommendations for Shell and the type of further research that can be conducted.

8.1. Recommendations to Shell

- The SCADA monitoring system, should have faults corresponding to the Alarm ID specified in the inverter manual. In the current SCADA system used in Moerdijk , the faults are mapped incorrectly. An external disturbance of grid causes an internal *equipment failure* fault to get triggered, which becomes confusing while analysing. Therefore the mapping of faults should not be intertwined, otherwise it may lead to erroneous analysis.

- After meeting with inverter experts, it has come to light that most of the recent solar plants have a monitoring frequency in range of milliseconds, whereas Moerdijk has a constant granularity of 5 minutes. This unavailability of data at a granular level creates a bottleneck to study the dynamics of the fault , as the fault occurrence is in the time range of 60-80 milliseconds. Hence it is recommended to arrange a separate channel monitoring the inverter data with a milliseconds granularity. In case this is a financially unsuitable option, a possibility of sourcing this granularity of data is through the manufacturer, as most manufacturers already have such data.

- An Application Programming Interface (API) to extract data from SCADA is needed for automatic extraction of data through scripts for building any algorithm. Currently a manual query has to be generated to download the data, this can slow the process of analysing the data and building suitable algorithms. An addition of API is recommended for the Moerdijk and future plants.

8.2. Research Recommendations

The main conclusions of the research show that no early signal was found partly due to issues in data quality leading to lack of training data. Therefore, it is recommend to obtain data from multiple plants/sources to avoid lack of training data. The research questions that need to be answered with the availability of more data are :

- Do inverters have an early signal before failure? Can it be identified using SCADA level data?

- How can a software failure be differentiated from a hardware failure?

- Should the normal operation of inverter be defined in the early stage of its operation, because as per the bathtub curve the initial faults are higher, this needs to be discussed with manufacturers.

Alternatively , if millisecond granularity of data is available then the faults can be modelled more accurately. It can be combined with inverter models to generate synthetic fault responses, thereby eliminating the need for more data. Although there should be enough data for validation of the response towards synthetic data.

This research did not focus on any method involving electrical signature analysis as performed for an electrical generator by [41]. The research applies fast fourier transforms (fft) to generate current and voltage

signatures. Then faults were simulated in a lab setup and the difference between the signatures was established to detect early signals for fault detection. The author recommends in performing a study based on this approach as an alternative to ML based early fault detection.

A

Hyper-parameter details

This appendix presents the basic theory regarding the hyper-parameters and their tuning. Additionally, it presents values of optimal hyper-parameters obtained in this research .

A.1. Hyper parameter optimisation

Hyper-parameter is a term defined in Bayesian statistics as the parameter of prior distribution. The term is used to distinguish them from model parameters. Hyperparameters can have a direct impact on the performance of machine learning algorithms. Thus, in order to achieve maximal performance, it is important to understand the optimization of hyper-parameters. The common strategies for optimizing hyper-parameters are listed below:

- **Grid Search** : Search a set of manually predefined hyper-parameters and iterate over all the combinations to find out the best configuration. This method is very simple but if the range of the hyper-parameters is very wide then it can be computationally very challenging.
- **Random Search** : It is very Similar to the grid search, but replaces the exhaustive search with a random search. This can outperform grid search when only a small number of hyper-parameters are needed to actually optimize the algorithm.
- **Bayesian Optimization** : This method builds a probabilistic model of the function by mapping hyper-parameter values to the target variable and is then evaluated on a validation set.
- **Gradient-Based Optimization** : This approach computes gradient using initially guessed hyper-parameters and tries to optimize by using gradient descent.

In this research grid search was first used to find the optimal values for hyper-parameters and it was cross validated with random search.

A.1.1. Elastic net hyper parameters

The Elastic net has two hyperparameters : α, l1 ratio. l1 ratio decides whether l1 norm dominates over l2 norm, as elastic net is a combination of ridge regression andLASSO . l1 ratio with value of zero corresponds to a LASSO and , value of one corresponds to a ridge regression. The optimal values obtained for both the elastic net models were:
$\alpha := 0.01$
$l1_{ratio} := 0.9$

A.1.2. GBM hyper-parameters

A GBM model contains two categories of hyperparameters: boosting hyperparameters and tree-specific hyperparameters.

The boosting hyper-parameters are listed below:

- Number of trees : There can be several thousand trees, it is important to optimise them to avoid over-fitting.
- Learning rate : A low learning rate is preffered to have a better performance, but if the dataset is very large then low learning rate can be computationally intensive.

The two main tree hyper-parameters in a GBM model include:

- Tree depth : Controls the depth of the individual trees. Typical values range from a depth of 3–8.
- Minimum number of observations in terminal nodes : It also controls the complexity of each tree. Since we tend to use shorter trees this rarely has a large impact on performance. Typical values range from 5–15.

The hyper parameters for both the quantiles are specified in table A.1 and A.2.

Loss	quantile
learning rate	0.4
max depth	15
min sample split	800
$n_{estimators}$	80

Table A.1: hyper-parameters for 0.01 quantile

Loss	quantile
learning rate	0.4
max depth	15
min sample split	400
$n_{estimators}$	100

Table A.2: hyper-parameters for 0.99 quantile

B

Code Base

The algorithms were built on Python 3.0 , the IDE used was Jupyter notebook. A copy of the final code is preserved on the Git Hub repository of Shell. The versions of the libraries used is shown in table B.1. All the notebooks were locally run, the specifications of the machine are given below:

- *RAM* : 16 GB
- *Processor* : i7 8665U, 1.9 Ghz CPU
- *Integrated graphic card*

Library	**Version**
scikit-learn	0.21.3
pandas	0.24.2
numpy	1.17.0
matplotlib	3.1.1
seaborn	0.9.0
statsmodel	0.10.1

Table B.1: Library versions

Pipelines were built in order to avoid data leakage and replication of the code. Minmaxscaler() was used to scale the features and target as it is the most commonly used scaling technique.

Bibliography

[1] Amsterdam, Netherlands — Sunrise, Sunset, and Daylength, June 2019. URL https://www.timeanddate.com/sun/netherlands/amsterdam?month=6.

[2] Mohammed Khorshed Alam, Faisal H. Khan, Jay Johnson, and Jack Flicker. PV faults: Overview, modeling, prevention and detection techniques. *2013 IEEE 14th Workshop on Control and Modeling for Power Electronics, COMPEL 2013*, 2013. doi: 10.1109/COMPEL.2013.6626400.

[3] Miguel Ángel, Fernández Fernández, Juan Luis, Carús Candás, Pablo Barredo Gil, Antonio Miranda, De Torre, Gabriel Díaz Orueta, and Ada Byron. An Industry 4 . 0 Approach for Photovoltaic Plants. pages 2–5, 2018. doi: 10.3390/proceedings2231409.

[4] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de Pison, and F. Antonanzas-Torres. Review of photovoltaic power forecasting, 2016. ISSN 0038092X.

[5] Rachad Antonius. Inferential Statistics: Hypothesis Testing. In *Interpreting Quantitative Data with SPSS*. 2011. doi: 10.4135/9781849209328.n10.

[6] Sharif Atique, Subrina Noureen, Vishwajit Roy, Vinitha Subburaj, Stephen Bayne, and Joshua MacFie. Forecasting of total daily solar energy generation using ARIMA: A case study. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference, CCWC 2019*, 2019. ISBN 9781728105543. doi: 10.1109/CCWC.2019.8666481.

[7] Alessandro (Enel Green Power Rome) Betti and Maria (Enel Green Power Rome) Lo Trovato. PREDICTIVE MAINTENANCE IN PHOTOVOLTAIC PLANTS WITH A BIG DATA APPROACH. In *EUPVSEC*, pages 1895 – 1900, sep 2017. ISBN 9788578110796. doi: 10.1088/1751-8113/44/8/085201. URL 10.4229/EUPVSEC20172017-6DP.2.4https://arxiv.org/pdf/1901.10855.

[8] Konrad Bogner, Florian Pappenberger, and Massimiliano Zappa. Machine Learning Techniques for Predicting the Energy Consumption/Production and Its Uncertainties Driven by Meteorological Observations and Forecasts. *Sustainability*, 11(12):3328, 2019. ISSN 2071-1050. doi: 10.3390/su11123328.

[9] Lisa B. Bosman, Walter D. Leon-Salas, William Hutzel, and Esteban A. Soto. PV system predictive maintenance: Challenges, current approaches, and opportunities. *Energies*, 16(3), 2020. ISSN 19961073. doi: 10.3390/en13061398.

[10] Elmer Collins, Michael Dvorack, Jeff Mahn, Michael Mundt, and Michael Quintana. Reliability and availability analysis of a fielded photovoltaic system. *Conference Record of the IEEE Photovoltaic Specialists Conference*, pages 002316–002321, 2009. ISSN 01608371. doi: 10.1109/PVSC.2009.5411343.

[11] Massimiliano De Benedetti, Fabio Leonardi, Fabrizio Messina, Corrado Santoro, and Athanasios Vasilakos. Anomaly detection and predictive maintenance for photovoltaic systems. *Neurocomputing*, 310: 59–68, 2018. ISSN 18728286. doi: 10.1016/j.neucom.2018.05.017. URL https://doi.org/10.1016/j.neucom.2018.05.017.

[12] Fault Detection, Protection In, Solar Photovoltaic, and A Dissertation Presented. *PROTECTION IN SOLAR PHOTOVOLTAIC ARRAYS Acknowledgments*. PhD thesis, North Eastern University, 2015. URL https://repository.library.northeastern.edu/downloads/neu:m039kr12f?datastream_id=content.

[13] Tyler J. Formica, Hassan Abbas Khan, and Michael G. Pecht. The Effect of Inverter Failures on the Return on Investment of Solar Photovoltaic Systems. *IEEE Access*, 5:21336–21343, 2017. ISSN 21693536. doi: 10.1109/ACCESS.2017.2753246.

[14] Ran Fu, David Feldman, Robert Margolis, Mike Woodhouse, Kristen Ardani, Ran Fu, David Feldman, Robert Margolis, Mike Woodhouse, and Kristen Ardani. U . S . Solar Photovoltaic System Cost Benchmark : Q1 2017 U . S . Solar Photovoltaic System Cost Benchmark : Q1 2017. *National Renewable Energy Laboratory*, (September):1–59, 2017. URL https://www.nrel.gov/docs/fy17osti/68925.pdf.

[15] Elyes Garoudja, Fouzi Harrou, Ying Sun, Kamel Kara, Aissa Chouder, and Santiago Silvestre. Statistical fault detection in photovoltaic systems. *Solar Energy*, 150:485–499, 2017. ISSN 0038092X. doi: 10.1016/j.solener.2017.04.043. URL http://dx.doi.org/10.1016/j.solener.2017.04.043.

[16] Thomas D Gauthier and Mark E Hawley. CHAPTER 5 - STATISTICAL METHODS. pages 129–183. Academic Press, Burlington, 2007. ISBN 978-0-12-369522-2. doi: https://doi.org/10.1016/B978-012369522-2/50006-3. URL http://www.sciencedirect.com/science/article/pii/B9780123695222500063.

[17] GitHub. No Title. URL http://ethen8181.github.io/machine-learning/model_selection/model_selection.html.

[18] Anastasios Golnas. PV system reliability: An operator's perspective. *Conference Record of the IEEE Photovoltaic Specialists Conference*, (PART 2):1–6, 2012. ISSN 01608371. doi: 10.1109/pvsc-vol2.2013.6656744.

[19] P Grover. No Title, 2017. URL https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d.

[20] Abdul Haq, Jennifer Brown, and Elena Moltchanova. A New Exponentially Weighted Moving Average Control Chart for Monitoring the Process Mean. *Quality and Reliability Engineering International*, 31 (8):1623–1640, 2015. ISSN 10991638. doi: 10.1002/qre.1696.

[21] Eric Haun. Predictive maintenance. Technical Report 5, Shell, 2019.

[22] H. S. Huang, J. C. Jao, K. L. Yen, and C. T. Tsai. Performance and availability analyses of PV generation systems in Taiwan. *World Academy of Science, Engineering and Technology*, 78(6):309–313, 2011. ISSN 2010376X. doi: 10.5281/zenodo.1333003.

[23] Robert Kaplar, Reinhard Brock, Sandeepan DasGupta, Matthew Marinella, Andrew Starbuck, Armando Fresquez, Sigifredo Gonzalez, Jennifer Granata, Michael Quintana, Mark Smith, and Stanley Atcitty. PV inverter performance and reliability: What is the role of the IGBT? *Conference Record of the IEEE Photovoltaic Specialists Conference*, pages 001842–001847, 2011. ISSN 01608371. doi: 10.1109/PVSC.2011.6186311.

[24] Joseph M. Kuitche, Rong Pan, and Govindasamy Tamizhmani. Investigation of dominant failure mode(s) for field-aged crystalline silicon PV modules under desert climatic conditions. *IEEE Journal of Photovoltaics*, 4(3):814–826, 2014. ISSN 21563381. doi: 10.1109/JPHOTOV.2014.2308720.

[25] S Kurtz, J Newmiller, A Kimber, R Flottemesch, E Riley, T Dierauf, J McKee, and P Krishnani. Analysis of Photovoltaic System Energy Performance Evaluation Method. *National Renewable Energy Laboratory*, TP-5200-60(November 2013):1–54, 2013.

[26] Yanting Li, Yan Su, and Lianjie Shu. An ARMAX model for forecasting the power output of a grid connected photovoltaic system. *Renewable Energy*, 2014. ISSN 09601481. doi: 10.1016/j.renene.2013.11.067.

[27] Isidoro Lillo-Bravo, Pablo González-Martínez, Miguel Larrañeta, and José Guasumba-Codena. Impact of energy losses due to failures on photovoltaic plant energy balance. *Energies*, 11(2), 2018. ISSN 19961073. doi: 10.3390/en11020363.

[28] Zhao Lin and Hao Ma. Modeling and analysis of three-phase inverter based on generalized state space averaging method. In *IECON Proc. (Industrial Electron. Conf.*, 2013. ISBN 9781479902248. doi: 10.1109/IECON.2013.6699271.

[29] Claudio Monteiro, L. Alfredo Fernandez-Jimenez, Ignacio J. Ramirez-Rosado, Andres Muñoz-Jimenez, and Pedro M. Lara-Santillan. Short-term forecasting models for photovoltaic plants: Analytical versus soft-computing techniques. *Mathematical Problems in Engineering*, 2013. ISSN 1024123X. doi: 10.1155/2013/767284.

[30] David Moser, Giorgio Belluardo, Matteo Del Buono, Walter Bresciani, Elisa Veronese, Ulrike Jahn, Magnus Herz, Eckart Janknecht, Erin Ndrio, Karel De Brabandere, and Mauricio Richter. Report on Technical Risks in PV Project Development and PV Plant Operation. *Solar Bankability WP1 Deliverable D1.1 and WP2 Deliverable D2.1*, 1:1–109, 2016. URL `http://www.solarbankability.org/results/technical-risks.html`.

[31] Alfredo Nespoli, Emanuele Ogliari, Sonia Leva, Alessandro Massi Pavan, Adel Mellit, Vanni Lughi, and Alberto Dolara. Day-ahead photovoltaic forecasting: A comparison of the most effective techniques. *Energies*, 12(9):1–15, 2019. ISSN 19961073. doi: 10.3390/en12091621.

[32] Tuan Anh Nguyen, Doina Bucur, Marco Aiello, and Kenji Tei. Applying time series analysis and neighbourhood voting in a decentralised approach for fault detection and classification in WSNs. In *ACM International Conference Proceeding Series*, 2013. ISBN 9781450324540. doi: 10.1145/2542050.2542080.

[33] Kristian Bonderup Pedersen and Kjeld Pedersen. Bond wire lift-off in IGBT modules due to thermomechanical induced stress. In *Proceedings - 2012 3rd IEEE International Symposium on Power Electronics for Distributed Generation Systems, PEDG 2012*, 2012. ISBN 9781467320238. doi: 10.1109/PEDG.2012.6254052.

[34] Kristian Bonderup Pedersen and Kjeld Pedersen. Dynamic Modeling Method of Electro-Thermo-Mechanical Degradation in IGBT Modules. *IEEE Transactions on Power Electronics*, 2016. ISSN 08858993. doi: 10.1109/TPEL.2015.2426013.

[35] Giovanni Petrone, Giovanni Spagnuolo, Remus Teodorescu, Mummadi Veerachary, and Massimo Vitelli. Reliability issues in photovoltaic power processing systems. *IEEE Transactions on Industrial Electronics*, 55(7):2569–2580, 2008. ISSN 02780046. doi: 10.1109/TIE.2008.924016.

[36] Alexander Phinikarides, George Makrides, Bastian Zinsser, Markus Schubert, and George E. Georghiou. Analysis of photovoltaic system performance time series: Seasonality and performance loss. *Renewable Energy*, 2015. ISSN 18790682. doi: 10.1016/j.renene.2014.11.091.

[37] Rakesh Ranjan. Largest solar park, 2019. URL `https://mercomindia.com/karnatakas-pavagada-solar-operational/`.

[38] Muhammad Qamar Raza, Mithulananthan Nadarajah, and Chandima Ekanayake. On recent advances in PV output power forecast. *Solar Energy*, 136(September 2019):125–144, 2016. ISSN 0038092X. doi: 10.1016/j.solener.2016.06.073.

[39] Gordon Reikard. Predicting solar radiation at high resolutions: A comparison of time series forecasts. *Solar Energy*, 2009. ISSN 0038092X. doi: 10.1016/j.solener.2008.08.007.

[40] Sourov Roy and Faisal Khan. Online Health Monitoring of Multiple MOSFETs in a Grid-Tied PV Inverter using Spread Spectrum Time Domain Reflectometry (SSTDR). pages 446–452, 2018.

[41] Camila P. Salomon, Wilson C. Santana, Erik L. Bonaldi, Levy E.L. De Oliveira, Jonas G. Borges Da Silva, Germano Lambert-Torres, Luiz E. Borges Da Silva, A. Pellicel, Marco A.A. Lopes, and Goncalo C. Figueiredo. A system for turbogenerator predictive maintenance based on Electrical Signature Analysis. In *Conf. Rec. - IEEE Instrum. Meas. Technol. Conf.*, 2015. ISBN 9781479961139. doi: 10.1109/I2MTC.2015.7151244.

[42] Abhishek B. Sharma, Leana Golubchik, and Ramesh Govindan. Sensor faults: Detection methods and prevalence in real-world datasets. *ACM Transactions on Sensor Networks*, 2010. ISSN 15504859. doi: 10.1145/1754414.1754419.

[43] Sobrina Sobri, Sam Koohi-Kamali, and Nasrudin Abd Rahim. Solar photovoltaic generation forecasting methods: A review. *Energy Conversion and Management*, 156(December 2017):459–497, 2018. ISSN 01968904. doi: 10.1016/j.enconman.2017.11.019.

[44] Katarzyna Stapor. Descriptive and Inferential Statistics. *Intelligent Systems Reference Library*, 176(1):63–131, 2020. ISSN 18684408. doi: 10.1007/978-3-030-45799-0{_}2.

[45] Suntech. Moerdijk Solar Project. 2018.

[46] Wan Syakirah, Wan Abdullah, Miszaina Osman, Mohd Zainal, and Abidin Ab. The Potential and Status of Renewable Energy. 2019.

[47] Joo Chuan Tong. *Statistical learning*. 2013. doi: 10.1007/978-1-4419-9863-7{_}941.

[48] Theresa L. Utlaut. Introduction to Time Series Analysis and Forecasting. *Journal of Quality Technology*, 2008. ISSN 0022-4065. doi: 10.1080/00224065.2008.11917751.

[49] Amit Kumar Yadav and S. S. Chandel. Solar radiation prediction using Artificial Neural Network techniques: A review, 2014. ISSN 13640321.

[50] Peng Zhang, Wenyuan Li, Sherwin Li, Yang Wang, and Weidong Xiao. Reliability assessment of photovoltaic power systems: Review of current status and future perspectives. *Applied Energy*, 104:822–833, 2013. ISSN 03062619. doi: 10.1016/j.apenergy.2012.12.010. URL `http://dx.doi.org/10.1016/j.apenergy.2012.12.010`.