THEFT DETECTION UNIT FOR PHOTO-VOLTAIC GENERATION IN SMART GRID NETWORKS

by

Nouf Ahmad Almadani

A Thesis presented to the Faculty of the
American University of Sharjah
College of Engineering
In Partial Fulfillment
of the Requirements
for the Degree of

Master of Science in
Electrical Engineering

Sharjah, United Arab Emirates

May 2020

Declaration of Authorship

I declare that this Thesis is my own work and, to the best of my knowledge and belief, it does not contain material published or written by a third party, except where permission has been obtained and/or appropriately cited through full and accurate referencing.

Signature: Nouf Ahmad Almadani

Date: April 27, 2020

Approval Signatures

We, the undersigned, approve the Master's Thesis of Nouf Ahmad Almadani

Thesis Title: Theft Detection Unit For Photo-Votaic Generation in Smart Grid

Date of Defense: 17-Apr-2019

Name, Title and Affiliation	Signature
Dr. Mostafa Shaaban Assistant Professor, Department of Electrical Engineering Thesis Advisor	
Dr. Usman Tariq Assistant Professor, Department of Electrical Engineering Thesis Co-Advisor	
Dr. Ahmed Osman Professor, Department of Electrical Engineering Thesis Committee Member	
Dr. Raafat Aburukba Assistant Professor, Department of Computer Science and Engineering Thesis Committee Member	
Dr. Nasser Qaddoumi Head Department of Electrical Engineering	
Dr. Lotfi Romdhane Associate Dean for Graduate Studies and Research College of Engineering	
Dr. Sirin Tekinay Dean College of Engineering	
Dr. Mohamed El-Tarhuni Vice Provost for Graduate Studies Office of Graduate Studies	

Acknowledgment

I would like to thank my Advisors Dr. Mostafa Shaaban and Dr. Usman Tariq for their continuous support and guidance throughout this research. It's been a hard time for me, I faced many problems, and without their help to steer me in the right direction I wouldn't be able to make it to the end. I am gratefully indebted to the valuable comments I received on this research whenever I nocked their doors.

Dedication

I dedicate my thesis work to my family, to my mother and father for the continuous support.

To my grandparents for the encouragement and the continuous prayers to finish this study.

And lastly, to my little siblings, the joy of my life

Abstract

While the increased connectivity of the power grid has allowed for the automation of its functionality, it has also led to a heightened vulnerability to cyber threats, putting the whole power system security at risk of energy theft through the manipulation of data. In addition, the introduction of the smart grid allows customers to have their own power-generating units, which are usually photovoltaic (PV) panels. With two-way communication under the smart grid paradigm, customers' local generation can be measured by smart meters and reported to the utility, which in turn pays customers for their generated electricity. Manipulating smart meters to report false generated electricity is a growing concern that can jeopardize a utility's revenues. Thus, the objective of this work is to design and build an intelligent theft detector unit for PV injection (TDUPV) that detects suspicious data flow from customers' solar smart meters to the back-end system within the utility. This topic contributes to the theft detection research community as it considers the injection of PV panels, which had not been considered in any previous research work. The detector is based on a regression tree model that utilizes weather information and customers' PV injections to predict the honesty of the injected power from customers' PV panels reported by the solar smart meters, assuming a data flow manipulated by cyberattacks. The mechanism of detection is based on the probability density function (PDF) of the error between the actual and predicted values. The performance of the TDUPV was evaluated by testing several case studies under different theft scenarios and shows the effectiveness of the proposed unit.

Keywords: ***Advanced Metering Infrastructure (AMI); Artificial Intelligence (AI); Cyber-attacks; Deep Learning (DL); Irradiance; Machine Learning (ML); Regression; Smart Grid***

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AMI	Advanced Metering Infrastructure
AI	Artificial Intelligence
FIT	Feed-In Tariff
ML	Machine Learning
NTL	Non-Technical Losses
RMT	Random Matrix Theory
SVM	Support Vector Machines
TDUPV	Theft Detection Unit for PV Injection

Chapter 1 . Introduction

This chapter intents to introduce the smart grid and distributed generation concepts. This chapter shall also address the challenges encountered in the field of smart grids and solar energy. The objective of this research shall then be presented. Lastly, we present the structure of this thesis.

1.1. Overview

According to the Federal Bureau of Investigation (FBI) and International Utilities Revenue Protection Association, energy theft cost utilities an estimated loss of around $6 billion in 2007, just in the U.S. [1]. Further, the world has started rushing toward automating each and every process in order to build a more reliable, intelligent, and efficient life in all aspects. Smart grids are becoming an essential innovation that leads to smarter cities. Smart grids keep growing and linking many smart sensors, consequently increasing the possibility of energy theft [1].

The term “smart grid” refers to the electricity grid, which is responsible for delivering electricity from the power plants to customers, including residential areas and industrial businesses. With the bidirectional communication platform and the embedded smart sensors such as smart meters, access points, and others, utilities can now ensure a more reliable, efficient, and economical power grid. Not only that, but the integration of distributed energy resources such as solar energy, wind energy, and many other renewable resources with the smart grid makes it environment-friendly. One of the main features of smart grids is self-healing, which allows the grid to heal itself by rerouting energy to feed faulty parts. This is done with the help of the embedded smart sensor, which provides real-time monitoring and control of the energy network, allowing the smart grid itself to repair its operations with minimal human intervention. Moreover, by using two-way communication to transmit consumption data to the utility and the dynamic energy pricing system, demand can be controlled in order to flatten the consumption curve during the peak hours, which results in lower operational costs.

Advanced metering infrastructure (AMI) is the backbone of smart grids. It consists mainly of smart energy meters with advanced communication capabilities, which form a network of information in the smart grid [2]. The communication is established through access points scattered throughout the area to constantly route the data from the meters all the way to the data concentrators and then to the back-end

system in the utility. Those meters can not only transmit the load profile but can also push access logs and several preprogrammed alarms that utilities want to monitor. AMI makes it possible to shift from the fixed tariff system in favor of a dynamic tariff system that seeks to improve consumption patterns and eliminates the peaks as much as is practical by incentivizing customers to shift their consumption, consequently making the grid more efficient.

As mentioned earlier, smart grids are capable of accommodating distributed energy resources (DERs) such as solar and wind, serving the required load with clean energy and, consequently, reducing carbon emissions. As a matter of fact, mounting PV panels on the roofs of houses and buildings has been gaining considerable momentum lately. This is due to the feed-in tariffs (FITs) policy that seeks to encourage people to produce green energy. FIT is referred to as clean energy cashback, where people get paid for the energy they produce and feed to the grid [3]. Some examples of FIT programs offered in various countries are illustrated in Table 1.1. An example of the energy exchange between customers with PV installations and the grid is illustrated in Figure 1.1.

Table 1.1: Feed-In Tariff rates for different countries.

SN	Country	FIT rate (currency/kWh)
1	United Kingdom	0.15-4.12 p/kWh [4]
2	India	0.15 USD/kWh[5]
3	Malaysia	0.16–0.34USD/kWh
4	Australia	0.17USD/kWh
5	Germany	0.450 Euro/kWh[6]
6	Italy	0.4Euro/kWh
7	Ontario-Canada	80.2c per kWh[7]
8	Japan	21 JPY/kWh[8]
9	China	0.42 RMB per kwh.[9]

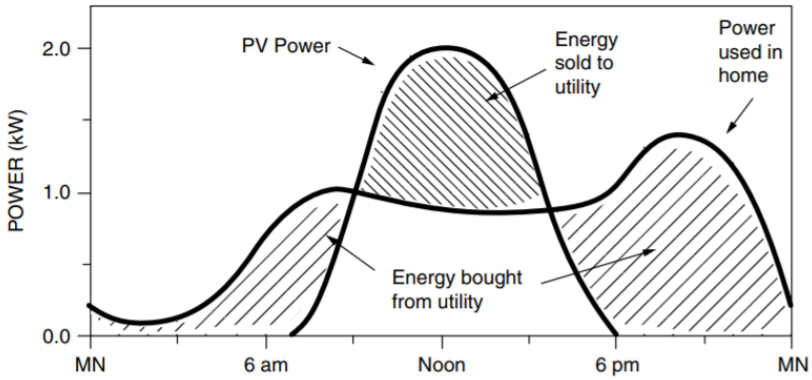

Figure 1.1: Grid-connected PV systems energy exchange [10].

Some countries, on the other hand, are implementing another approach called the Net Metering System, which is the case in the United Arab Emirates (UAE). In this paradigm, clients feed the excess of the generated solar energy to the grid and receive a reduction on the next bill and not cash [11]. An example of the net metering is shown in Figure 1.2, which requires one meter versus the FIT scheme, which requires two meters [10].

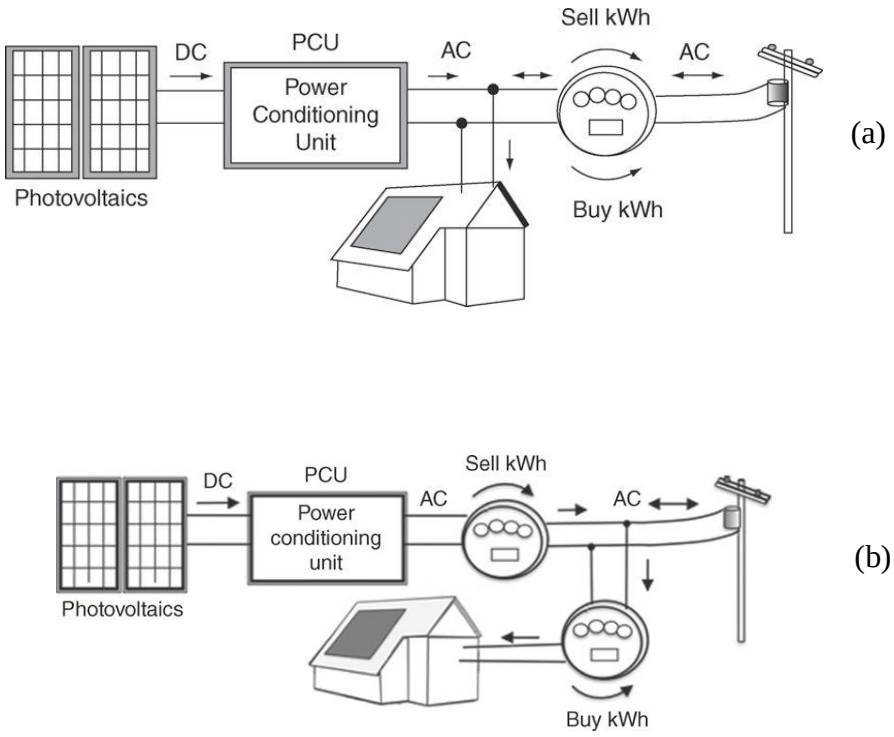

Figure 1.2: Grid-connected PV systems (a) Net metering scheme and (b) FIT scheme.

The mechanism of net metering accounting is based on an agreement between the producers and the utility that comes in the form of a contact, including prices. Some advantages of net metering policy are [12]:

- The billing is based on a net scheme over a long period.
- Promoting and encouraging the installation of renewable resources.
- Lower network losses.
- Lower peak load (flatter consumption curve).
- Single bidirectional electricity smart meter measures the net of production and consumption.
- It does not cost much.
- Easy to implement technically.
- Customers can guarantee the purchase of excess electricity.
- Price competitiveness.

Although smart grids together with integrated renewable DERs bring many remarkable benefits, they open the door to other risks. One of those risks is the ability to steal from the utility by exploiting the network and manipulating data through cyberattacks to reduce consumption or increase renewable energy injection. As the connectivity of the smart grid with different systems such as smart meters, electric vehicle (EV) chargers, and other renewable energy resources increases, the probability of network exploitation rapidly increases [13]. Hackers can penetrate the network and not only reduce their bills but also increase the amount of solar energy believed to have been injected into the grid, hence getting paid for energy they did not generate.

1.2. Thesis Objectives

The aim of this research is to design a theft detection unit for PV injections (TDUPV) into the smart grid. The unit will detect suspicious readings caused by cyberattacks by comparing the claimed PV injection with the forecasted PV injection, taking into account the type of PV panels installed. The TDUPV will be able to detect under the below assumed condition:

- Smart grid network being hacked to increase the injected power from a PV panel into the utility's grid.

1.3. Research Contribution

The main contribution of this thesis is to focus on theft incidents caused by cyberattacks to manipulate the injection of PV panels from the customer's premises to the utility grid using machine learning. Most of the previous work focuses solely on the power consumption reported by smart meters.

1.4. Thesis Organization

The rest of the thesis is organized as follows: Background and literature review are presented in chapter 2. Methodology and problem formulation are explained in chapter 3. Results and multiple case studies will be presented and discussed in chapter 4. Finally, the conclusion and future work are presented in chapter 5.

Chapter 2 . Background and Literature Review

This chapter forms the base upon which the TDUPV was developed. We introduce different probability distributions that were used for the theft detection mechanism, as well as introducing different machine learning algorithms that were used for the PV injection prediction. In addition, the previous related work in this field shall be presented.

2.1. Probability distribution

The theory of probability has been used to cope with the process of decision making in many AI systems divers field of study including engineering, science, management and many other fields [14]. Probability distribution is an essential tool that is widely used to measure the uncertainty of different aspects like weather forecasting [15].

Continuous random variables can be described by two types of density functions, probability density functions (PDFs) and Cumulative Density Functions (CDFs). The PDF is a statistical expression used to characterize the probability distribution of a continuous random variable. In other words, it is used to describe the relative likelihood for a certain random variable, which can be a set of observations, to take on a given value which is expressed in equation 2-1 [16]. Figure 2.1 shows a sample PDF of a Normal distribution. It can be clearly observed that the shape of PDFs for Normal distribution tends to have a peak. The location of the peak is expressed by the mean (μ) while the width of the graph is expressed by the variance (σ) [16].

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{2-1}$$

where

$f(x)$	The PDF of x
x	The random variable
a, b	Interval
$P(a \leq X \leq b)$	Probability that x lies between the interval [a,b]

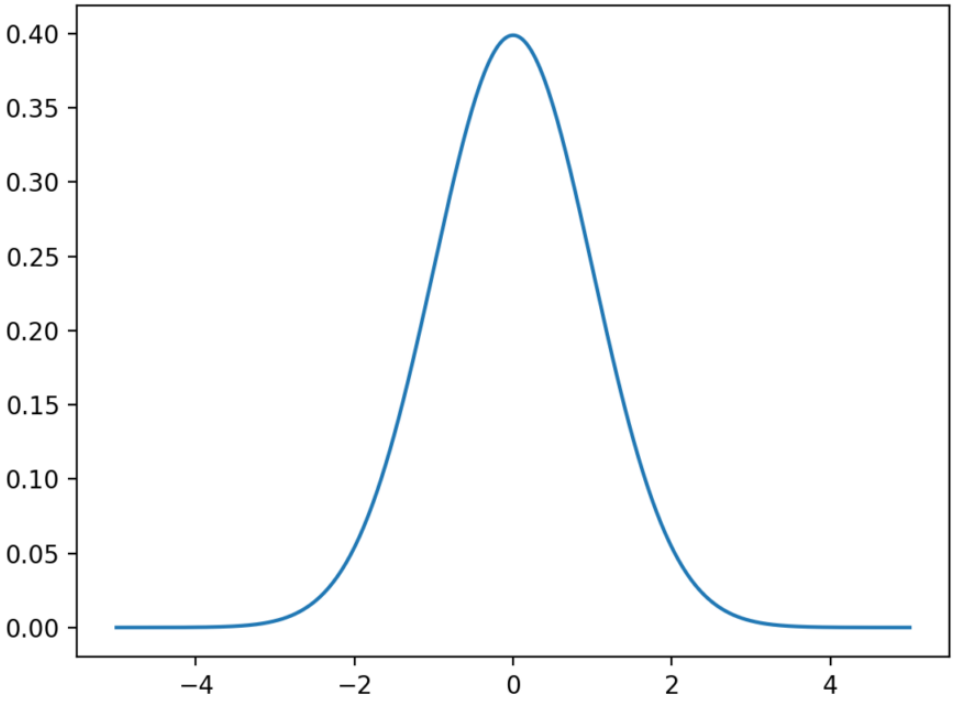

Figure 2.1: A sample PDF of a Normal distribution.

On the other hand, every distribution has a common characterization through its Cumulative Density Function (CDF) [15], which is the integral of the PDF. CDFs are used to find the probability that a certain variable lies exactly at or less than a certain value. Equation 2-2 shows the mathematical expression of CDFs [17]. A CDF typically varies from 0 to 1 to show the probability of occurrence [18]. Figure 2.2 shows the histogram of CDFs.

$$P(X \leq x) = \int_{-\infty}^{x} f(\mu) d\mu \tag{2-2}$$

where

$P(X \leq x)$ denotes the probability that X is less than or equal to x.

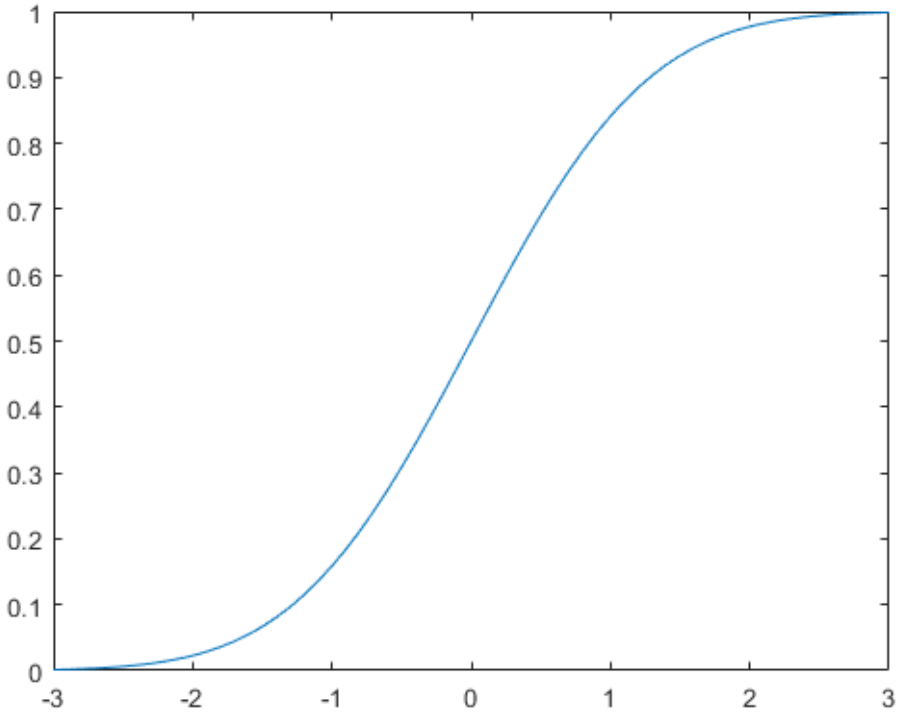

Figure 2.2: A sample CDF of a Normal distribution.

Moreover, density distributions are further classified based on the shape they follow. Beta distribution, Uniform distribution, Normal distribution, Lognormal distribution, Exponential distribution and Weibull distribution are briefly explained below.

2.1.1. Beta distribution. The beta distribution can be defined as a bounded continuous distribution. It is normally used to express an uncertainty in an observation that is between 0 and 1. Beta distribution has four parameters: x and y, which are positive values, as well as two arbitrary bounds, which are lower and upper. Other terminologies are also used like Pert distribution. The Beta distribution can follow multiple shapes for various values of $\alpha \; and \; \beta$ as shown in Figure 2.3, like bell-shaped unimodal, uniform, bimodal, exponentially increasing or exponentially decaying [19]. Equation 2-3 expresses the density function of beta distribution [20].

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{(\alpha-1)}\,(1-x)^{(\beta-1)} & a \le y \le b,\ \alpha \ge 0,\ \beta \ge 0 \\ 0 & otherwise \end{cases} \tag{2-3}$$

where

$\Gamma(z)$ is the Gamma function and it is expressed as:

$$\Gamma(z) = \int_0^\infty w^{z-1}\, e^{-w} dw \tag{2-4}$$

where

$f(x)$	PDF
w	Auxiliary variable
α, β	Shape parameters
a, b	Interval (lower and upper bounds respectively)

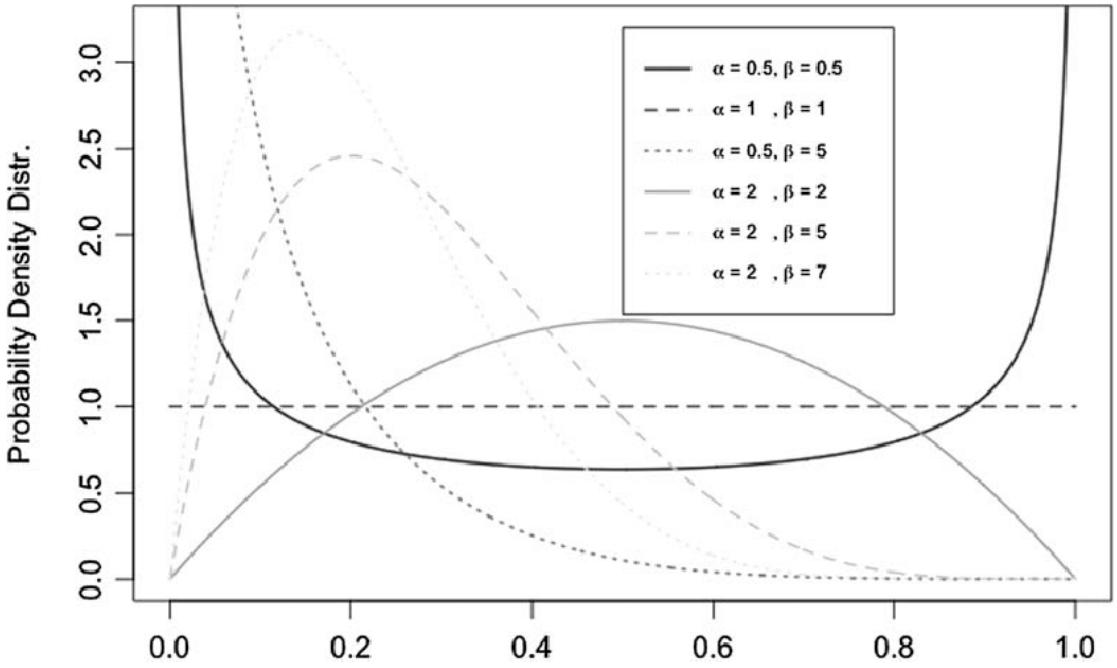

Figure 2.3: Beta PDF plot for different α and β [21].

2.1.2. Uniform distribution. A random variable is said to follow a uniform probability distribution when any two sub-intervals of equal length between a defined interval from ***a*** to ***b*** are equality likely. Equation 2-5 is used to mathematically express the probability of a certain random variable X. Figure 2.4 shows the probability density function PDF of a uniform probability distribution since any value of the random variable between the defined interval [a,b] is equally likely, the graph is a rectangle. The Figure 2.4 below corresponds to the PDF when the random variable x is between

0 and 60, as mentioned earlier, the area under the curve must add to 1, which is one of PDF conditions, therefore, the height of the curve is 1 divided by the width [22].

$$f(x) = \frac{1}{b-a} \quad a \leq x \leq b \tag{2-5}$$

where

a, b Interval

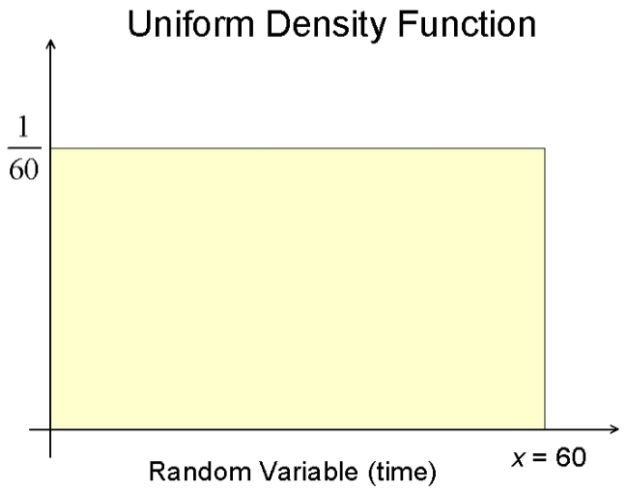

Figure 2.4: Uniform PDF.

2.1.3. Normal distribution. A continuous random variable x is said to be normally distributed or has a normal distribution if its corresponding histogram frequency has the shape of a normal curve like in Figure 2.5 and 2.6. The Figure also demonstrates the role of two important parameters of PDFs, which are the mean μ and the standard deviation σ and how they contribute to the shape of any distribution. The mean μ represents the balancing point of the graph with respect to the x-axis while the standard deviation plays with the height of the peak. Mathematically, the normal distribution can be expressed using equation 2-6 [22].

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad for \ -\infty<x<\infty \tag{2-6}$$

where

μ Mean

σ Standard deviation

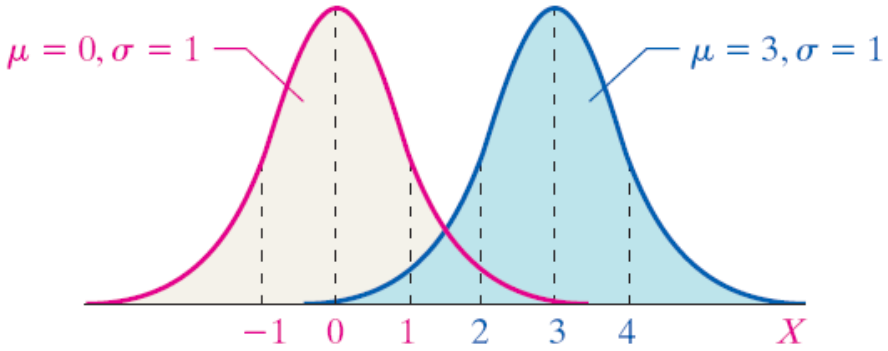

Figure 2.5: Normal distribution PDF for different μ values.

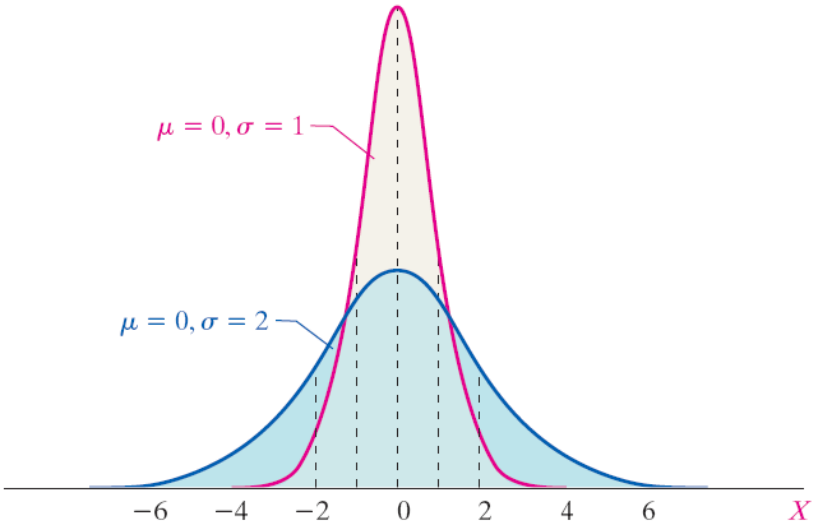

Figure 2.6: Normal distribution PDF for different σ values [22].

2.1.4. Log-normal distribution. A random variable *x* is said to be log-normally distributed when $\boldsymbol{Y = ln(X)}$ is normally distributed with the natural logarithm. Equation 2-7 expresses the general formula for the probability density function PDF of the log-normal distribution. The log-normal distribution is mathematically expressed in equation 2-8. The log-normal distribution when plotted, tend to look like Figure 2.7, it can be clearly observed that as sigma approaches zero, the PDF of the lognormal distribution becomes more like the PDF of the normal distribution [23].

$$f(x) = \frac{(e^{-((\frac{In(x-\theta)}{m})})^2}{(x-\theta)\sigma\sqrt{2\pi}} \quad x > \theta; m, \sigma > 0 \tag{2-7}$$

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{In\,x-\mu}{\sigma}\right)^2} \tag{2-8}$$

where

μ	Mean of the log of the distribution
σ	Standard deviation or shape parameter
θ	Location parameter
m	Scale parameter

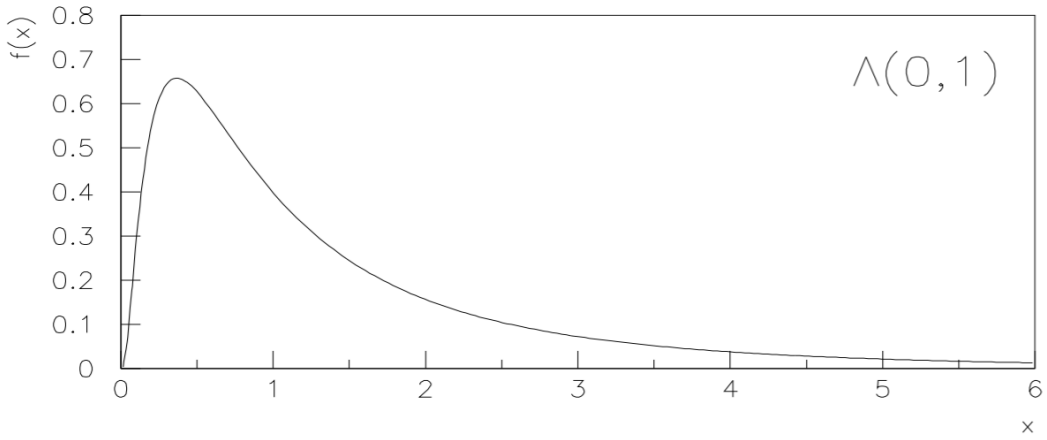

Figure 2.7: Log-normal distribution PDF for μ=0 and σ=1 [24].

2.1.5. Exponential distribution. Exponential distributions are used widely in engineering and science disciplines. The exponential distribution expression describing the PDF of a random variable X=*x* is shown in (2-9) [25] and has a shape similar to Figure 2.8.

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & otherwise \end{cases} \tag{2-9}$$

where

λ	Rate parameter

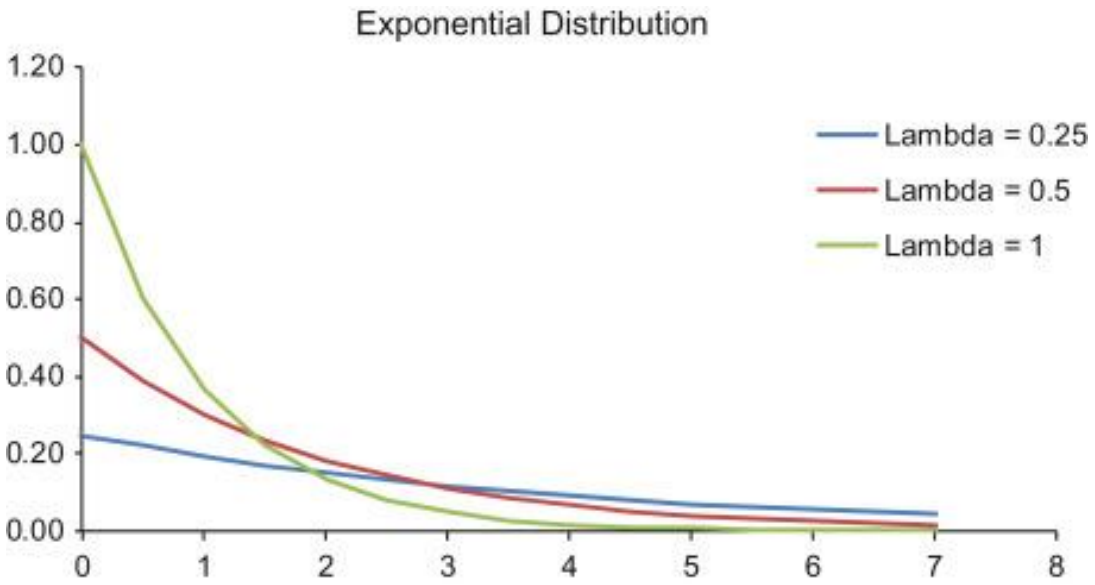

Figure 2.8: Exponential distribution PDF for different values of λ [26].

2.1.6. Weibull distribution. The Weibull distribution is similar to the exponential distribution; however, it has an additional parameter that introduces the flexibility. A random variable V=*v* is said to have a Weibull distribution only if its PDF can be expressed using equation 2-10 and has a shape similar to the curves in Figure 2.9. Different curves can be plotted with different values of the shape parameter (*k*) which is also knows as the Weibull slope, as well as different values of scale parameter (*c*) as illustrated in the Figure.

$$f(v) = \frac{k}{c}\left(\frac{x}{c}\right)^{k-1} exp\left(-\left(\frac{x}{c}\right)^{k}\right) \tag{2-10}$$

where

k	Shape parameter
c	Scale parameter

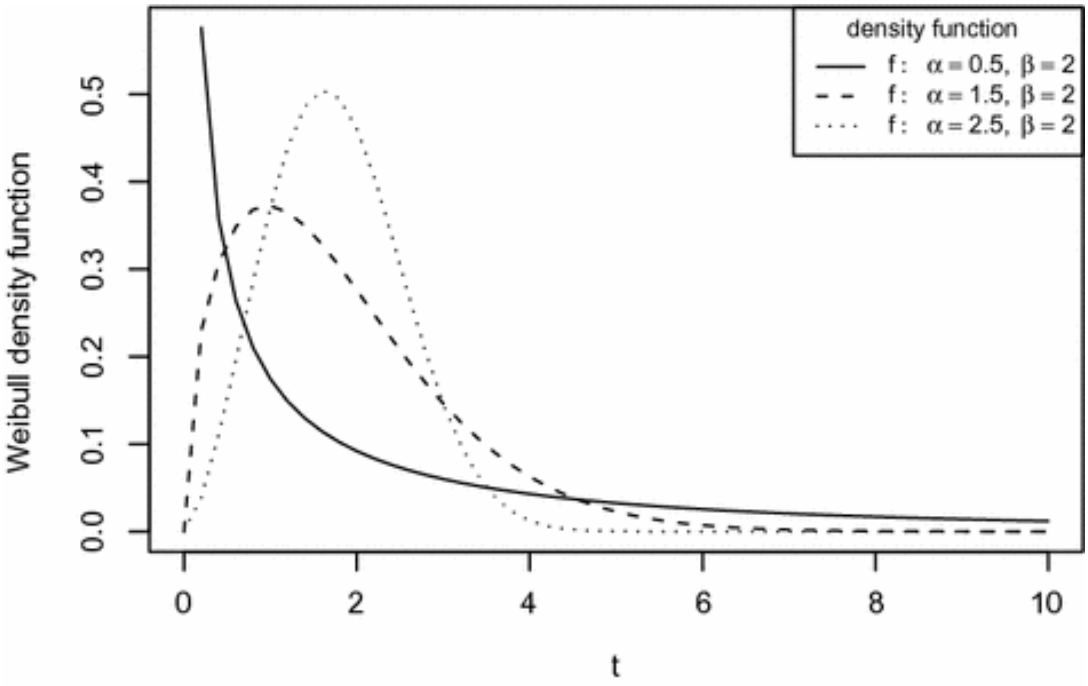

Figure 2.9: Weibull distribution PDF for different values of *c* and *k* [27].

2.2. The revolution of Machine Learning (ML)

Machine learning (ML) is an integral part of Artificial Intelligence (AI) [28]. It has gained incredible momentum in recent years. The basic idea here is to learn the implicit rules from past data to help predict the future occurrences. Machine learning has become more popular than ever. Many applications are in development today that use different machine learning algorithms, such as self-driving cars, content recommendations and fraud detection. A subset of machine learning is deep learning, which has gained a lot of traction in media in recent years. In the following, we will briefly review it.

2.2.1. Deep learning. The relationship between AI, machine learning and Deep Learning can be best described as a Figure in Figure 2.10. Deep Learning is essentially a subset of Machine Learning (ML). It has lately received a lot of attention as it can solve many sophisticated real-world problems with state-of-the-art performance. It takes advantage of two key aspects, lots of data and faster and parallel computational hardware.

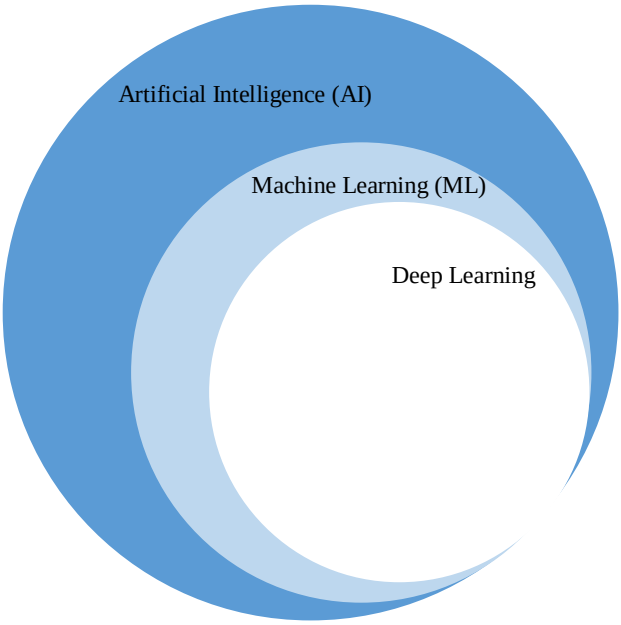

Figure 2.10: Artificial Intelligence, Machine Learning and Deep Learning.

It is essentially a layered approach to some algorithms in machine learning where the parameters in successive layers are jointly trained through various optimization algorithms. Most of the deep learning models are variants of neural networks. Neural networks were invented in the beginning to model human brain but advanced deep learning algorithms (layered neural networks with or without sparse connectivities or skipped connections etc.) are not built with the human brain in mind. The layered structure of deep learning models and joint training of parameters across layers gives these models greater flexibility compared to traditional machine learning algorithms. The real power of deep learning algorithms is unleashed when one has lots of data to learn from. The more useful data you feed, the more accurate the conclusions would be. In fact, it's now much easier to build some complex models due to the availability of large datasets and computational power.

Machine learning problems can be regarded and classified into three categories based on the type of data being processed:

2.2.1.1. ***Supervised machine learning.*** Problems dealing with labeled data fall under this category. In this case, the algorithm maps the input with the output to best approximate the function which apparently will look like $y = f(x)$. Some of the algorithms that come under the supervised machine learning are random forests, logistic regression, support vector machines, Naive Bayes and artificial neural networks

(ANNs). Supervised machine learning, problems can be further classified into two groups:

1. Classification: when the input is mapped to discrete output points.
2. Regression: when the input is mapped to a continuous output.

2.2.1.2. *Unsupervised machine learning.* In some problems, only the input data is available, hence, the algorithm goal is to find a hidden structure in unlabeled data. Under the unsupervised machine learning, problems can be further classified into two groups:

1. Clustering: This aims to group up the data based on specific things like behavior of Customers.
2. Association: This aims to find out rules and associate data based on those rules [29].

2.3. Regression Analysis

Regression is one of the machine learning approaches that lies under the supervised machine learning umbrella. It is basically used to find the mathematical relationship between one or multiple independent variables or sometimes called predictors and a single dependent variable called the response variable [30]. As a matter of fact, there are several types of regression models and the selection of the best model that fits a certain problem depends on the type of data available. Regression models can be classified into two groups based on the type of the independent variables or the predictors. Variables can be continuous or nonnegative integers count data [31].

In our case study and since we are predicting how much solar energy each customer should produce from their PV panels, we shall consider the main factors affecting the solar energy production. In this research, we are taking into account three independent variables or predictors: temperature, solar irradiance and time. All the three predictors are continuous. Therefore, we will start our analysis with a multiple linear regression model.

2.3.1. Multiple linear regression. One of the most common and straightforward approaches that is advisable to start with is the linear regression or Ordinary Least Squares. Although it's called linear, linear regression can model curves

using polynomials by introducing quadratic and cubic terms of the predictor variables [31].

2.3.2. Multiple linear regression. Classification and Regression Trees (CART). CART is also one of the most useful and simple algorithms that are widely used in Machine Learning [32]. As the name indicates, a tree-like models are used for decision-making to visually represent decisions. This algorithm can be utilized to solve both classification and regression problems and it can handle both categorical and numerical data [33]. Regression trees are used for prediction problems, when the response variable or target variable is continuous, while classification trees are used for classification problems when the target is to be classified into several classes, for instance, yes and no or male and female.

This algorithm works by breaking down a specific dataset into smaller subsets in an iterative process called Binary Recursive Partitioning (BRP), while at the same time an associated tree-structured vector is gradually developed [34]. The structure of decision trees consists of three main parts: The first part if the Root node; which is the starting point of the tree. The second part is the Leaf node; which contains a decision on the numerical target. The third part is the branch; which is simply, an arrow connecting nodes to show the direction from the question to the answer. From each node, two or more nodes can be extended depending on the outcome of the test performed at each branch as shown in Figure 2.11 [35].

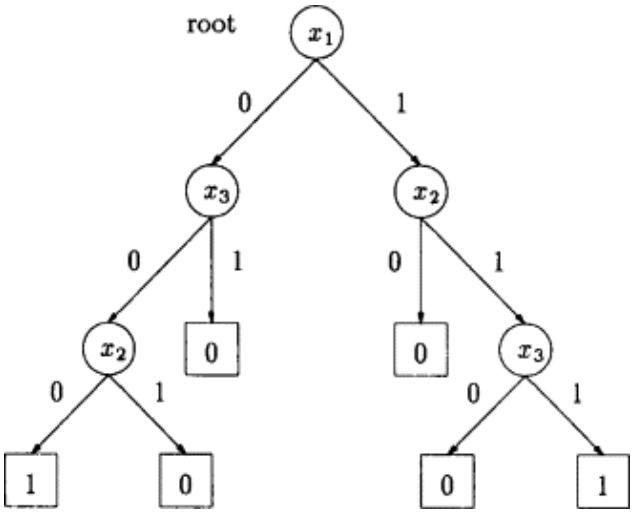

Figure 2.11: Regression trees structure [36].

2.3.3. Support vector machine (SVM). Support vector machine or simply abbreviated as SVM has been found to outperform many traditional machine learning algorithms. This is due to its significant accuracy that requires less computation power compared to other algorithms. SVM algorithms are used in classification-kind of problems, however, it can work for regression problems as well. To better understand SVM algorithms, let us first look into the linear classifier.

The linear classifier is a classifier that uses a hyperplane to separate or partition a set of data into N-groups corresponding to N-features or classes. Figure 2.12 shows a two-class (blue and orange) illustration of a linear classifier. This line, also known as hyperplane, can be defined as a boundary between the different groups that distinctly classifies the data points. The objective of SVM algorithms is to maximize the margin between the different classes or groups, which in return provides a sort of reinforcement such that future data points can be attributed easily to the classes with less error. It is good to mention that when the number of classes exceeds three, it becomes difficult to visualize the hyperplane [37].

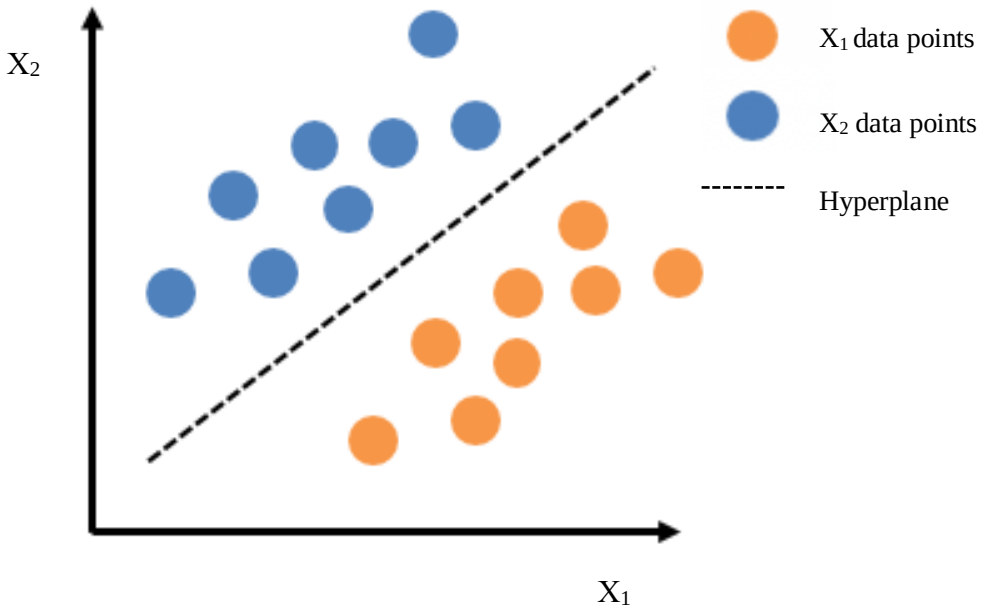

Figure 2.12: A two-class linear classifier [37].

The linear classifier is used only when the boundary between different classes or groups is a straight line; however, most of the classification problems are not that

simple. Hence, SVM algorithms are particularly used to handle such complex classifications. The basic idea behind SVM algorithms is shown in Figure 2.13, where the data is non-uniformly arranged with a nonlinear boundary. SVM algorithms work by making sure the original data is rearranged such that a straight line can fit between the classes. This rearrangement is referred to as Mapping or Transformation and its executed using a set of mathematical functions, known as kernels [38]. Equation 2-10 expresses a separation function that is used in SVM algorithms [37].

$$f(x) = \sum_{x_j \in S} \alpha_j \, y_j K(x_j, x) + b \qquad \textit{(2-10)}$$

where

x_j	Support vectors
x	Unknown variable
$K(x_j, x)$	Kernel function
b	Offset value
S	Set of support vectors SVs
α_j	Corresponding coefficients
y_j	Class label

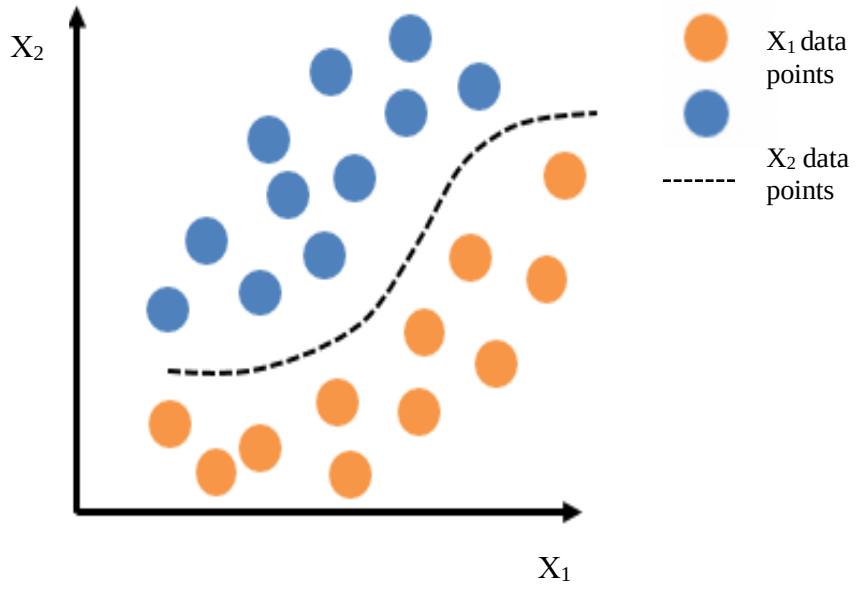

Figure 2.13: A two-class nonlinear classifier [37].

2.4. Related Work

Electricity theft detection has been previously investigated by researchers. However, the scope of the previous research work did not take into account solar energy injection, which is a major vulnerability through which hackers can penetrate the network and change the injected PV data. A literature review of the key efforts done previously in theft detection is given below:

In [39], the researcher used Artificial Neural Network (ANN) method to model a detector that detects suspicious load profiles of customers assuming several scenarios of cyber-attacks like: Assuming the attacker will reduce the consumption by random number for each slot of time or will reduce the consumption with a fixed number for a specific period of time and other scenarios.

In [40], an electricity theft detector based on Random Matrix Theory (RMT) along with the cost-effective Distributed Meter Data Management DMDM solution was developed. The researcher used real power data and simulated power data to validate and test the detector under several theft scenarios and the results were satisfactory.

In [41], the researcher proposed to solve electricity theft in smart grid networks using a linear system of equations (LSE) for the customers' "honesty coefficients", which is equal to 1 when the customer is honest and larger than 1 when the customer is

dishonest. The proposed model is based on the privacy-preserving electricity theft detection in smart grids using LU decomposing; which is called LUD/LUPD algorithms.

The authors in [42] have used some classification and clustering techniques to find the probability of energy theft were suspicious clients are identified by monitoring abnormalities in consumption patterns. A new algorithm was introduced to predict the normal and suspicious consumption, which is called consumption pattern-based electricity theft detection CPBETD algorithms along with SVM algorithm, the proposed model used silhouette plots for the analysis of the non-technical losses dataset that was extracted from distribution transformer meters.

In [43], non-technical losses (NTL) problem was addressed, were a fraud detection model (FDM) based on support vector machine SVM was developed with a purpose of extracting suspicious customers for inspection onsite based on an irregular consumption patterns. The used approach, in addition, provides data mining method, which uses historical customer consumption data to extract features. Moreover, this approach exposes abnormal behavior, which is highly correlated with non-technical losses NTL using load profile data of different customers and other attributes. After then, customers are classified to shortlist potential abnormal data for onsite inspection. The model was tested for consumption data of three towns within west Malaysia.

The work presented in [44] analyzes the NTL due to cyberattacks, that aims to manipulate the reported energy usage data to reduce the bill. A detection technique based on partially observable Markov decision process (POMDP) and Bollinger bands was used along with an adaptive dynamic programming technique to improve the efficiency. The proposed model can detect 92.55% of the malicious data.

In [45], an energy theft detection scheme is proposed using energy privacy preservation in the smart grid network. In addition, combined convolutional neural networks (CNN) technique was used for the detection of abnormal measurements of the smart meters data. The performance of the proposed approached was studied by observing the energy theft behavior from a user group perspective within a specific location taking into account the time of the data.

The work presented in [46] used Decision Trees (DT) and Support Vector Machine (SVM) to build a comprehensive top-down scheme. The proposed algorithm can detect, as well as locate energy theft in real-time in both power transmission level and power distribution level. The Decision Tree algorithm was used to process the input data, which are the number of persons, the season, the time slot and the temperature. Using those data, the DT algorithm calculates the predicted electricity consumption of customers and during a particular time slot. The output of the DT algorithms, along with other features are fed to the SVM classifier for training. Therefore, the proposed scheme can be regarded as a two-level data processing and analysis approach.

In [47], a novel energy theft detector is proposed named NFD (NTL Fraud Detection). The NFD is based on Lagrange polynomial interpolation to generate a polynomial for each smart meter using a small set of data to model the behavior of an adversary. By comparing the polynomials of tampered and normal smart meters, it can detect a tampered smart meter. In addition, mathematical models are built for each adversary which in return, helps in the detection mechanism.

In [48], a false data detection system was developed by integrating two techniques. Those techniques are tailored to fit different types of attacks. By adopting anomaly-based detection, strong attacks like injection of large amounts of data in a short time can be detected. The anomaly detection mechanism was integrated with a watermarking-based detection scheme to prevent attacks that involve subtle manipulation of the measurement data.

None of the previous work focused on the injection of PV panels to the grid, which is the target of this research.

Chapter 3 . Proposed Research and Methodology

In this chapter, the problem of electricity theft in smart grid networks shall be stated and formulated. In addition, the proposed design of the detector shall be presented.

3.1. Problem Statement

In order to predict the amount of electricity that can be generated from the solar energy, we should consider all the possible known factors that can affect the electricity generation wither by reducing it or increasing it. In this study, we are assuming a residential area where the geographical location of the customer is fixed, hence, the geographical location won't have a significant impact on the generated energy. On the other hand, temperature, solar irradiance and time of the day are the main factors that are significantly affecting the amount of generated electricity. Other variable factors, such as the characteristics of individual solar panel are not considered in this research as they cannot be measured. Therefore, in this research, we shall consider three independent variables of predictors to predict a single dependant variable or response or target variable.

3.2. Proposed Methodology

The methodology shown in Figure 3.1 was used in this research to achieve the goal and design the Theft Detector Unit of PV injections (TDUPV).

Since machine learning is utilized to build the TDUPV, the first and vital step was to gather all the required data upon which the model will learn the behaviour of the PV panels and the energy generation patterns. Then, the gathered data was used to train not only one model, but several types of regression models. After training different set of models, performance and the accuracy of the tested models have been studies and compared. This step was vital to make sure that the most suitable model for the kind of problem we are studying is used. However, all the trained models gave unrealistic results. Therefore, it was required to enhance the models and retrain them again. After doing so, the theft detection mechanism was developed and the criterion of theft was set. Finally, the TDUPV was tested for its accuracy throughout different stealing scenarios.

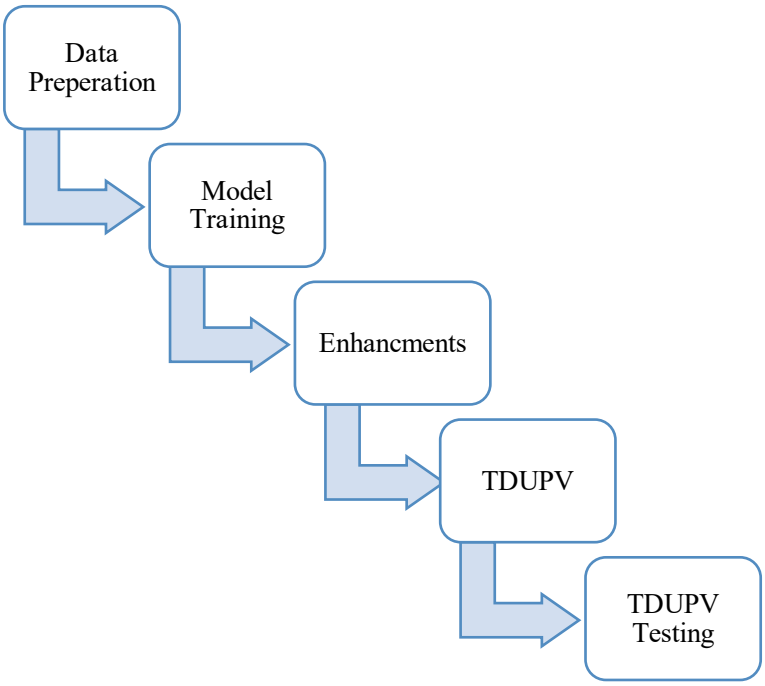

Figure 3.1: Proposed research methodology.

3.2.1. Data preparation. The first stage of this research was to prepare the data to be able to feed it the regression learner tool in MATLAB. In other words, the regression learner tool requires the data to be arranged in a matrix that starts with all the predictors and ends with the response variable in the last column. It is good to mention that since we are considering multiple types of PV panels with different specifications, the data should to be normalized to avoid any sort of bias and produce a fair prediction. As mentioned earlier, we will be using three input variables as the predictors and one output variable as the response or the target. The three predictors are time, solar irradiance and temperature. Solar irradiance data are historical hourly data from Canada. The temperature data are also real hourly data from Toronto, Canada that was extracted from [49]. In this research, we study the behaviour of 400 customers assuming that each customer will not only mount different number of panels based on their roof size, but also the type of panel will defer from one customer to another. In this research, we considered 11 different types of PV panels available in the market [50] and [51]. These types are randomly assigned to each one of the 400 customers. The time step is assumed to be hourly and historical data for 6 years are available. As a result, and since we are predicting the hourly output power, the total number of available observations for each customer is $8760 \times 6 = 52560$ observations.

Furthermore, to get the output power of each customer, we first calculate the hourly output power for each type separately for one year, then assigned each customer to one of the 11 types and assign a random number of mounted PV panels using uniformly distributed pseudorandom integers.

Using the data sheet characteristics of the 11 PV panels that are shown in Table 3-1 and the relations (3-1 to 3-5) [52], [53], the output power of each type of the PV panels is calculated for each of the 52560 observations.

$$T_{cell} = T_A + \frac{S\ (T_{NOCT} - 20)}{0.8\ \ kW/m^2} \tag{3-1}$$

$$I_{PV} = S\ \left(I_{sc}\left(1 + K_i\ (T_{cell} - 25)\right)\right) \tag{3-2}$$

$$V_{PV} = V_{oc}\ (1 + K_v\ (T_{cell} - 25)) \tag{3-3}$$

$$FF = \frac{V_{MPP}\ I_{MPP}}{V_{oc}\ I_{sc}} \tag{3-4}$$

$$P_{PV} = FF\ \ V_{PV}\ \ I_{PV} \tag{3-5}$$

where

T_{cell}	Cell temperature in ^oC
T_A	Ambient temperature in ^oC
S	Solar irradiation in kW/m^2
T_{NOCT}	Nominal operating cell temperature in ^oC
I_{PV} and V_{PV}	Current and voltage of the PV module
K_i and K_v	Current and voltage temperature coefficients
V_{MPP} and I_{MPP}	Maximum voltage and maximum current of the panel
V_{oc} and I_{sc}	Open circuit voltage and short circuit current
FF	Fill factor
P_{PV}	Output power of the panel

Table 3.1: Characteristics of the 11 PV panels.

PV panel type	Max Power point (W)	NOCT (^oC)	I_{max} (A)	V_{max} (V)	V_{OC} (V)	I_{SC} (A)
Type 1	435	45	5.97	72.9	85.6	6.43
Type 2	245	46	8.11	30.2	37.8	8.63
Type 3	87.5	45	1.78	49.2	61	1.98
Type 4	230	47	6	40.2	50.7	6.7
Type 5	135	45	2.88	47	61.3	3.41
Type 6	240	47	4.86	49.38	59.23	5.44
Type 7	245	47	4.95	49.51	59.45	5.54
Type 8	250	47	5.01	49.91	59.92	5.61
Type 9	255	47	5.09	50.11	60.36	5.70
Type 10	260	47	5.17	50.30	60.36	5.79
Type 11	265	47	5.25	50.48	60.60	5.88

The overall inputs and outputs of the data preparation stage are illustrated in Figure 3.2.

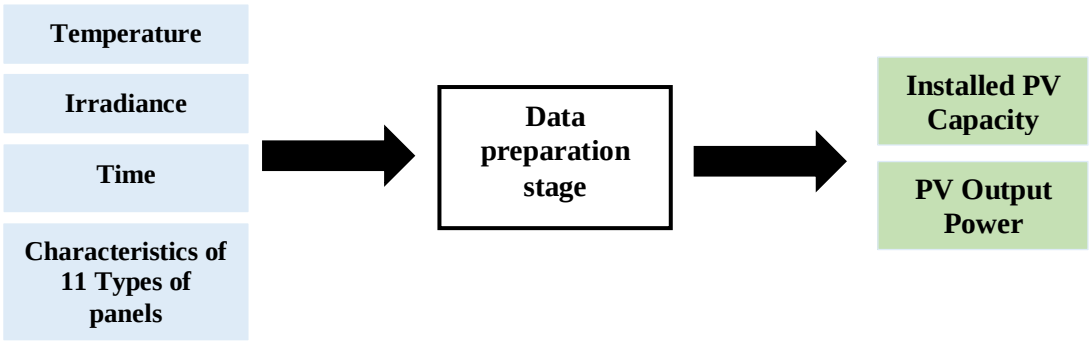

Figure 3.2: Output power and installed capacity using MATLAB.

Since we have different types of panels, different specifications and efficiencies, the training data will not be useful in determining the dishonest customers; therefore, the output power must be normalized with respect to the installed capacity of each customer. Figure 3.3 shows the output power waveform of type 1 (orange) and type 2

(blue) of PV panels before normalizing, whereas Figure 3.4 and 3.5 shows the normalized output power of two different panels.

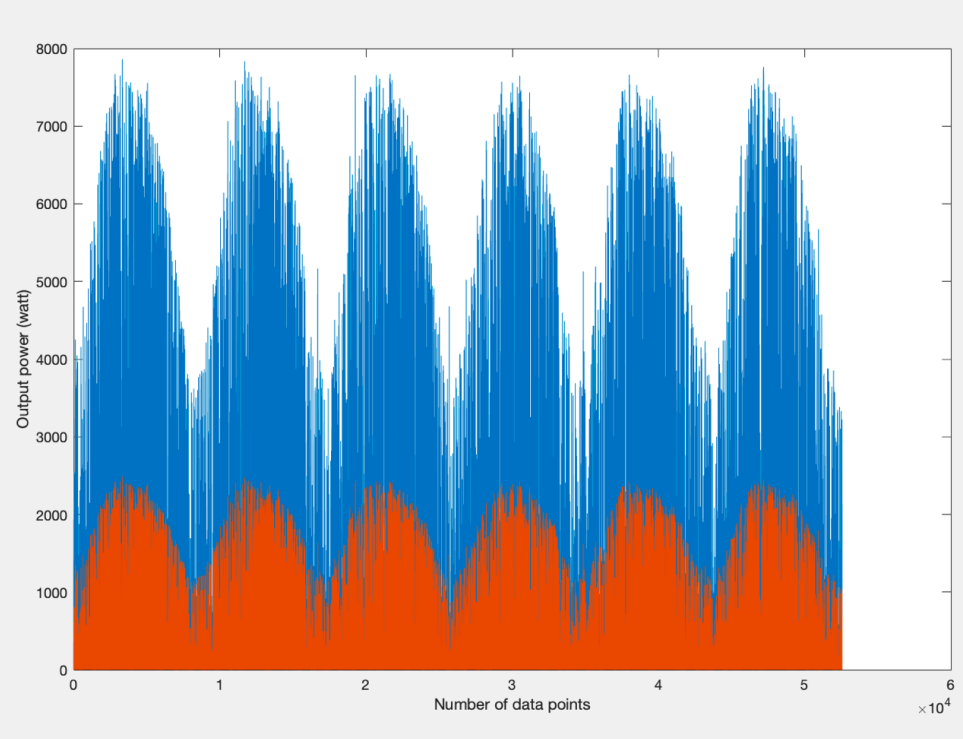

Figure 3.3: Output power before normalizing.

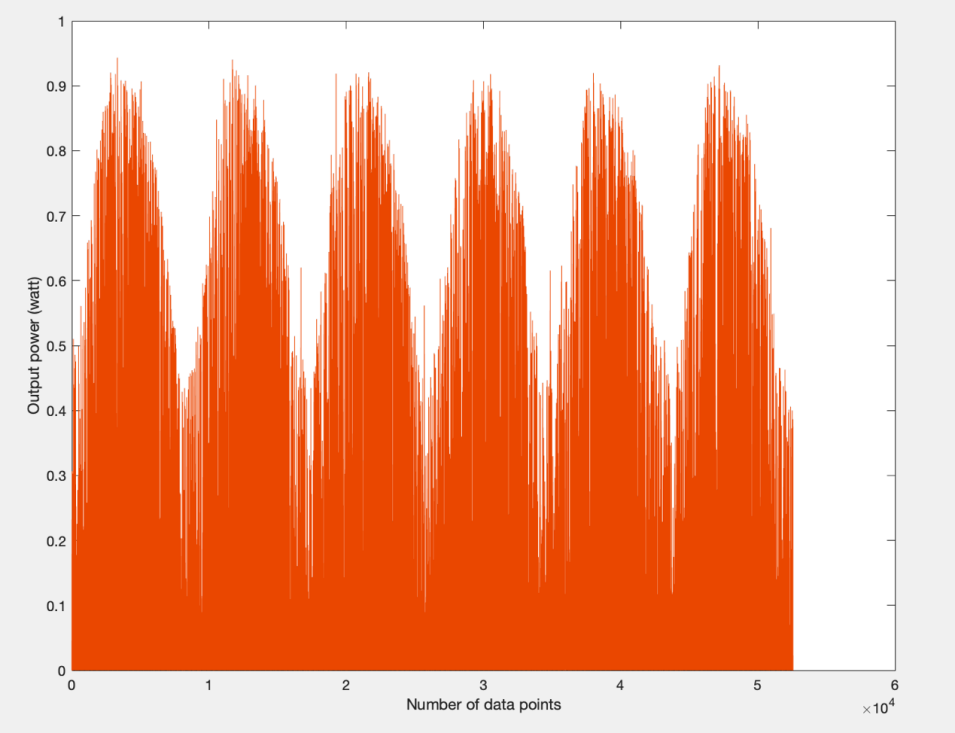

Figure 3.4: The normalized output power of type 1 of PV panels.

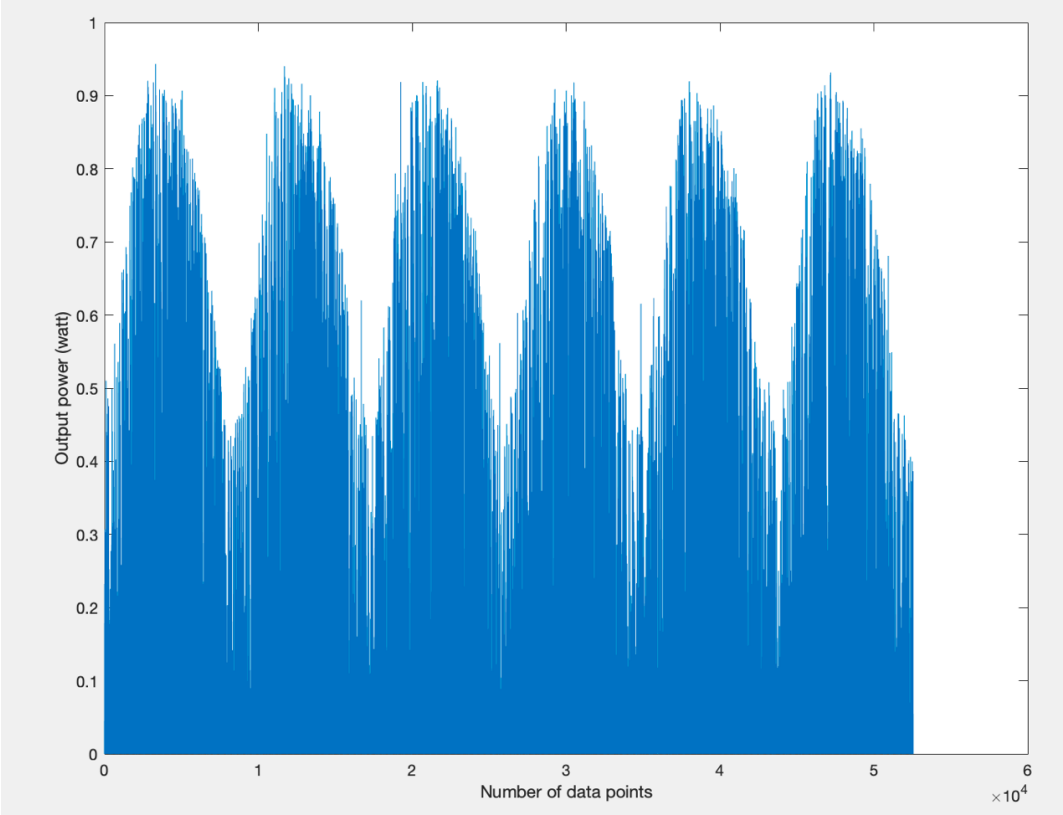

Figure 3.5: The normalized output power of type 2 of PV panels.

3.2.2. Model training. To be able to predict the outcome of the installed solar panels for each customer, we will need to train the model with historical data of PV generation. In this research, historical data of the irradiance and temperature in an hourly basis was used to generate virtual historical data of the output power for a group of customers given the installed capacity for each customer. Bear in mind that each customer has different type and number of panels.

A portion of the known predictors and the known response or target were supplied to train different models. Then, the generated model can be used to predict the responses of a new data set as shown in Figure 3.6. The validation error of each model will be used to choose the best regression model type [54]. The different models that are tested include regression trees, linear regression models, SVM and ensembles of regression trees. The model that will give the least error will be selected to build the detector.

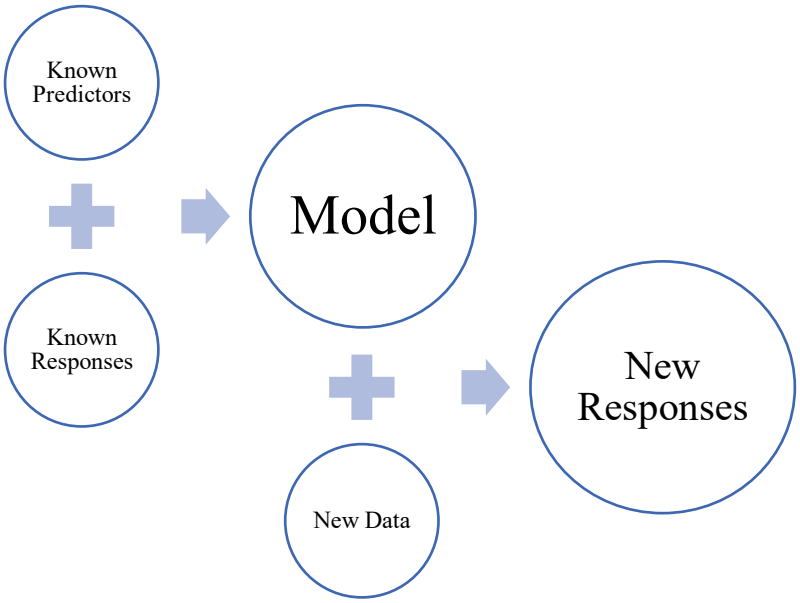

Figure 3.6: Regression learner tool process in MATLAB.

3.2.2.1. ***K-fold cross-validation technique.*** K-fold cross-validation is one of the statistical techniques used to evaluate and estimate the misclassification error in machine learning models [55]. This technique is widely used for its simplicity and for the fact that it can prevent overfitting since not all the data is used to build the model. The way this method works is by partitioning the whole data set intro K folds and use K-1 folds to train the model while a single fold is used to test the model. This procedure is repeated multiple times with different folds to ensure that each fold has been used exactly once to validate and test the model. As a consequence, the average error gives a good indicator of the performance of the model. It's good to bear in mind that the number of folds has a significant impact on the computational time of the algorithm. Therefore, more folds can be inconvenient [56]. In this research, 5-fold cross validation method was used, in another words, the data set has been divided into 5 equal folds randomly and the average error was taken.

3.2.3. Model enhancement. After training and testing the model, it was noticed that most of the predictions are exactly matching the actual values, in other words, the distribution was noticeably biased toward zero and as a result, the prediction was not accurate. In fact, the reason behind that bias is the data taken during the night time when the solar irradiance is zero and that actually makes sense as the prediction will always match the actual values resulting in a zero error for the night data points. In

addition to that, reporting an injection during the night time can easily be detected which is something considered rare to happen by any customer. Therefore, all the zero data points were curtailed from the original data set and the model was trained and tested all over again. The new model gave satisfactory results with a reasonable distribution histogram with no bias.

3.2.4. Theft detection unit (TDU). The theft detection mechanism depends on the probability density function (PDF) of the error between the predicted output power and the actual power. The idea is to generate a distribution of the error that can be used to compare between the actual values of the power and the predicted power generated from the model by defining a limit or threshold that indicates the acceptance range. For example, if the customer data was found to be laying after x number of standard deviations; which is equivalent to a very small percentage of occurrence, in other words, unlikely to happen, then the data that was reported by the customer might be incorrect. It's good to mention that we are only focusing on the positive probability which indicates that the customer is reporting an injection more than what the panels are supposed to generate.

3.2.4.1. ***Error distribution fitting***. The goal at this stage is to fit the extracted error from the regression learner tool with a proper PDF. The error between the predictor output and the virtual generated output power for the 400 customers was extracted from the regression learner model to generate the PDF of the error for each of the mentioned types above. Choosing the best PDF will be determined based on the maximum likelihood of each type.

Maximum likelihood (ML) estimation is widely used to estimate information measures. It is a method used to find the set of parameters $\hat{\theta}$ that maximizes the probability of getting the data we observed. To mathematically construct the likelihood function, the product of all individual likelihoods of each corresponding point x_i within a dataset is calculated using equation 3-6.

$$F(\hat{\theta}) = \prod_{i=1}^{N} P(x_i|\theta) \tag{3-6}$$

The ML estimation can be described as:

$$\hat{\theta} = arg\,\max_{\theta} \prod_{i=1}^{I} P(x_i|\theta) \tag{3-7}$$

The derivative of equation 3-16 with respect to each parameter in θ can be equated to zero to find the parameters at which ML occurs; however, due to the complexity of solving the products, the Log likelihood is used instead as in equation 3-8. Figure 3.7 illustrates different histograms of different distributions likelihood including lognormal, Cauchy, Weibull and Rayleigh.

$$\hat{\theta} = arg\,\max_{\theta} \sum_{i=1}^{I} log\,[P(x_i|\theta)] \tag{3-8}$$

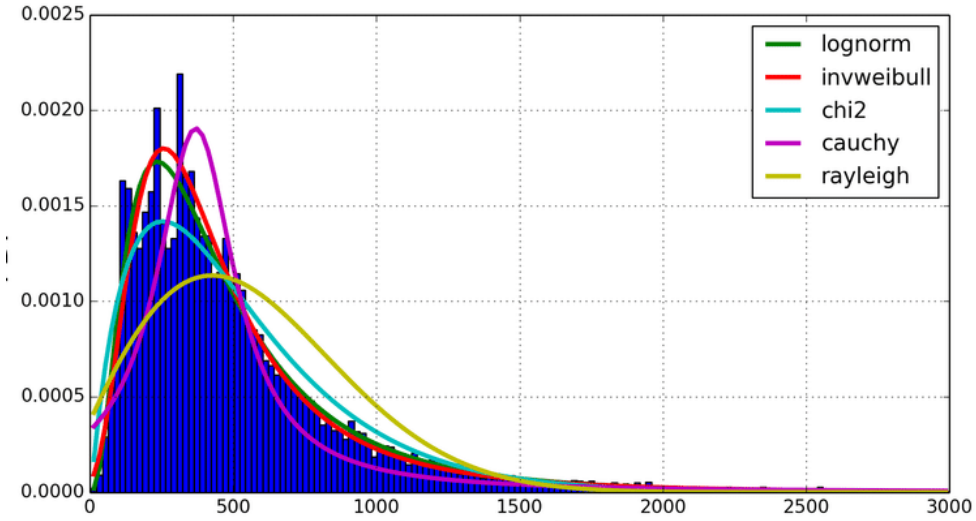

Figure 3.7: Maximum likelihood for different PDFs [57].

Then, to classify a customer, the measured output power (from the smart meter) will be compared to the predicted output power. The error between the measured output and the predicted output will be checked using the fitted PDF of the extracted error and an alarm can be triggered to indicate that a theft is detected if a certain criterion is met. This criterion is chosen based on the probability of occurrence of the error between the measured output and the predicted output. If the error's probability of occurrence is below a certain threshold (for example 5%), it will be classified as a suspicious customer [58]. Only large errors have small probability to occur, which is our interest as they are more unlikely to happen (customer might be stealing and shall be under the investigations).

Chapter 4 . Results and Discussions

In this chapter, the results will be presented and discussed along with some case studies. Table 4.1 shows the list of variables and assumptions that were used in this research. The data was used to train several models to find the best model that can be used to accurately detect suspicious data using the three predictors: temperature, solar irradiance and time of the day as input parameters and the virtual output power of the PV panels of each type after normalizing. Eight different models were trained and the Root Mean Square Error (RMSE) of each type is shown in Figure 4.1. According to the RMSE, decision tree model gave the least error (0.0027612) and therefore was used in this study.

Table 4.1: Parameter values for all case studies.

Parameter	Value
Number of customers	400
Location	Toronto, Canada
Data point per customer	48
Type of PV panels	11 randomly assigned
Training Data	Customers with types 1 to 10
Testing data	Customers with type 11
Number of PV installed	5-50 randomly assigned
Time step	1 hour
Solar irradiance	From Toronto
Temperature	From Toronto
Error Threshold	5%

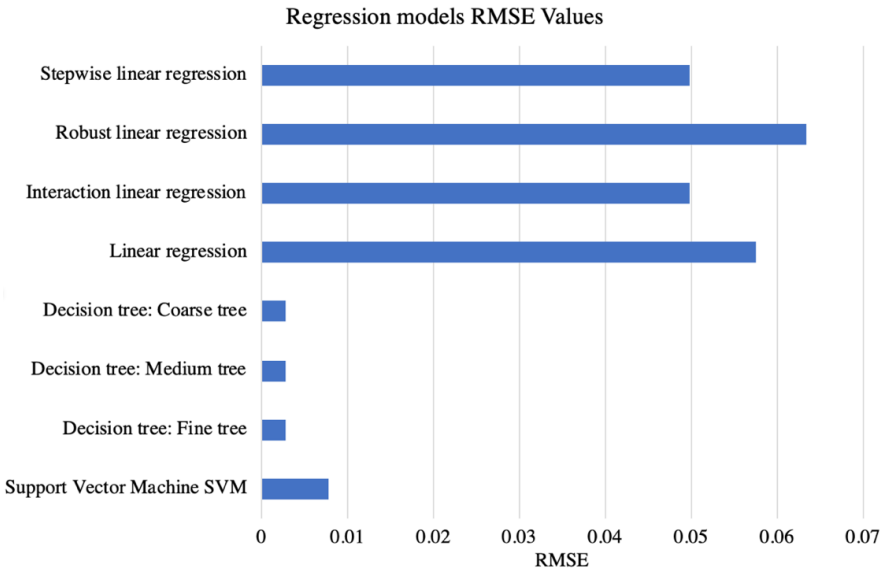

Figure 4.1: Regression models RMSE Values.

Figure 4.2 shows the original data set which is the output power or the response variable that was generated using the three predictors: time, temperate and solar irradiance, whereas Figure 4.3 shows the trained model using fine tree algorithm. The blue represents the actual data points, the yellow represents the predicted value while the red represents the error.

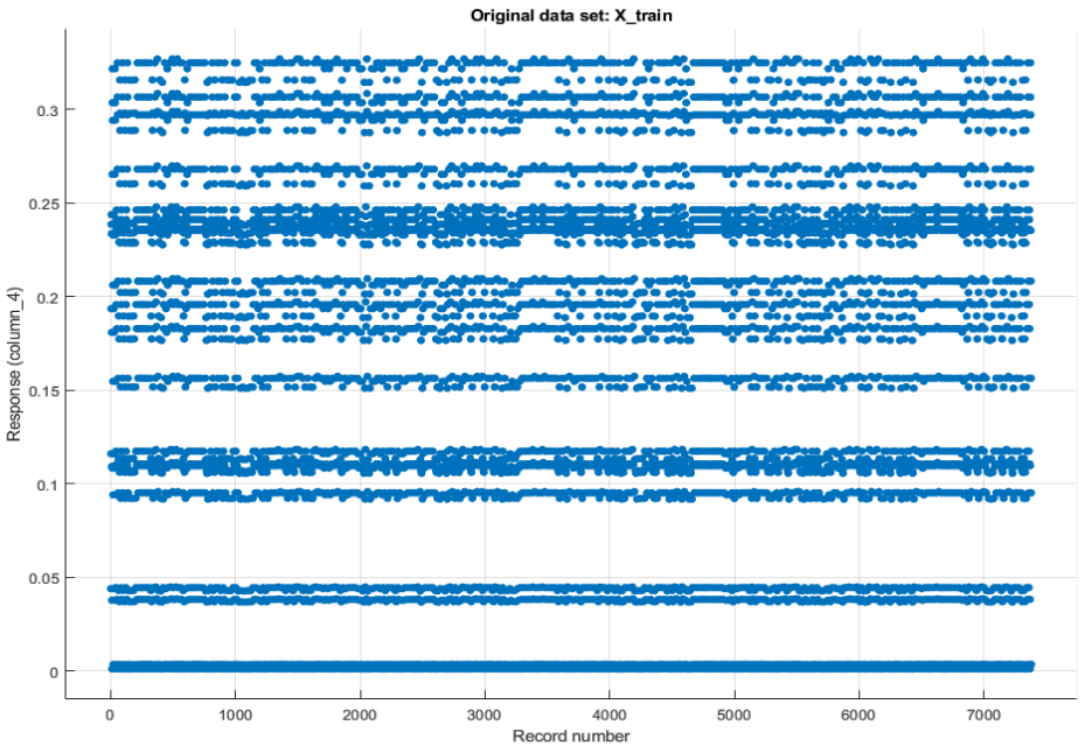

Figure 4.2: Original data points before training from the regression learner tool.

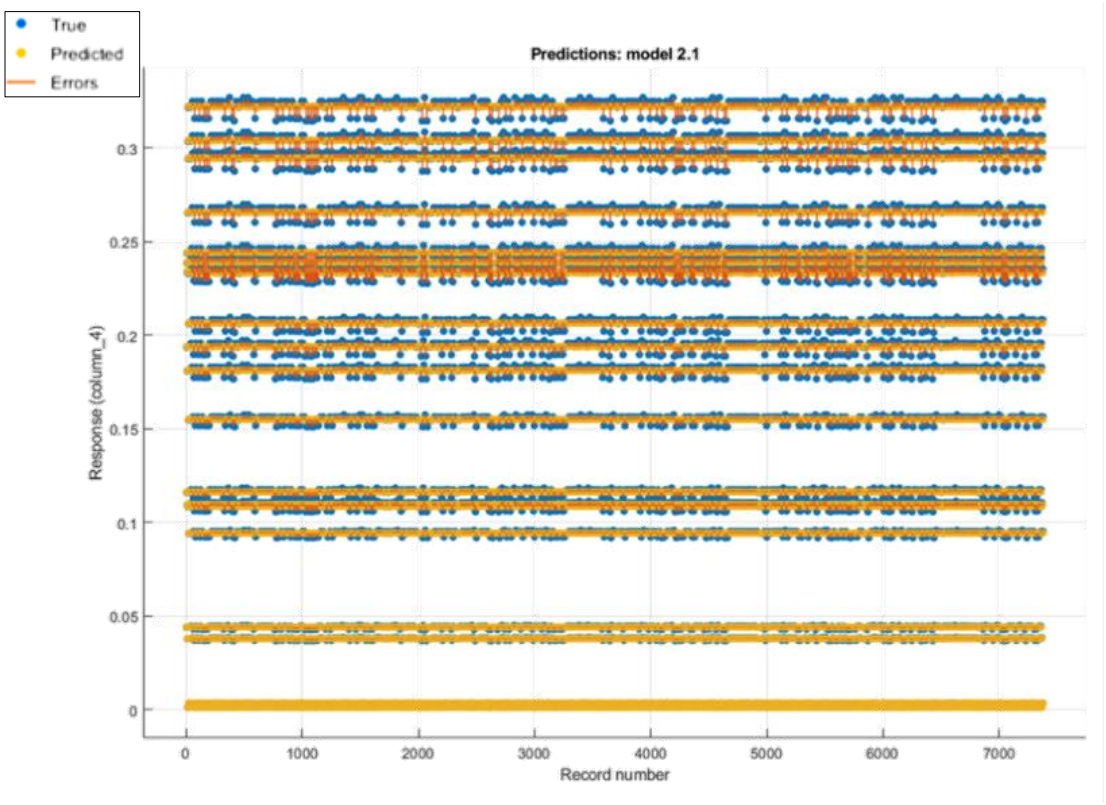

Figure 4.3: Trained model using fine tree algorithm.

The error of the model was then extracted from the regression learner tool to find the best PDF fit for the theft detection unit. The extracted error of the trained model was used to plot the histogram of different distribution types including Beta, Lognormal, Weibull and exponential. Figure 4.4 and 4.5 illustrates the error histogram and the different density distributions respectively. Comparing the likelihood of each histogram, Beta distribution gave the best (highest) likelihood which was 35837 and therefore, was used to build the detection unit. The PDF alone is useless as the probability for a random variable to occur at a given point is zero, therefore, integrating the PDF gives the CDF which gives the probability up to a particular value.

After extracting the model and building the theft detection unit, the accuracy of the model was tested. Using a set of data that the model has not seen before is important to avoid any bias and accurately evaluate the performance of the model. Therefore, all data points of the Type 11 of PV panels where separated from the data set previously and was not used for training. Testing the model using Type 11 data, which we know is within the acceptance threshold (no theft is introduced yet), shows that the customer is not sealing which is correct. Figure 4.6 illustrates the probability of occurrence after

testing. It is clear that none of the data points have a probability less than the threshold, which is 5% in this research.

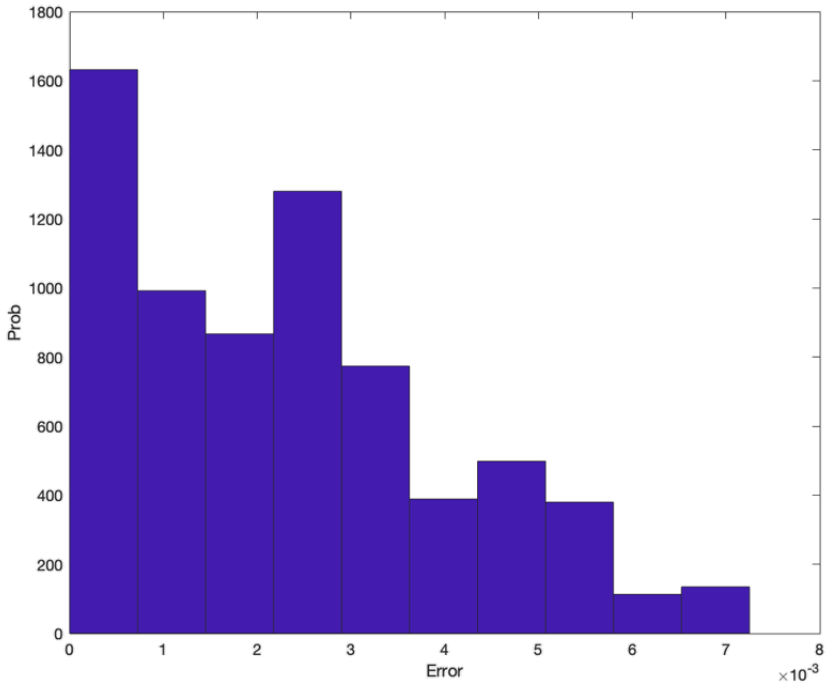

Figure 4.4: Error histogram.

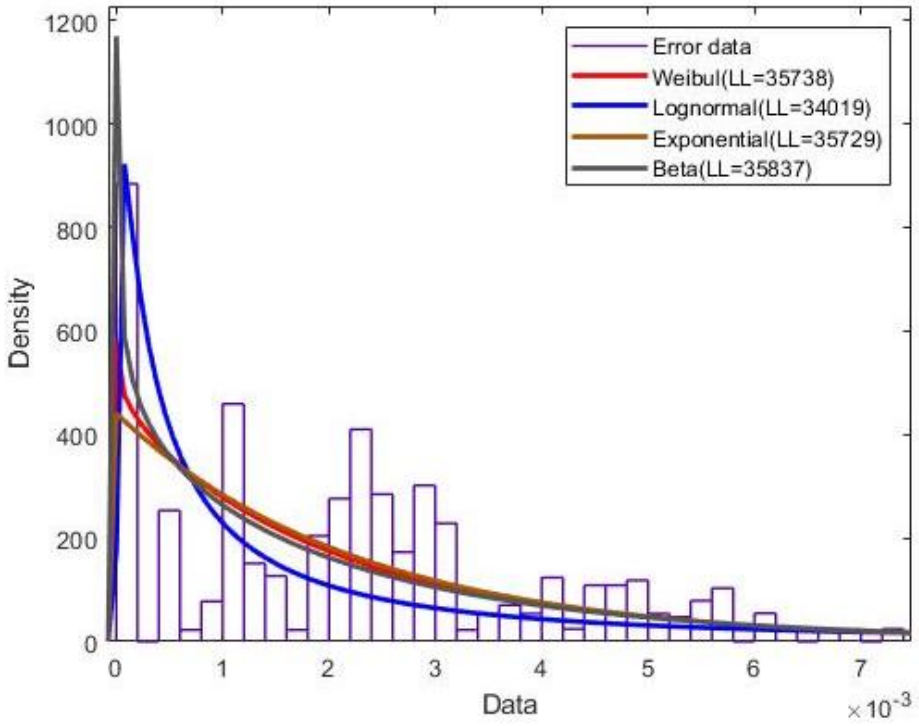

Figure 4.5: PDF of different distribution of the error.

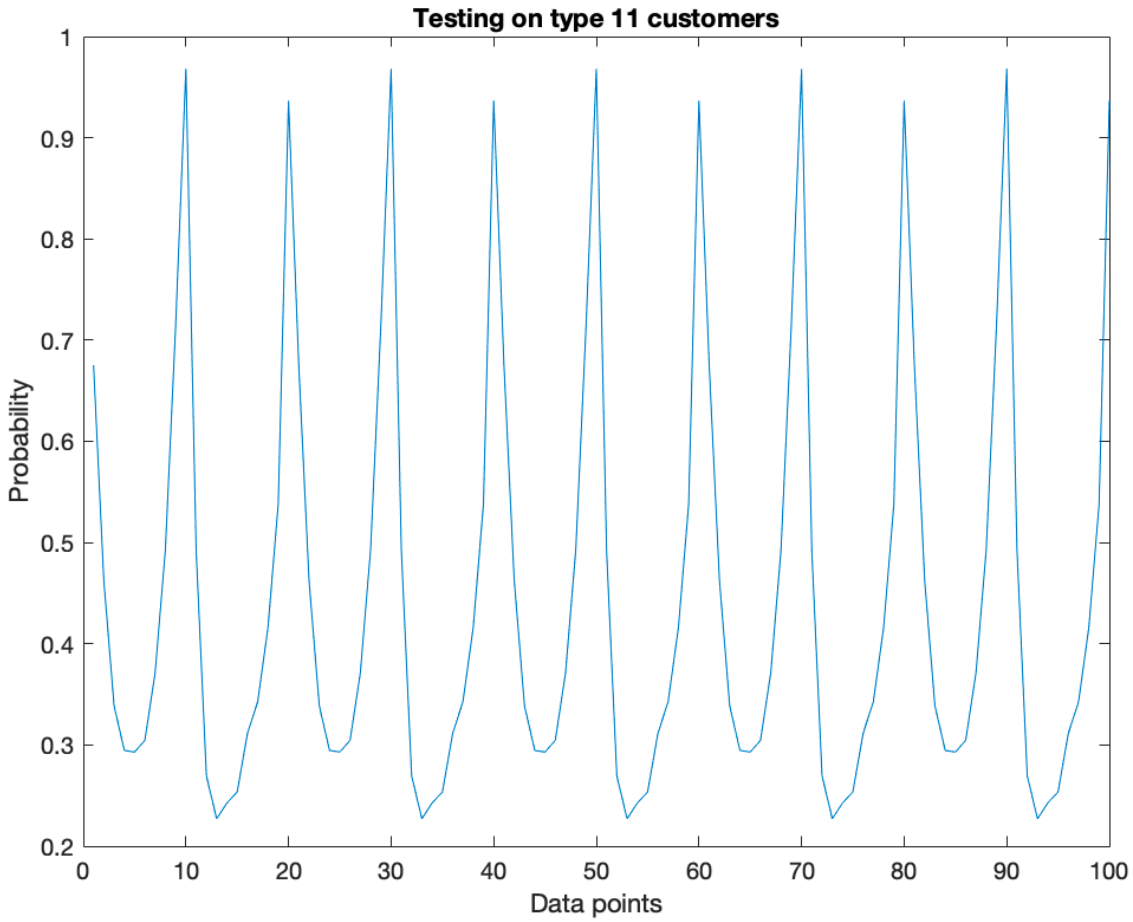

Figure 4.6: The probability of Type 11 data points occurrence.

To study the accuracy of the theft detector unit, several case studies for different scenarios of theft where simulated:

Case 4.1: Stealing by a fixed multiplier 2.5%, 5%, 10%

Assuming that a customer will maliciously increase the injection by a fixed multiplier. Figure 4.7, 4.8, 4.9 shows the probability of occurrence when the injection is increased by 2.5%, 5% and 10% respectively. It can be clearly observed that most of the data points are having a very low probability (less than the threshold 5%) which indicates that this injection is unlikely to happen. The error probability is high only when the injection from the PV is high, which occurs at the middle of the day time. Whereas during the night time, the injection is zero. Hence, the amount of theft is not significant when the PV injection is low. Thefrore, customers will focus of the daytime periods to manuplate the data and increas the injection.

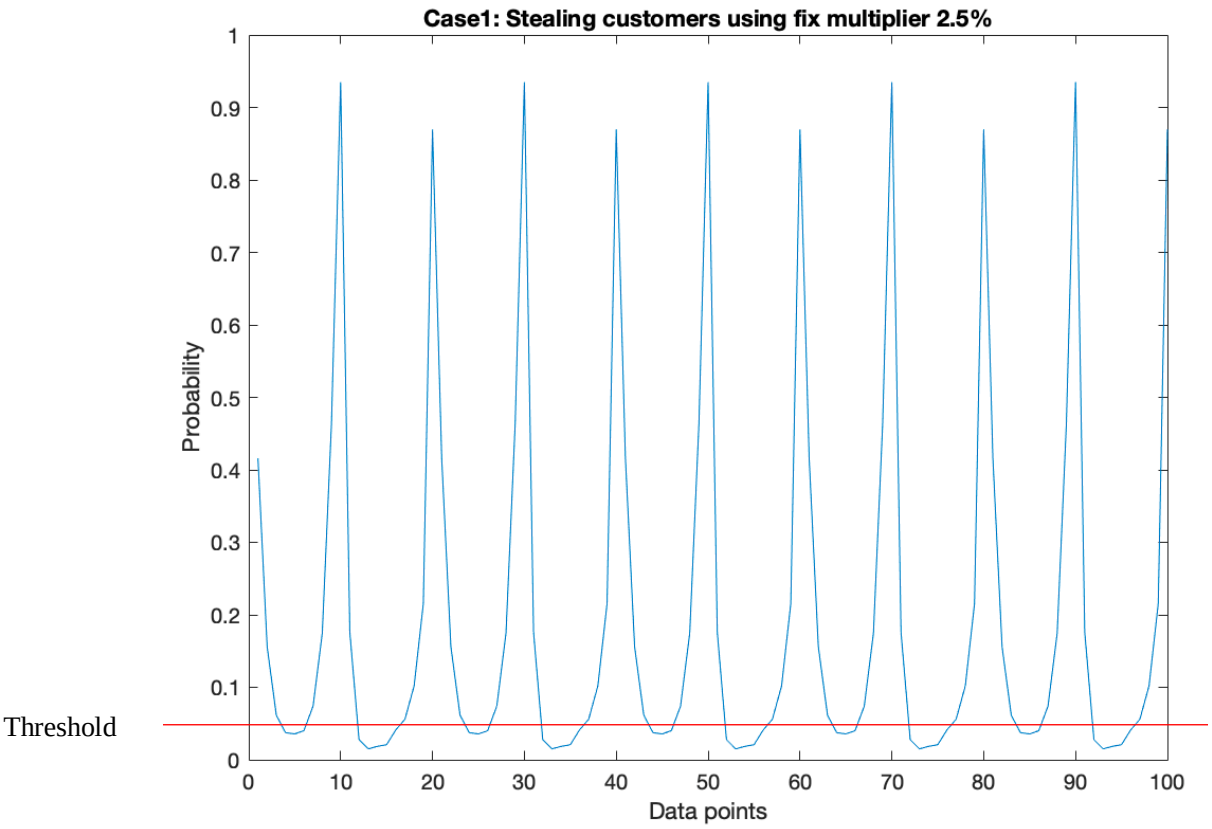

Figure 4.7: The probability of occurrence when the injection in multiplied by 2.5%.

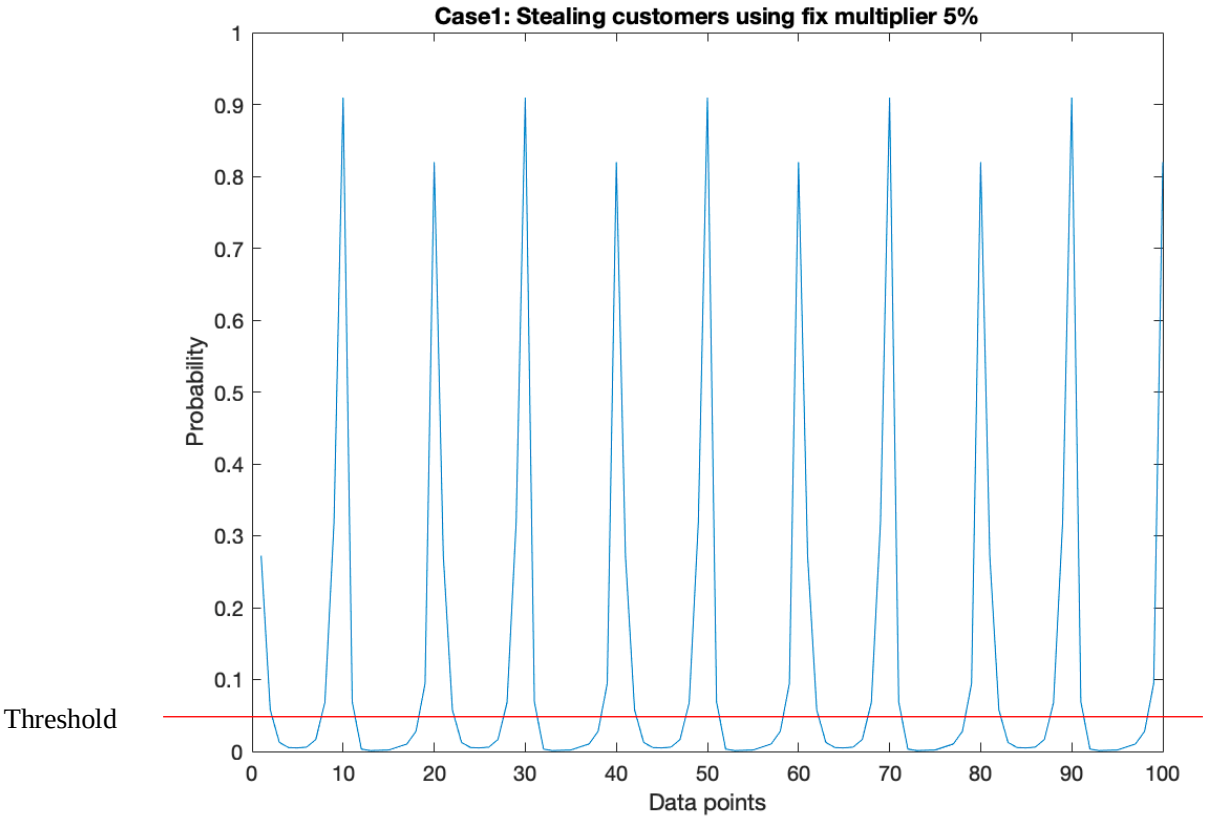

Figure 4.8: The probability of occurrence when the injection in multiplied by 5%.

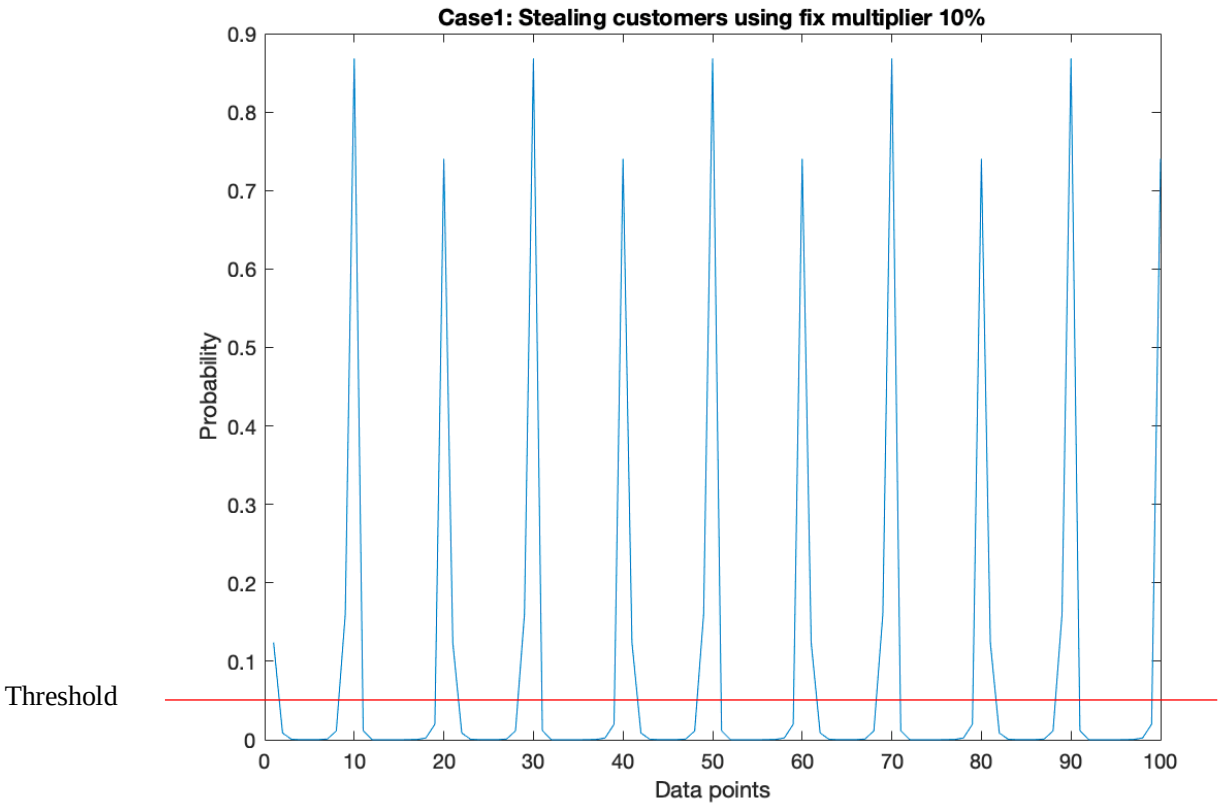

Figure 4.9: The probability of occurrence when the injection in multiplied by 10%.

Case 4.2: Stealing by a random multiplier.

Another scenario is assumed when the customer adds a random increase in the injection to make it difficult for the utility to recognize the abnormality. Figure 4.10 shows the probability. The random multiplier is generated in MATLAB using the command rand which generates pseudorandom values drawn from the standard uniform distribution on the open interval (0,1).

Case 4.3: Stealing by a fixed multiplier with partial zero elimination.

The main purpose of this use case is to test the performance of the model without curtailing all zero. This will measure how capable is the system in detecting solar injection during night times. Figure 4.11 shows the waveform considering 2.5% injecting including night time slots.

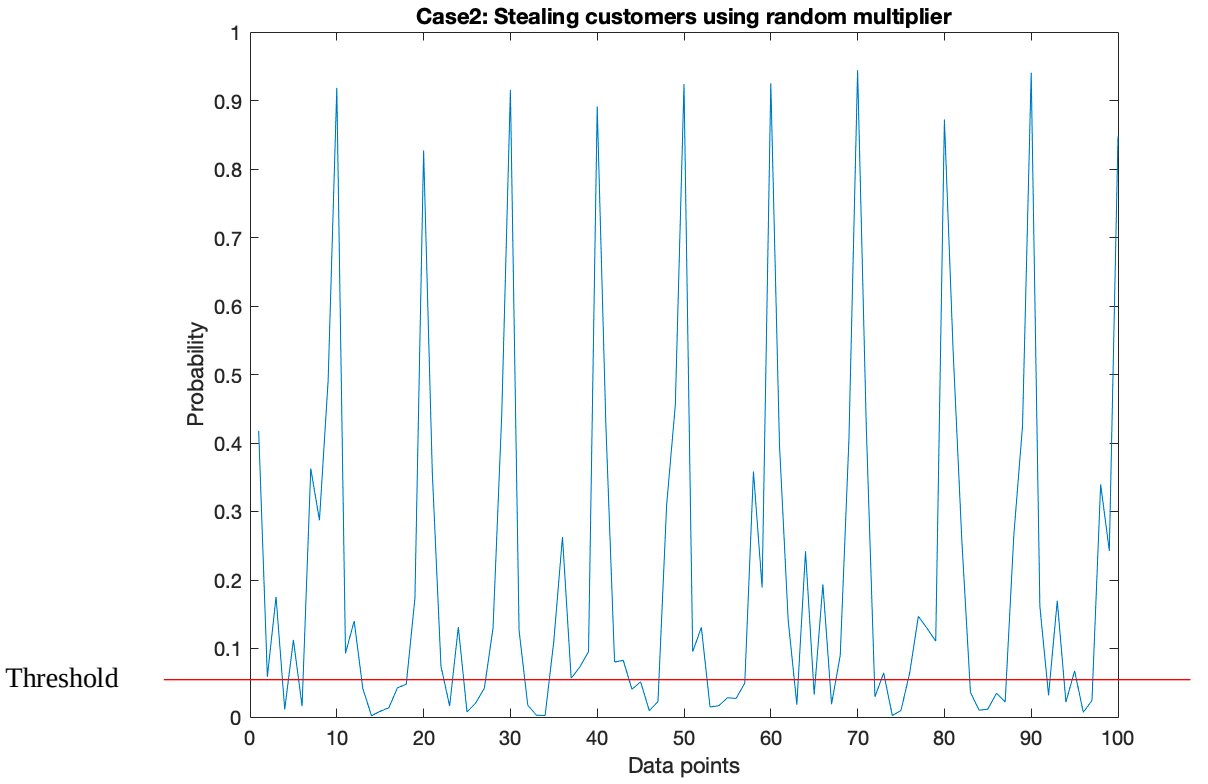

Figure 4.10: The probability of occurrence when the injection in randomly increased.

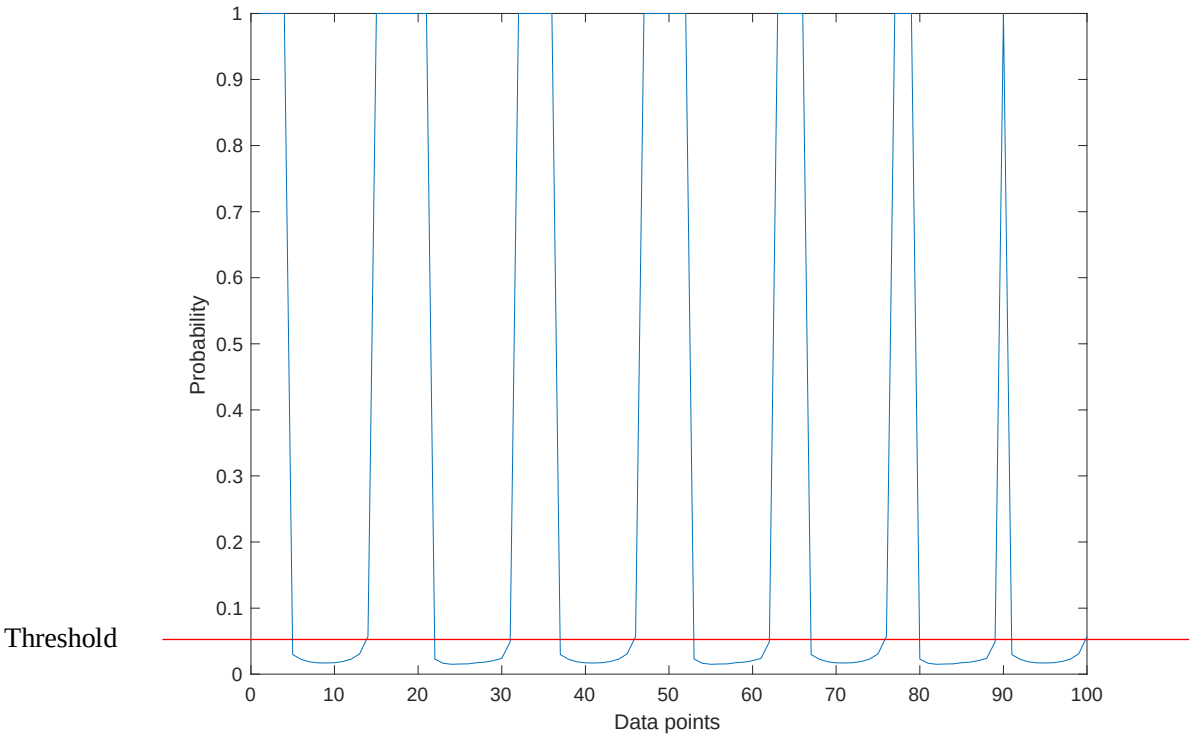

Figure 4.11: Testing with zeros.

Chapter 5 . Conclusion and Future Work

In this chapter, the conclusion of this research work results is presented, as well as the future work that can be done to enhance the model.

5.1. Conclusion

In this research, a regression-based algorithm was used to build a model that predicts the normalized output of PV modules using three predictors. After gathering the required data for training, several regression models were trained. The predicted output was compared to the actual PV module output power and the root mean square error (RMSE) was calculated for each corresponding model. The (RMSE) value of each trained model was used in evaluating the performance and accuracy. Regression tree model gave the least RMSE value; hence, was used to build the Theft Detection Unit for PV (TDUPV).

As for the detection mechanism, several probability distributions were used to fit the error values extracted previously. The error of a large data set was found to follow Beta distribution as it gave the highest likelihood. However, the challenge was to make sure that the error is not influenced by any sort of bias. The huge amount of zero error values that was corresponding to the night time observations when the irradiance is zero caused that bias. Therefore, to remove the bias, 90% of the zero observations were removed. The remaining 10% was kept to test the model against the rare scenarios of night time injection by malicious customers. Furthermore. The criterion of the detection depended on a predefined threshold value, which is 5%, in other words, if the probability of the error is below 5%, i.e. unlikely to happen, the customer is considered suspicious.

Simulation results of three use cases on typical theft scenarios assuming cyber-attacks proved the effectiveness of the proposed approach.

5.2. Future Work

To further enhance the results, experimental work can be conducted to showcase some practical numbers regarding the amount of generated solar energy for each of the 11 types. This will give a realistic comparison between the actual and the predicted

output of the panels; consequently, it will validate the threshold that was set in this research work. For instance, if the PV injection was found to be experimentally varying beyond 5% then this indicates the threshold must be something more that 5% for customer to be suspected.

Another future work can be done by considering the injection from Plugged in Hybrid Electric Vehicles (PHEV) in what is called vehicle to grid (V2G) services. This will include the behaviour of customers within a specific city and the amount of energy they feed the grid through their electric Vehicles (EVS) and predicting their future behaviour.

References

[1] "5 Ways Energy & Utilities Firms Can Capitalize On Streaming Analytics", *Vitria.com*, 2014. [Online]. Available: https://www.vitria.com/pdf/WP-Energy-041514.pdf. [Accessed: 18- Jan- 2019].

[2] H. Sun, Chatzēargyriu Nikos, H. V. Poor, and L. Carpanini, *Smarter energy: From smart metering to the smart grid*. London: The Institution of Engineering and Technology, 2016, pp. 1,2.

[3] Cory, Karlynn, Toby Couture, and Claire Kreycik. "Feed-in tariff policy: design, implementation, and RPS policy interactions", Report No. NREL/TP-6A2-45549. National Renewable Energy Lab.(NREL), Golden, CO (United States), 2009.

[4] Department for Business, Energy and Industrial Strategy (BEIS), "Feed-In Tariff (FIT) rates", ofgem.gov.uk, 2019. [Accessed: 2-May-2019].

[5] M. Pocci, "Feed-In-Tariff Handbook for Asian Renewable Energy Systems", *Winston.com*, 2014. [Online]. Available: https://www.winston.com/images/content/9/1/v2/91697/Feed-In-Tariff-Handbook-for-Asian-Renewable-Energy-Systems.pdf. [Accessed: 15- Jan- 2019].

[6] A. Campoccia, L. Dusonchet, E. Telaretti and G. Zizzo, "Feed-in Tariffs for Grid-connected PV Systems: The Situation in the European Community," 2007 IEEE Lausanne Power Tech, Lausanne, 2007.

[7] M. Boxwell, *Solar electricity handbook*. Greenstream Publishing, 2019.

[8] [7]N. Author, "Japan to more than halve its solar power feed-in tariffs | The Japan Times", *The Japan Times*, 2018. [Online]. Available: https://www.japantimes.co.jp/news/2018/09/15/business/japan-halve-solar-power-feed-tariffs/#.Xs5PrWgzY2x. [Accessed: 28- Mar- 2018].

[9] "IEA - China." [Online]. Available: https://www.iea.org/policiesandmeasures/pams/china. [Accessed: 29-Mar-2019].

[10] G. M. Masters, *Renewable and efficient electric power systems*, Second edition. Hoboken, New Jersey: John Wiley & Sons Inc, 2013.

[11] "Calls for Dubai to introduce feed-in-tariff to aid rooftop solar," *The National*. [Online]. Available: https://www.thenational.ae/business/calls-for-dubai-to-introduce-feed-in-tariff-to-aid-rooftop-solar-1.62624. [Accessed: 13-Nov-2018].

[12] S. Seme, K. Sredensek and Z. Praunseis, "Smart grids and net metering for photovoltaic systems," 2017 *International Conference on Modern Electrical and Energy Systems (MEES)*, Kremenchuk, 2017, pp. 188-191,

[13] [8]"Will smart grids be vulnerable to cyber attacks?", *Power Technology | Energy News and Market Analysis*, 2018. [Online]. Available: https://www.power-technology.com/comment/will-smart-grids-vulnerable-cyber-attacks/. [Accessed: 13- Jun- 2018].

[14] J. T. Nutter, "Uncertainty and Probability", In 1987 Proceedings of *The Tenth International Joint Conference on Artificial Intelligence (IJCAI)*, Virginia, 1987, pp. 373–379.

[15] M. Evans and J. Rosenthal, *Probability and statistics*, 2nd ed. Toronto: University of Toronto, 2019, pp. 62, 63.

[16] G. Grimmett and D. Stirzaker, *Probability and random processes*, 3rd ed. Oxford: Oxford Univ. Press, 2009, p. 90.
[17] "NIST/SEMATECH e-Handbook of Statistical Methods", *Itl.nist.gov*, 2003. [Online]. Available: http://www.itl.nist.gov/div898/handbook/. [Accessed: 26- Mar- 2019].
[18] J. Brownlee, "A Gentle Introduction to Statistical Data Distributions", *Machine Learning Mastery*, 2018. [Online]. Available: https://machinelearningmastery.com/statistical-data-distributions/. [Accessed: 07- Jun- 2018].
[19] M. Stephens, "The Beta Distribution", *fiveMinuteStats*, 2017. [Online]. Available: https://stephens999.github.io/fiveMinuteStats/beta.html. [Accessed: 06- Oct- 2019].
[20] S. Ghahramani, *Fundamentals of probability with stochastic processes*. Upper Saddle River, N.J.: Pearson/Prentice Hall, 2005, pp. 273-297.
[21] D. Lallemant and A. Kiremidjian, "*A Beta Distribution Model for Characterizing Earthquake Damage State Distribution*", *Earthquake Spectra*, vol. 31, no. 3, pp. 1337-1352, 2015.
[22] M. Sullivan, *Statistics: Informed decisions using data*. Harlow, Essex: Pearson, 2014.
[23] S. S. Gokhale and R. E. Mullen, "*Application of the Lognormal Distribution to Software Reliability Engineering," in Handbook of Performability Engineering*, K. B. Misra, Ed. London: Springer London, 2008, pp. 1209–1225.
[24] C. Walck, *Hand-book on STATISTICAL DISTRIBUTIONS for experimentalists*. Stockholm: University of Stockholm, 1996, p. 86.
[25] D. Montgomery, G. Runger and N. Hubele, *Engineering statistics*. New York: Wiley, 2001, p. 101.
[26] F. GABBIANI and S. J. COX, *Mathematics for Neuroscientists*. Science direct, 2010.
[27] C. Lai, *Generalized Weibull Distributions*. Berlin: Springer-Verlag, 2014.
[28] Langley, P. (1996). Elements of machine learning. San Francisco, Calif.: Morgan Kaufmann, p.1.
[29] J. Brownlee, "Supervised and Unsupervised Machine Learning Algorithms", *Machine Learning Mastery*, 2016. [Online]. Available: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/. [Accessed: 15- Mar- 2019].
[30] [6]G. Tso and K. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks", *Energy*, vol. 32, no. 9, pp. 1761-1768, 2007. Available: 10.1016/j.energy.2006.11.010 [Accessed 15 March 2015].
[31] J. Frost, "Choosing the Correct Type of Regression Analysis - Statistics By Jim", *Statistics By Jim*, 2017. [Online]. Available: https://statisticsbyjim.com/regression/choosing-regression-analysis/. [Accessed: 14- Oct- 2018].
[32] Z. Yang, *Machine learning approaches to bioinformatics*. Hackensack, NJ: World Scientific, 2010.
[33] R. Barga, V. Fontama and W. Tok, *Predictive analytics with Microsoft Azure Machine Learning*. Berkeley,CA: Apress, 2015.

[34] A. Nobel, "Recursive partitioning to reduce distortion", *IEEE Transactions on Information Theory*, vol. 43, no. 4, pp. 1122-1133, 1997. Available: 10.1109/18.605573.
[35] X. Ma, *Using Classification and Regression Trees : A Practical Primer*. Information Age Publishing, Incorporated, 2018, p. 19.
[36] C. Bird, T. Menzies and T. Zimmermann, *The Art and science of analyzing software data*. Amsterdam: Elsevier, Morgan Kaufmann, 2015.
[37] S. V. M. Vishwanathan and M. Narasimha Murty, "SSVM: a simple SVM algorithm," *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, Honolulu, HI, USA, 2002, pp. 2393-2398 vol.3, doi: 10.1109/IJCNN.2002.1007516.
[38] H. Lam, S. Ling and H. Nguyen, *Computational intelligence and its applications*. London, UK: Imperial College Press, 2012, p. 10.
[39] M. Ismail, M. Shahin, M. F. Shaaban, E. Serpedin and K. Qaraqe, "Efficient detection of electricity theft cyber attacks in AMI networks," *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, Barcelona, 2018, pp. 1-6, doi: 10.1109/WCNC.2018.8377010.
[40] F. Xiao and Q. Ai, "Electricity theft detection in smart grid using random matrix theory," *in IET Generation, Transmission & Distribution*, vol. 12, no. 2, pp. 371-378, 30 1 2018, doi: 10.1049/iet-gtd.2017.0898.
[41] S. Salinas, M. Li and P. Li, "Privacy-preserving energy theft detection in smart grids," *2012 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, Seoul, 2012, pp. 605-613, doi: 10.1109/SECON.2012.6275834.
[42] P. Jokar, N. Arianpoo and V. C. M. Leung, "Electricity Theft Detection in AMI Using Customers' Consumption Patterns," in *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216-226, Jan. 2016, doi: 10.1109/TSG.2015.2425222.
[43] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines," *IEEE Transaction on Power Delivery*., vol. 25, no. 2, pp. 1162–1171, Apr. 2010.
[44] Y. Liu and S. Hu, "Cyberthreat Analysis and Detection for Energy Theft in Social Networking of Smart Homes," *IEEE Transactions on Computational Social Systems*, vol. 2, no. 4, pp. 148–158, Dec. 2015.
[45] Yao, M. Wen, X. Liang, Z. Fu, K. Zhang and B. Yang, "Energy Theft Detection with Energy Privacy Preservation in the Smart Grid", *IEEE Internet of Things Journal*, pp. 1-1, 2019. Available: 10.1109/jiot.2019.2903312.
[46] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar and S. Mishra, "Decision Tree and SVM-Based Data Analytics for Theft Detection in Smart Grid", *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1005-1016, 2016. Available: 10.1109/tii.2016.2543145.
[47] W. Han and Y. Xiao, "NFD: Non-technical loss fraud detection in Smart Grid", *Computers & Security*, vol. 65, pp. 187-201, 2017. Available: 10.1016/j.cose.2016.11.009.
[48] W. Yu, D. Griffith, L. Ge, S. Bhattarai and N. Golmie, "An integrated detection system against false data injection attacks in the Smart Grid", *Security and Communication Networks*, vol. 8, no. 2, pp. 91-109, 2014.
[49] "Historical Data - Climate - Environment and Climate Change Canada", *Climate.weather.gc.ca*. [Online]. Available:

https://climate.weather.gc.ca/historical_data/search_historic_data_e.html. [Accessed: 07- Jan- 2019].
[50] "PV Solar Panels | Photovoltaic Panels | Solar Electric Panels & Modules." [Online]. Available: http://www.solarpanelsplus.com/products/solar-pv-panels/. [Accessed: 06-Feb-2019].
[51] "Solar Water Heaters | Solar Air Conditioning | PV Solar Panels | Solar Heating | Solar Thermal." [Online]. Available: http://www.solarpanelsplus.com/. [Accessed: 06-Feb-2019].
[52] C. Chen, *Physics of solar energy*. Hoboken, NJ: John Wiley & Sons, 2011, p. 89.
[53] A. T. Umoette, E. A. Ubom, and I. E. Akpan, "Comparative Analysis of Three NOCT-Based Cell Temperature Models," Int. J. Syst. Sci. Appl. Math., vol. 1, no. 4, p. 69, Dec. 2016.
[54] "Regression Learner App - MATLAB & Simulink." [Online]. Available: https://www.mathworks.com/help/stats/regression-learner-app.html. [Accessed: 16-Jan-2019].
[55] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation", *Machine Learning Mastery*, 2018. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/. [Accessed: 22- May- 2018].
[56] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, Bhimavaram, 2016, pp. 78-83, doi: 10.1109/IACC.2016.25.
[57] "FIGURE 2 | Probability distributions fit to histogram of forecast...," *ResearchGate*. [Online]. Available: https://www.researchgate.net/Figure/Probability-distributions-fit-to-histogram-of-forecast-errors-in-Germany-in-2017_fig2_326614522. [Accessed: 08-Apr-2019].
[58] "Gaussian Distribution." [Online]. Available: http://hyperphysics.phy-astr.gsu.edu/hbase/Math/gaufcn.html. [Accessed: 16-Jan-2019].

Vita

Nouf Ahmad Almadani was born in 1994, in Dubai, United Arab Emirates. She was educated in local public schools and graduated from Sakina Bint Al Hussain High School with a cumulative percentage of 95.1% in 2012. She received TRA scholarship to the University of Sharjah in Sharjah, United Arab Emirates, from which she graduated with a degree of Bachelor of Science in Electrical and Electronics Engineering.

Ms. Almadani begun a Master's program in Electrical Engineering at the American University of Sharjah right after her graduation in 2017. In August 2017, Ms. Almadani started working in Dubai Electricity and Water Authority (DEWA) as an electrical engineer in the Smart Grid Cyber Security Project. In 2019, Ms. Almadani moved to the Future Accelerators team within the Innovation department in DEWA. Ms. Almadani have received the distinguished new employee award and currently is a member of DEWA Youth Council, DEWA Future Shaping committee and knowledge committee. Ms. Almadani was nominated for a four months mission in UK and was selected by UC Barkley to commence the second Master's program.