QATAR UNIVERSITY

COLLEGE OF ENGINEERING

ACCURATE CLASSIFICATION OF PARTIAL DISCHARGE PHENOMENA IN POWER TRANSFORMERS IN THE PRESENCE OF NOISE

BY

RACHAEL FERNANDEZ

A Thesis Submitted to the Faculty of

the College of Engineering

in Partial Fulfillment

of the Requirements

for the Degree of

Masters of Science in Computing

June 2017

COMMITTEE PAGE

The members of the Committee approve the Thesis of Rachael Fernandez defended on 21/05/2017.

_____

Dr. Khaled Bashir Shaban
Thesis/Dissertation Supervisor

_____

Dr. Ehab Fahmy El-Saadany
Committee Member

_____

Dr. Tamer Elsayed
Committee Member

_____

Dr. Mohammad Saleh
Committee Member

Approved:

_____

Khalifa Al-Khalifa, Dean, College of Engineering

ABSTRACT

FERNANDEZ, RACHAEL, K.., Masters:

June : 2017, Masters of Science in Computing

Title: Accurate Classification of Partial Discharge Phenomena in Power Transformers in the Presence of Noise

Supervisor of Thesis: Khaled, Bashir, Shaban.

The objective of this research is to accurately classify different types of Partial Discharge (PD) phenomenon that occurs in transformers in the presence of noise. A PD is an electrical discharge or spark that bridges a small portion of the insulation in electrical equipment, which causes progressive deterioration of high voltage equipment and could potentially lead to flashover. The data for the study is generated from a laboratory setup and it is 300 time series signals each with 2016 attributes corresponding to 3 types of PDs; namely: Porcelain, Cable and Corona. The data is collected from two sensors with different bandwidths, in which Channel A signals refer to the data collected from the higher frequency sensor and signals from Channel B refer to data of the lower frequency sensor. Different feature engineering approaches are investigated in order to find the set of the most discriminant features which help to achieve high levels of classification accuracy for Channel A and Channel B signals. First, features that describe the shape and pulse of signals in the time domain are extracted. Then frequency domain based statistical features are generated. In comparison with classification accuracies using frequency domain features, time domain based features gave higher accuracy of more than 90% on average for both channels in the absence of noise while frequency domain features allowed classification accuracy up to 80% on average. However, in the presence of noise, both

methods degraded. To overcome this, Regularization techniques were applied on the features from the frequency domain which helped to maintain classification accuracy even in the presence of high levels of noise.

DEDICATION

“For my family and friends, for their constant support, love and faith in me.”

ACKNOWLEDGMENTS

“I would like to thank Dr. Khaled Shaban for his constant support and patience.

I would also like to thank Ramy Hussein for his valuable suggestions.

Finally, I would like to thank everyone who made this journey, a memorable one.”

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of abbreviations

AE	Acoustic Emission
CT	Current Transformer
CWT	Continuous Wavelet Transform
DFT	Discrete Fourier Transform
DGA	Dissolved Gas Analysis
DSP	Digital Signal Processing
DWT	Discrete Wavelet Transform
EEG	Eelectroencephalogram
FFT	Fast Fourier Transform
GOOSE	Generic Object Oriented Substation Events
HPLC	High Performance Liquid Chromatography
HV	High Voltage
IED	Intelligent Electronic Device
KNN	K Nearest Neighbour
MU	Merging Unit
PCA	Principal Component Analysis
PD	Partial Discharge
PSA	Pulse Sequence Analysis
QP	Quadratic Programming
RF	Radio Frequency
RTAC	Real-Time Automation Controller
SCADA	Supervisory Control And Data Acquisition
SMO	Sequence Mining Optimization
SNR	Signal To Noise Ratio
SOM	Self-Organised Map
SVM	Support Vector Machine
UHF	Ultra-High Frequency
VHF	Very High Frequency

Chapter 1

1. Introduction

1.1 Overview and Motivation

The conventional power grid is a rigidly hierarchical system in which power plants at the top of the chain ensuring power delivery to all the customers in the network. The communication is essentially one-way and requires manual restoration in case of failure of any of the components in the network [1]. This conventional grid is now being revolutionized to become 'Smart' by the "integration and application of Real-time monitoring, advanced sensing, communications, analytics and control" [2]. The smart grid supports "dynamic flow of energy and information to accommodate existing and new forms of energy supply" and also incorporates renewable sources of energy like solar panels, wind mills, etc. [3] . The communication is now a two-way pipeline to promote various functions like self-coordination, self-awareness, self-healing, and self-reconfiguration, boosting the deployment of renewable energy sources, augmenting the efficiency of power generation, transmission, and usage [4]. The most important function of which is the self-healing which refers to continuous self- assessments to detect and analyse faults and respond quickly to restore grid components [5].

Power transformers are the main components of the grid and are numerous in number in any grid. They are responsible for transferring electrical energy from one electrical circuit to another without changing the frequency. Transformers and other high-voltage (HV) equipment exhibit a phenomenon of partial discharge (PD) which is "an electrical discharge or spark that bridges a small portion of the insulation in the electrical equipment" [6]. This activity leads to progressive deterioration of components and is a

common symptom of impending faults. The faults are sometimes catastrophic and usually lead to irreversible internal damage. Moreover, the transient and intermittent nature of PDs disables detection and protection relay until disastrous flashovers are developed [7]. Monitoring, detection and classification of different types of PD lead to partial implementation of the self-healing function.

PD activity is an electrical breakdown that is found in high-voltage equipment which is confined to regions in the insulator. It causes a slow but progressive degradation of the insulation material. Thus, PD measurement is one of the important methods for determining the health of the HV equipment. Also, different types of PD have different effects on the equipment and is therefore necessary to distinguish between these different types of PD. For example, a PD activity on the surface of the insulator is less harmful than PD activity found in an internal void in the insulator [7].

The health of the equipment can be continuously monitored by recording vast amount of data captured from sensors that are either attached or in close proximity to the equipment. The data necessary for classifying PD types are extracted. This necessitates the need for applying the principles of data mining for efficient extraction of useful knowledge from the raw data. Further, data mining can be classified into descriptive modelling and predictive modelling. The former is used for identifying patterns in the data, whereas the latter is used for making predictions for new data based on a set of historical data [7][8].

Before predictive modelling is used on the data to assess the condition of the HV equipment, the PD signal has to be de-noised, which helps in removing the noise due to external electrical or magnetic interferences that corrupt the data. Thereafter, the de-noised data are analysed in the time or frequency domain in order to differentiate between the

different PD types. Once the PD type is identified, the frequency spectrum of the PD signal can be analysed to assess/evaluate the severity of the PD activity based on the magnitude of the signal. The obtained information can then be used to interpret the equipment's health.

The detection of PD activity can be either online or offline. Traditionally, PD measurements were done offline using a decoupling capacitator which can decouple the transformer from the electrical network. This method necessitates the transformer to be disconnected from the network and PD measurement is typically done during the maintenance period. However, in online monitoring, the PD measurement can be carried out during regular operation of the transformer [8]. Physio-chemical analysis of the oil in oil-filled transformers is one of the popular methods for online monitoring in which the degradation and properties of the oil and the gases that are released can be used to predict the type of the PD activity. In this thesis, we will perform online monitoring of the transformer using non-invasive current transformer (CT) sensors, so that the PD activities can be classified in real time. Specifically, we focus on performing accurate PD classification in a noisy environment, where often the performances of traditional approaches degrade.

Although experts have been investigating PD activity in HV equipment for decades now, there is no universally accepted method for PD detection [19]. This is due to the different types of PD activity that can be generated. In addition, different types of sensors could be used for acquisition of the signal. This leads to unlimited variations in the type of the PD detection system that can be constructed. Hence, a method that is tailored to our needs must be implemented to detect and classify PD activity [8].

1.2 Problem Statement

We are tackling the problem of PD classification in the presence of noise, and aiming to perform that with high accuracy regardless of the level of noise. This problem poses various challenges including:

1) **Hard to detect PD activity:** The PD activity in the signal does not occur continuously nor does it appear in regular intervals. This makes it hard to detect harmful PD activity in transformers due to their transient and intermittent nature. The only way to detect these PD signals are by monitoring the equipment continuously so that variations in the signal can be identified. These variations could also be due to the noisy environment surrounding the transformers that is susceptible to magnetic and electrical interferences that often interfere with PD detection. Hence, a method for processing the data that can tolerate these inaccuracies have to be implemented [9].
2. **Hard to achieve high classification accuracies, particularly in the presence of high noise levels**: Many PD detection systems have been reported in the literature [13]. These systems however are not robust in imperfect conditions, since their performance is seriously affected in noisy environments. This work represents one step forward towards robust detection of PD signals. *Figure 1* shows the different levels of noise (different Signal-to-Noise Ratio (SNR) levels with which our data is corrupted) with which we will test our system.

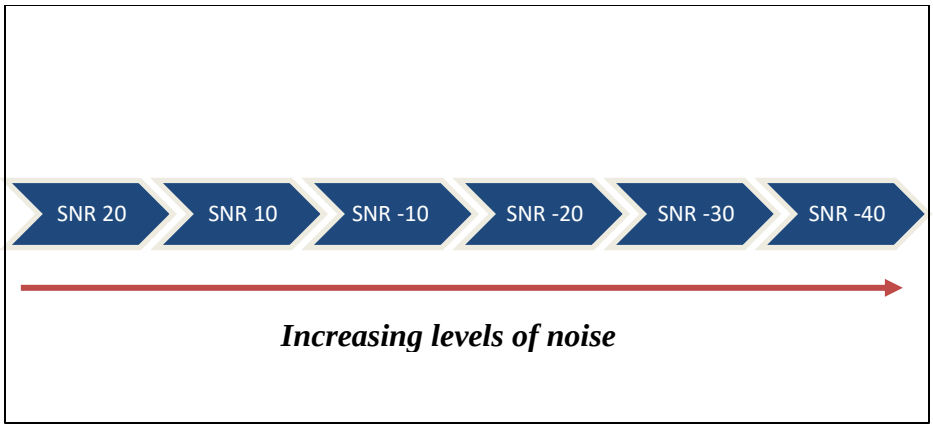

Figure 1: Different levels of noise with which we will test our system

In the this thesis, we overcome these challenges by converting the signal in the time domain to a signal in the frequency domain, as it is less susceptible to noise when compared to the signal in the time domain. Unlike the previous methods, we introduce an automated feature learning method that relaxes the need for data de-noising and maintains high classification accuracies in imperfect conditions. Regularization techniques are applied to the spectrum of the PD signals in order to adaptively select the most discriminant features of PD signals and preserve high classification accuracies.

1.3 Thesis Objectives

The aim of this research is to devise a method for online monitoring and accurate classification of PD activity in power transformers such that it is immune to noise. In this thesis we will classify three common types of PDs that are external to the transformer, namely: Porcelain, Cable and Corona. The main objectives of the research is to construct a classification model that is robust to noise. Moreover, data has to be processed in real time.

1.4 Contributions of the thesis

In this thesis, we have contributed to PD research by:

1) Creating two sets of novel features for processing the signal in the Time domain.
2) Applying Regularization techniques for classifying the different PD types. To the best of our knowledge, no one has applied Regularization techniques for diagnosing PD types before.
3) Using a novel form of pessimistic testing for measuring classification accuracy.

1.5 Thesis Overview

This chapter provided an overview of the research and presented the research problem and objectives. In Chapter 2, a background to the main concepts used in this research are introduced, followed by a review of the related work in PD classification. In Chapter 3, we describe the methodology that is followed in the research. Chapter 4 will give a detailed description of the experimental setup, datasets and the results that were achieved. We finally conclude with Chapter 5 that will provide a summary of our research and the future work that will be followed to construct a complete PD monitoring system.

Chapter 2

2. Background and Related Work

In this chapter, we will review some of the previous work on detecting PD along with some background information on the concepts that are related to this thesis.

2.1 Background

In this section, we will review some of the main concepts related to electrical signals and the machine learning approaches that are required for better understanding.

2.1.1 Background on Electrical Signals

Some of the main concepts related to electrical signals like the domains for signal processing, signal de-noising and phase-resolved analysis will be reviewed in this section.

A signal is a function that carries information by mapping to a domain, often time or space into a range such as air pressure, light intensity, etc. [11].

Each sinusoidal signal can vary with time over 3 parameters, namely:

2. ***Frequency:*** Measure of how many times the sinusoidal wave completes a full cycle, where one full cycle is 360 degrees of phase.
3. ***Amplitude:*** Measure of the intensity of the wave.
4. ***Phase:*** Measure of the relative displacement between or among waves having the same frequency.

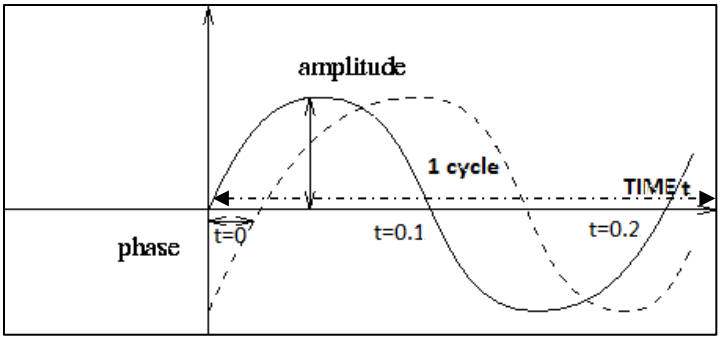

Figure 2: Sinusoidal wave that completes a cycle in 0.2 s

Figure 2 depicts a sinusoidal wave that completes a full cycle in 0.2 seconds, (i.e.) a frequency of 5 for every second.

2.1.1.1 Signal Domains

The PD signal can be processed in two different domains for analysing PD, where each domain has its own set of features: a) Time domain b) Frequency domain c) Time-Frequency domain. .

a) Time Domain

The signal in the time domain displays the displacement over a period of time, e.g., stock market values. It is plotted with time in the X-axis and the amplitude of the signal in the Y-axis. *Figure 3* shows the plot of the signal in the time domain.

The signal in the time domain is generally not regarded as suitable for detecting PD types, as it is suitable for recognising only certain types of PD and is highly sensitive to noise. These disadvantages lead to the transition to the frequency domain to process the data which is better suited for detecting PD type and location [14].

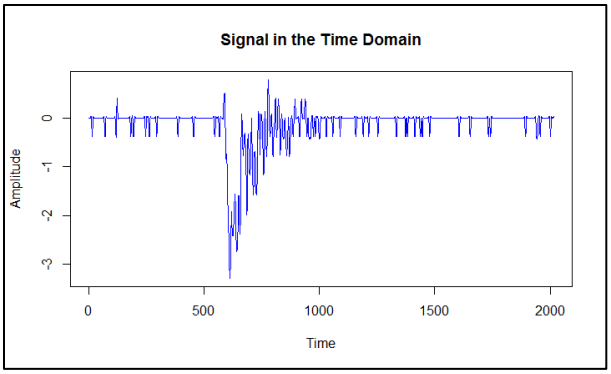

Figure 3: PD signal in the time domain

b) Frequency Domain

While the signal in the time-domain shows how the amplitude of the signal changes over time, the frequency-domain shows how the signal's energy is distributed over a range of frequencies. In other words, it measures how many times the signal cycle is repeated over a certain period of time. It is plotted by the range of frequencies over which the signal exists in the X-axis and the magnitude of the signal in the Y- axis. This representation of the frequency domain is commonly called the frequency spectrum [15].

Discrete Fourier Transform (DFT) or Fast Fourier Transform (FFT) can be applied on the signal in the time domain to convert into a signal in the frequency domain. Processing the signal in this domain is one of the most popular approaches for PD detection [8][16][17][21], in which each sample in the frequency spectra can be considered as a feature. The frequency domain is less susceptible to noise when compared with the time domain [22].

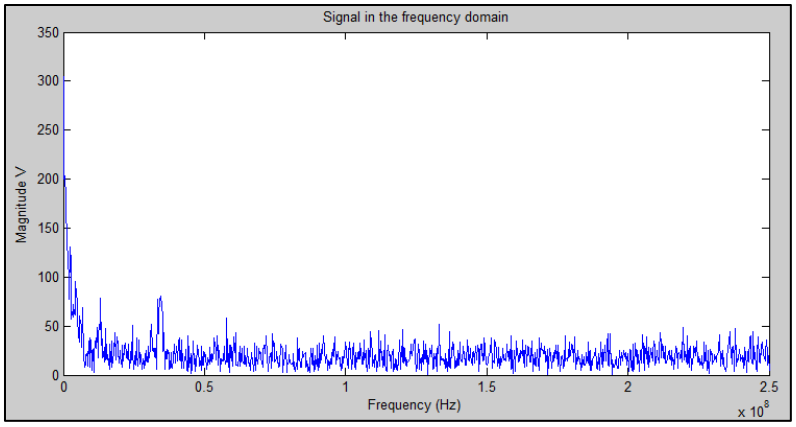

Figure 4: PD signal in the frequency domain

c) Time-Frequency domain

The signal in the wavelet domain combines the information from the time and frequency domains. Discrete Wavelet Transform (DWT) or Continuous Wavelet Transform (CWT) converts the signal into a Wavelet "by decomposing the time-series signal into a series of coefficients, where each series of coefficients represents a particular division in the time-frequency plane" [27]. The DWT of a signal can be computed by passing it through a list of filters. The signal is decomposed simultaneously with low-pass and high-pass filters. The low-pass filter produces the details coefficients and the high-pass filter produces the approximate coefficients as output [58].

$$y_{low}[n] = \sum_{k=-\infty}^{\infty} x\,[k]\; g[2n-k] \tag{2.1}$$

$$y_{high}[n] = \sum_{k=-\infty}^{\infty} x\,[k]\; h[2n-k] \tag{2.2}$$

The output of both filters are subsampled by two (i.e.) length of the signal is halved. These outputs are further processed by passing it again through the high and low pass filters for *n* levels [28].

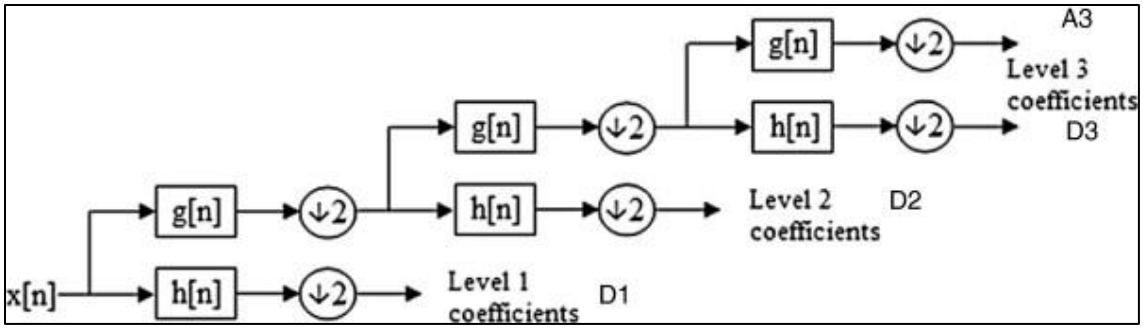

Figure 5: Decomposition of a signal through 3 levels using DWT [23]

2.1.1.2 PD Signal De-noising

De-noising the signal is one of the necessary steps for extracting the PD pulse as the PD pulses are very weak and the environment from where they are captured are generally noisy as well [26]. The de-noising of the signal can be done by decomposing the signal into a wavelet. This is done by choosing a level N and then computing wavelet decomposition coefficients of the signal for levels 1 to N [58]. This step is followed by selecting threshold detail coefficient for de-noising the signal. Finally, the de-noised time domain signal can be reconstructed from the wavelet by applying the Inverse Discrete Wavelet Transform (IDWT) [28].

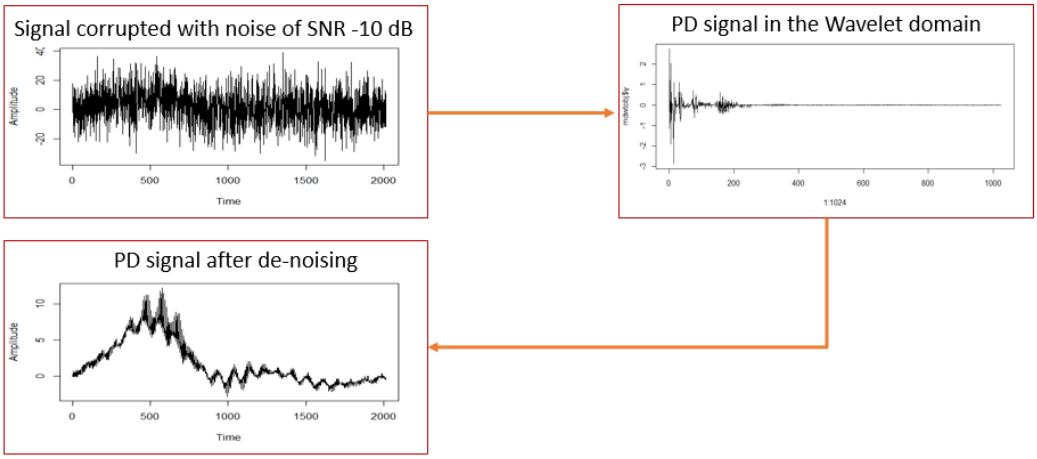

Figure 6: De-noising the PD signal using wavelet analysis

2.1.1.3 Phase Resolved Analysis

From PD measurement, some basic parameters like the individual PD magnitude (apparent change q) and their corresponding phase of occurrence (φ) with respect to the AC cycle can be calculated. Using these basic parameters, integrated parameters like peak discharge magnitude (q_m), average discharge magnitude (q_a), discharge current (i) and discharge rate (n) i.e. number of PD pulses per second can be calculated [25][61]. From these parameters, we can calculate the variation of any of these integrated parameters with respect to the phase φ. The four commonly used univariate distributions that are calculated for each phase window are **Error! Reference source not found.**[24]:

1. (n - φ): total number of PD pulses detected vs. φ.

2. (q_a - φ): average discharge magnitude vs. phase position φ.

3. (q_m - φ): peak discharge magnitude vs. φ.

4. (i - φ): average discharge current vs. φ.

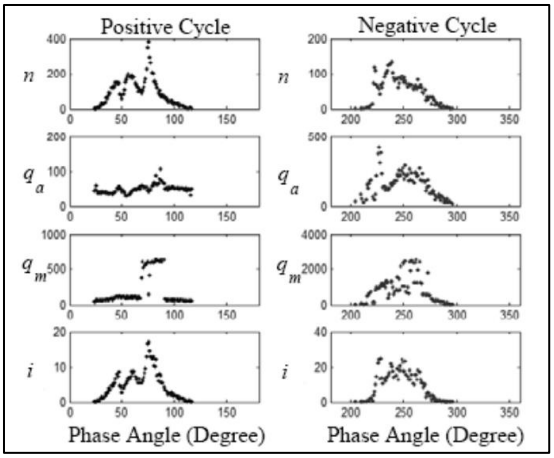

Figure 7: Phase resolved distribution in positive and negative half cycles [25]

2.1.2 Background on PD detection and classification

In this section, main concepts related to PD detection are presented. This includes background on the data acquisition methods, feature engineering processes and the different classification models.

2.1.2.1 Data Acquisition

Modern PD detection methods can be broadly classified into the following categories of their data acquisition methods [12]:

i. ***Analysis of Chemical Compounds:*** This method works by detecting the chemical compounds that are produced as a result of PD activity The two main methods that are based on this principle are [29]:

   a) ***High Performance Liquid Chromatography (HPLC):*** It identifies the products that are released due to the insulation breakdown of the transformer wall. These by-products are primarily glucose and degraded

forms of glucose. This test is performed by evaluating the oil samples at an offsite lab.

b) ***Dissolved Gas Analysis (DGA):*** This method measures the gas levels that are produced due to the breakdown of oil in oil-filled transformers. The gas levels vary due to equipment ageing and the component in which the PD occurs and is measured by taking oil samples from the transformer tanks [18].

ii. ***Optical Occurrence:*** It works by detecting light that is emitted as a result of PD activity. The optical signals are influenced by various factors like insulation material, temperature, intensity and pressure. These signals are measured using optical sensors [29].

iii. ***Acoustic Emission (AE):*** This is an online method for PD detection that picks up mechanical vibrations during PD activity, using devices like piezoelectric transducers, fibre optic sensors etc. It is extremely effective for identifying the location of the PD source due its immunity against electromagnetic noise [30].

iv. ***Ultra-High Frequency Method:*** This is also an online method for PD detection and uses high frequency sensors that can cover a wide range of frequency spectrum and can be divided into 2 types based on the frequency ranges that they cover; a) Very High Frequency (VHF) (3MHz – 300 MHz) b) Ultra High Frequency (UHF) (300 MHz – 3 GHz) [30].

v. ***Electrical Methods:*** It is mainly used for PD detection in cables and primarily uses inductive and capacitive sensors for capturing the PD signals. They are grouped into 2 categories, namely [29]:

1. ***Direct Probing***: Capacitive couplers are connected to the phase terminals of the transformers for measuring the PD activity.
2. ***Radio Frequency (RF) Emission Testing***: Antennas that are connected in the area of the transformers measure the PD activity.

Both the methods use a time domain recording device like a data storage oscilloscope to record the PD activity which is then detected using digital processing methods.

2.1.2.2 Feature Engineering

Feature Engineering is the process of extracting features using the domain knowledge of the data and then applying feature selection methods to reduce the dimensionality of the features. This section covers the main concepts related to Feature Extraction and Selection.

2.1.2.2.1 Feature Extraction

Feature Extraction is the process of extracting information from the raw data that is most suitable for applying a classification model. This section reviews some of the popular feature extraction methods. The type of features that are extracted differ with respect to the domain in which the signal is processed.

A survey of PD detection methods identified Pulse Characteristic and Signal Processing Tools as the most popular method for feature extraction in the time domain [11].

a) Pulse Sequence Analysis

Pulse Sequence Analysis (PSA) parameters which are captured from the time-series signal using Pulse Characteristic Tools are used as the features. These features work on the principle that the discharges that occur during PD, occur in a sequence, which are different

for each PD type. The two main parameters of PSA are dU (change in applied voltage between consecutive discharges) and dT (time difference between consecutive discharges) [31].

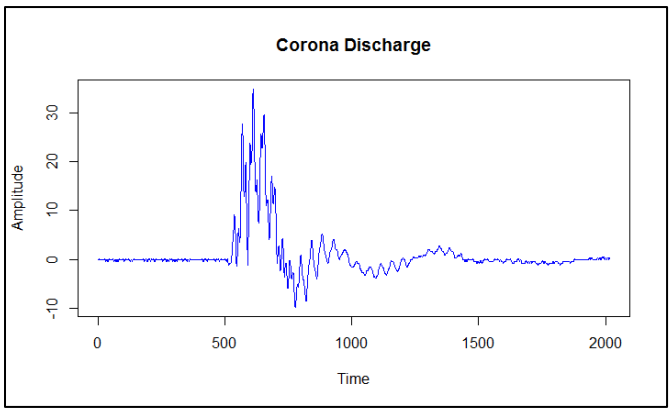

Figure 8: Corona discharge with primary pulse and decaying pulse

b) Pulse Shape Analysis

The shape of the pulse can provide additional information about the type of the PD to be detected, when compared with Pulse Sequence Analysis. In the latter method, the actual shape of the pulse is not considered, which is actually a good feature for defining the characteristic of the pulse [59]. Some of the features that describe the pulse shape are [13]:

i. *Amplitude:* The average amplitude of the signal.

ii. *Rise Time:* The mean rise time of the pulse.

iii. *Fall Time:* The mean fall time of the pulse.

iv. *Pulse Width:* The width of the pulse, which is calculated at 50% of the pulse amplitude.

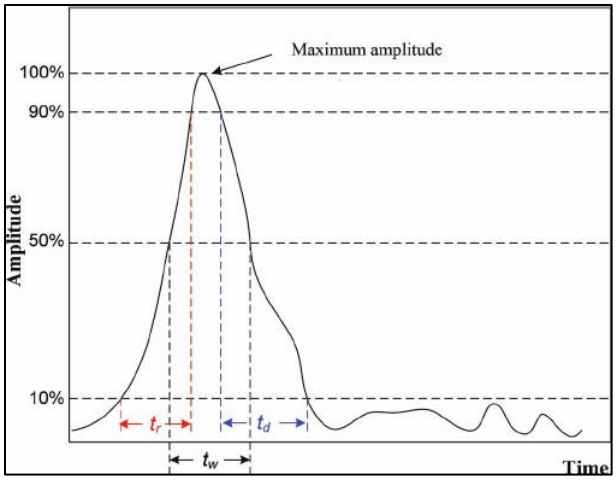

Figure 9: Pulse Shape Analysis parameters; tr-Rise time, td-Decay/Fall time, tw-Pulse Width [62]

c) Statistical Features

Statistical features can be extracted from the signal, irrespective of the domain. Some of the statistical features that can be extracted are [13]:

a) ***Mean μ:*** It is the estimate of the central value under the distribution clusters.

b) ***Variance σ^2:*** Standard Deviation σ is the estimate of the width of the distribution around its mean value. Variance is calculated as the square of Standard Deviation.

c) ***Skewness S_k:*** Measure of the degree of asymmetry of the distribution around the mean value

d) ***Kurtosis K_u:*** Measure of the degree of flatness of a distribution relative to the normal distribution.

$$\textit{Average (mean) value: } \mu = \frac{\sum_{I=1}^{N} x_i \, f(x_i)}{\sum_{i=1}^{N} f(x_i)} \qquad (2.3)$$

$$\textit{Variance: } \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2 f(x_i)}{\sum_{i=1}^{N} f(x_i)} \quad (2.4)$$

$$\textit{Skewness: } S_k = \frac{\sum_{i=1}^{N}(x_i - \mu)^3 f(x_i)}{\sigma^3 \sum_{i=1}^{N} f(x_i)} \quad (2.5)$$

$$\textit{Kurtosis: } K_u = \frac{\sum_{i=1}^{N}(x_i - \mu)^4 f(x_i)}{\sigma^4 \sum_{i=1}^{N} f(x_i)} - 3.0 \quad (2.6)$$

2.1.2.2.2 Feature Selection

Feature Selection is used for selecting a subset of the features that are best for the classification model. Principal Component Analysis (PCA) is the most commonly used method for feature selection in PD detection

a) Principal Component Analysis (PCA)

PCA is a "*multivariate technique that analyses datasets in which observations are described by several inter-correlated quantitative dependent variables*" [10]. PCA extracts the 'principal components' or the important information from the data. This is done by reducing the dimensionality while maintaining the maximum possible variation in the set. This is achieved by transforming the data into a lower dimensional space in the direction of maximum variance. It is calculated by eigenvalue decomposition of a data covariance matrix [8].

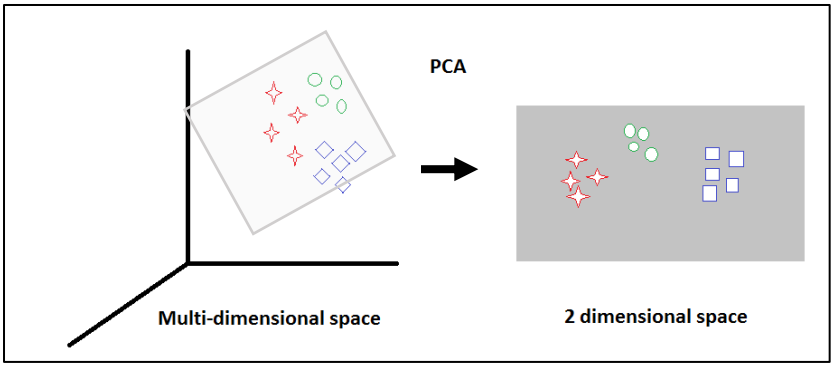

Figure 10: Transforming the data to a lower dimensional space to identify patterns using PCA.

2.1.3 Classification Model

This is the last step in data mining, where the types of PD are classified. In this section we will review three of the commonly used classifiers for PD detection.

a) Support Vector Machine (SVM)

SVM is the most commonly used classifier for PD detection. It works by trying to construct the best hyperplane between the samples of two classes [60]. It was essentially constructed for binary classification. However, it works for multi-class data as well by reducing it into a set of binary classification problems [17].

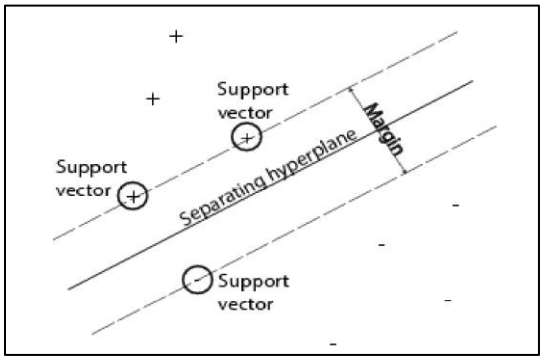

Figure 11: SVM Example [17]

b) K – Nearest Neighbour (KNN)

KNN is better suited for real world applications where the data is not linearly separable. It works by calculating the distance between each sample in the test data that is to be classified, and the training data that has already been classified. The class of the data to be classified is predicted by taking the majority of the class of the 'K' nearest neighbours. Some of the distance calculating measures are Euclidean, Manhattan, Hamming, City Block, Cosine and Distance Correlation [8][57].

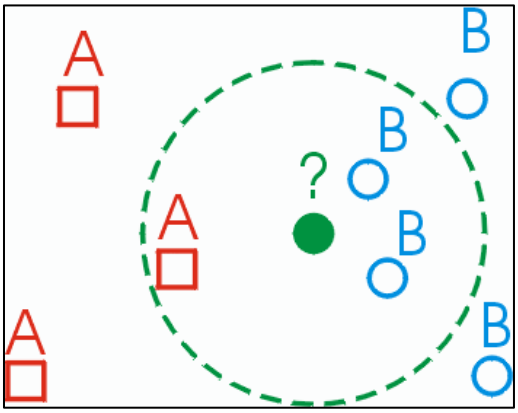

Figure 12: In 1NN, the green circle will be classified as B.

Source: [*http://www.fsanchezcv.com/*](http://www.fsanchezcv.com/)

c) Self-Organising Map (SOM)

SOM is an unsupervised neural network in which the data in a multidimensional space is mapped onto a 2D space, while preserving the original order of the inputs. It also forms clusters of similar data together. Therefore, it can be said that SOM reduces dimensions and groups similar items together.

Each node in the network is made up of a weight vector whose dimension is the same as that of the input vector. The input data is classified by calculating the distance between the vector of the input data and the weight vector of each node in the network [34].

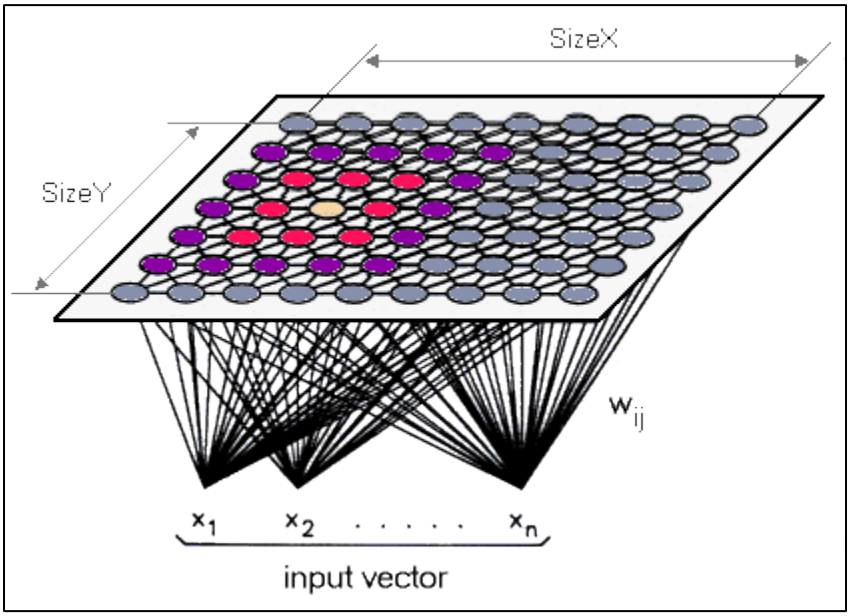

Figure 13: SOM - Mapping multidimensional data into 2D data [35]

2.2 Related Work

In this section, we will review some of the work that has been done in PD classification. The papers that have been reviewed are categorised based on their:

1. Data Acquisition System
2. Feature Engineering
3. Feature Classification

2.2.1 Data Acquisition

Data acquisition systems for PDs can be broadly categorised into: 1) Optical Methods 2) Chemical Methods 3) Electrical Methods 4) Acoustic Methods 5) Ultra High Frequency (UHF) Methods.

Optical sensors can detect PD by measuring the optical emissions from PD. This method enjoys immunity from electromagnetic interference and high sensitivity when compared with other conventional methods [30]. However, this method is not highly used as it is difficult to implement it in HV transformers due to the opaque nature of oil.

PDs can be detected using chemical analysis by breaking down the materials surrounding the PD activity into different chemical compounds. Chemical testing has two well-known tests, namely, 1) Dissolved Gas Analysis (DGA) 2) High Performance Chromatography Analysis (HPLC).

DGA is a widely used method for detecting PD activity by correlating different levels of released gases to a specific type of fault in the HV transformer. However, there are several debates as to whether the levels of dissolved gases really correlate to a specific type of fault. HPLC the other well-known test for chemical analysis measures the levels of glucose and degraded forms of glucose. However, HPLC and DGA suffer from the problem of not having standard values for mapping glucose concentration and their correlation to HV transformer fault [20].

Overall, chemical testing suffers from two main disadvantages. This form of testing does not provide any details regarding the location of the PD activity. Mostly, the transformers should be disconnected from the network to collect oil samples on which the

testing can be done. In the case of HPLC, the oil sample has to be sent out of the HV transformer site for analysis and can take a long time to obtain the results. These limitations lead us to search for a better PD detection method [17].

On the other hand, electrical detection records the electrical pulse that is contrived during any PD activity. The electrical detection method can be carried out both online and offline and is better than the chemical method in this aspect. However, this method is highly susceptible to noise. Using this method, it is difficult to distinguish noise from a PD signal, because the width of the PD pulse is short which is similar to the characteristics of 'noise'. This leads to false detection in online PD detection. This can be overcome by disconnecting the transformer from the network and connecting to an external power source to reduce the noise levels. However, this can result in a loss for the company [30]. In spite of this disadvantage, this form of detection is widely used in many power plants across the world.

Acoustic detection is similar to electrical discharge in the sense that it captures the signal related to the PD activity. The electrical method records the electrical signal of the PD, whereas, the acoustic method records the acoustic signal [37]. The acoustic signal is better than the chemical and electrical methods as it can identify the type of the PD in addition to the source of the PD activity, by attaching multiple sensors to different parts of the transformer. Knowing the location of the PD can help the plant technicians for repairing the transformers quickly. However, this method suffers from sensitivity as the acoustic signals are very weak. Hence, the sensors must be highly responsive to detect even minor changes to the amplitude of the signal identify a PD activity [30].

UHF detection method is another online PD detection method in transformer. It uses UHF sensors, RF sensors or antennas to measure the PD activity. This method enjoys

immunity against electrical noise and can be used for detecting PD type and location. However, this method is costly and its design is complex in nature [37]. In this thesis, we use the UHF detection method as it overcomes the critical disadvantages like sensitivity and supports online detection of PD activity.

2.2.2 Feature Engineering

The features for PD detection system are often obtained by processing the signal in the 1) Time domain 2) Frequency Domain 3) Wavelet Domain.

The pulse of a signal is defined as a "rapid change in the amplitude of a signal to a higher or lower value than the baseline value followed by a rapid return to the baseline value" [31]. The Pulse Sequence Analysis (PSA) and Pulse Shape Analysis methods can be used for extracting features from the signal in the time domain.

In PSA, the space charges that are created after each pulse in the PD source location modify the electric field, which affects the delay time between consecutive pulses. This method works on the principle that PD data is obtained as phase-resolved plots, where phase-resolved means that the PD pulses are time domain signals that are superimposed over the applied sinusoidal voltage. The PSA correlates signal from one source only, so this method will fail if the pulse is superimposed from multiple sources. In addition, these type of features are highly sensitive to noise [32].

On the other hand, Pulse Shape Analysis can be applied for the signal in the time and frequency domain [32]. However, this method requires advanced analysis systems for classification in real time.

The signal time domain is highly susceptible to noise and these features usually suffer when the signal is subjected to highly noisy conditions. In this case, features can be extracted from the signal in the frequency domain which is less susceptible to noise.

The signal in the time domain is converted to the frequency domain by applying DFT or FFT. Each sample in the frequency spectra are then used as a feature. The features can be normalized to reduce the influence of the electric field stress on the signal. As opposed to the approaches where the raw signal in the frequency domain is processed as is, some statistical parameters can be extracted as features [56][57]. These features are described in the *Section 2.1.2.2.1*.

If the entire feature spectra is used as the feature vector, it leads to complex analysis which is not good for classification of PD type in real time. To overcome this, PCA which is a feature selection method can be applied to the signal in the time and frequency domain. It was applied to the features derived from the frequency spectra by Pattanadech et.al [56]. On the other hand, Firuzi et. al applied it to the raw signal in the frequency domain, but used it as a feature extracting method [16].

Some papers process the signal in the wavelet domain, as this domain contains the information about the time and frequency components in the signal simultaneously. This method is good for signals with irregular and transition features which is similar to the characteristics of the PD signals. However, selecting a mother wavelet and number of levels for decomposition are quite difficult to choose. It is also computationally intensive which is not good for PD detection system that have to possess low processing time [38].

In this thesis, we use the frequency domain to process the signals as it offers a balance between processing the signal in the time and wavelet domain. It is not as sensitive as the time domain and doesn't require complex processing which the wavelet domain demands.

2.2.3 Feature Classification

Various classification models have been used for classifying the features that are extracted from the signal in the time or frequency domain. The classification models that usually did well are, SVM, KNN and SOM. These classifiers were introduced in the last section. PCA has often been used in conjunction with other classifiers and has usually provided good results. This ascertains that PCA produces separable representation that is suitable for classification.

The performance of a classifier depends on the data and there are no rules to map the best classifiers for a given dataset. Multiple classifiers are usually run on the data to identify the classifier that best suits our data.

Chapter 3

3. Methodology

This chapter will describe the methodology that has been followed to predict the type of the PD activity. The signal has been processed in the time and frequency domains using different features, classification models and validation methods.

3.1 Data Processing

The raw PD signal that was generated in the lab is a time-series signal. For processing the signal in the frequency domain, this time domain signal is converted into the frequency domain by applying Fast Fourier transform (FFT).

3.1.1 Fast Fourier Transform

The FFT decomposes an N sample time domain signal into N time domain signals, each of which is made up of a single sample. The frequency spectra corresponding to each of these N time domain signals is then computed. Finally, the N spectra are fused to form one frequency spectrum. The FFT is calculated exactly the same as DFT but is much faster than the latter. Let $x_{0,}\, x_{1\ldots..XN-1}$ be complex numbers. The DFT can be computed as:

$$X_k = \sum_{n=0}^{N-1} x_n \, e^{i2\tau kn/N} \;, \quad where\ k = 0,1 \ldots . N-1 \tag{3.1}$$

$e^{-2\pi ik/N}$ is known as the twiddle factor. It is a "rotating vector", which rotates in increments according to the number of samples, N.

The FFT algorithm can be represented by a Butterfly Diagram. The simplest butterfly, which is made up of two inputs A= X_0 and B= X_1 and 2 is depicted in *Figure 14*.

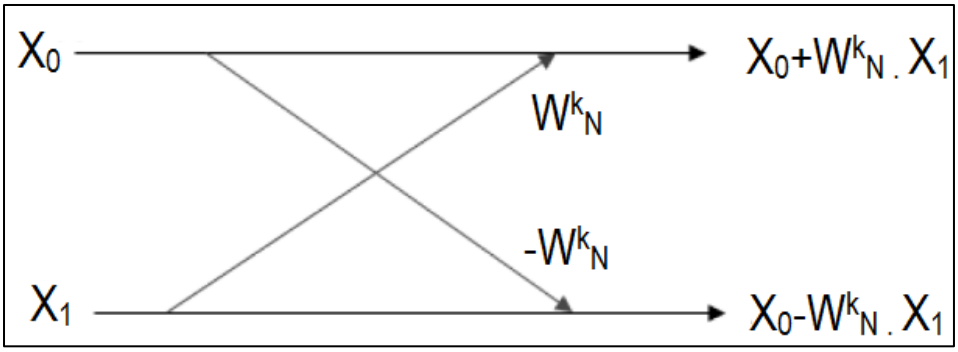

Figure 14: Simple FFT butterfly

The number of stages required for computing an N-point FFT is $\log_2 N$. This is the reason why FFT sizes are usually in powers of two, and is also known as the "Radix 2 FFT". For an 8-point signal, we require $\log_2 8$ stages (i.e.) 3 stages to compute the FFT . The overview of these stages is depicted in **Error! Reference source not found.**.

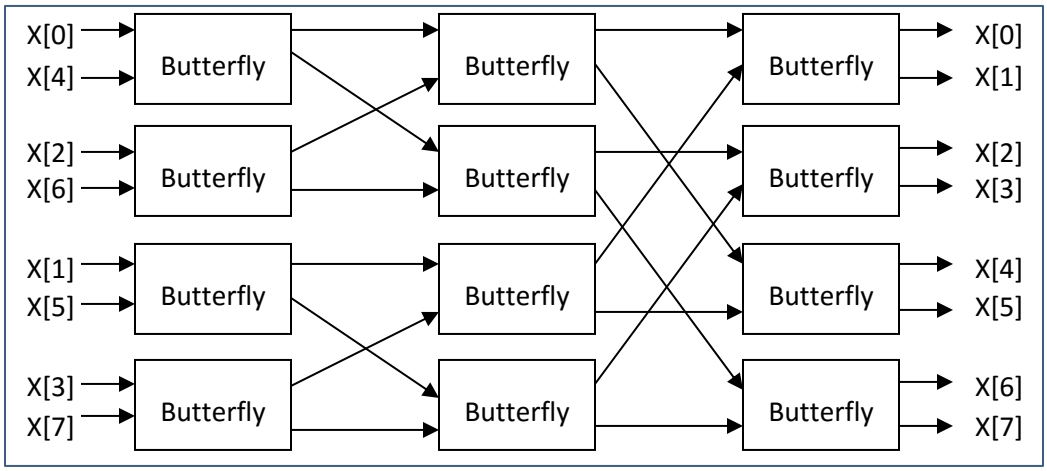

Figure 15: 3 stages of FFT of an 8-point signal

The function *fft()* in R and Matlab can compute the value of X_k for each sample in the signal, by providing the entire signal as a vector of values. The *fft()* function returns the signal in the frequency domain which is a vector made up of complex numbers. This

complex vector is converted into a vector of real numbers by applying the *abs()* function which calculates the absolute value of the complex number.

3.2 Feature Engineering

We know that a feature is a piece of information that is potentially used for prediction and that feature engineering is the process of designing features, which are fed as inputs to machine learning algorithms. In this section, we will explain the feature engineering methods that were followed for the PD signal in the time and frequency domains.

3.2.1 Feature Extraction

Feature Extraction is the process of creating a subset of new features by combinations of the existing features (i.e.) we build a new set of features from the original feature set. It involves a transformation of the features, which is often not reversible because some information is lost during the transformation.

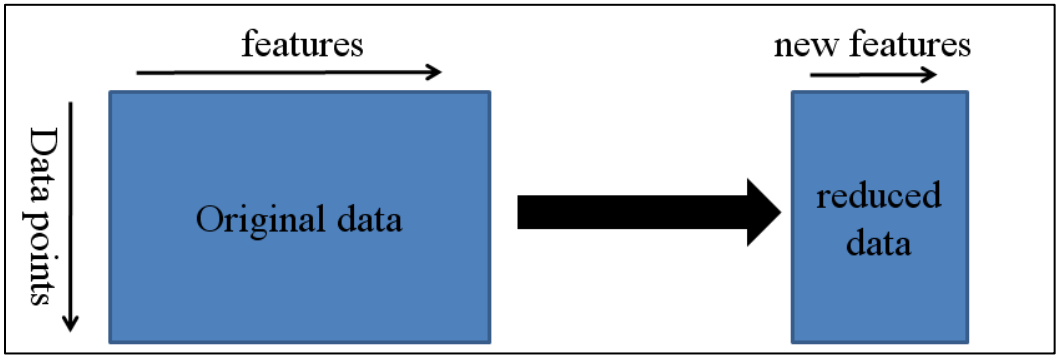

Figure 16: Feature Extraction process

Feature Extraction in this thesis is done on the signal in the time domain. We extract features, which provide information regarding the peaks in the dataset. Some of the features

that we have extracted from the signal in the time domain are loosely based on the Pulse Shape Features that were described in *Section 2.1.2.2.1*.

3.3 Build Classification Models

Once we have defined our problem and prepared the data for processing, we need to apply machine-learning algorithms to the data in order to solve our problem.

3.3.1 Naïve Bayes Classifier

It is a probability-based classifier that applies Bayes' theory by assuming that there is independence between the features. In other words, the effect of a feature on a given class is considered to be free from the influence of other features. It uses the maximum-likelihood estimation to approximate the feature learning process in the training stage. After this, it computes the posterior probability of each sample and assigns the sample to the class with the highest posterior probability [39][44].

Given a sample **x**, the classifier will predict that **x** belongs to the class with the highest posteriori probability [36] (i.e.) **x** is predicted to belong to class C_i if and only if,

$$P(C_i \mid \mathbf{x}) > P(C_j \mid \mathbf{x}) \text{ for } 1 \leq j \leq m, j \neq i. \quad (3.2)$$

3.3.2 J48

It is a tree-based model, which is one of the most simple classification models. In this classifier, the rules are optimally inferred from decision trees. The decision tree model can be constructed without complicated computations and this makes them computationally efficient. However, the tree model sometimes declines to include all discriminant features, which in turn affects the detection accuracy. This classifier is

appropriate for the applications of PD detection where quick diagnosis and detection is needed [40].

3.3.3 Bagging

Bagging is a "bootstrap" ensemble method that learns many classification models, each with only a part of the data. The bootstrap sample is obtained by creating random subsets of the data by sampling. For training a dataset *D* of size *n*, bagging generates *m* new training sets, each of size *n'*, by sampling *D* uniformly (with replacement). Bagging then creates k bootstrap samples and trains k classifiers on each bootstrap sample. To test the model, each classifier votes on every instance the test set. The majority of the votes decide the class of the instance [41].

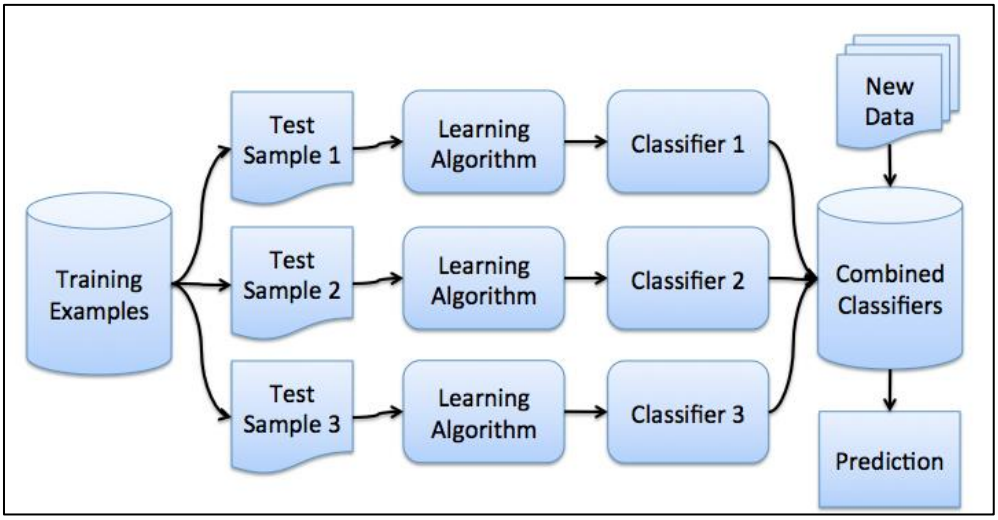

Figure 17: Overview of the Bagging Algorithm

Source: http://cse-wiki.unl.edu/wiki/index.php/Bagging_and_Boosting

3.3.4 Multilayer Perceptron (MLP)

An MLP is a "network of simple neurons called perceptrons. The MLP accepts multiple values as inputs by forming a linear combination of its input weights" [33]. It then computes as output, the class of the instance using a non-linear activation function. Mathematically this can be written as :

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \varphi\left(w^T x + b\right) \tag{3.3}$$

Where,

- w is the vector of weights,
- x is the vector of inputs,
- b is the bias,
- φ is the activation function.

A typical MLP network contains "an *input layer* made up of source nodes, one or more *hidden layers* of computation nodes and an *output layer* of nodes" [42]. *Figure 18* shows the flow graph of the perceptron.

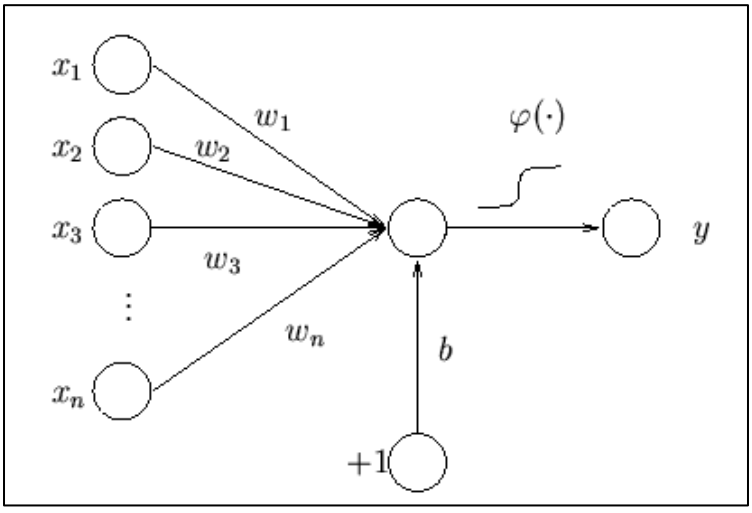

Figure 18: Flow graph of the perceptron

Source: http://users.ics.aalto.fi/ahonkela/dippa/node41.html

3.3.5 Sequential Mining Optimization (SMO)

The SMO is an algorithm that solves the Quadratic Programming (QP) problem which is faced during training of the SVM classifier. The SVM classifier is described in *Section 2.1.3.*

Training the SVM requires large matrix operations that are time-consuming and very slow. SMO is a simple iterative algorithm that can solve these problem without any extra storage or optimization steps [41] [43]. SMO does this by breaking the problem into a series of smaller problems which are then solved analytically [44].

3.3.6 Random Forest (RF)

RF is a popular and efficient algorithm that is based on the principle of model aggregation. It fits many classification trees to a dataset and then combines predictions from all the trees. Each tree is considered to be an individual classifier and has its own weight. At each node in the tree, a subset of features are chosen randomly and the best split for the tree is calculated from this subset [45]. The final classification of a sample is decided based on the majority of the classification output of all the trees [39].

3.3.7 JRip and OneR

It is an inference and rule-based learning method that tries to come up with propositional rules which can be used to classify elements [46]. It has a lower computational cost than other rule learning methods. OneR is another simple classification method that uses a single feature to classify the elements. In other words it uses only one-rule (OneR) to classify the elements.

Both JRip and OneR are computationally efficient. However, like the J48 method, they may not use all of the discriminant features which negatively impacts their classification accuracy [39].

3.3.8 Logistic Regression

It is a method for fitting a regression curve y = f(x) where x is a set of features and y is the predictor variable. The logistic function, also called the sigmoid function is an 'S' shaped curve, that predicts the probability (value between 0 and 1) of an observation belonging to a given class. [47].

$$\sigma(z) = 1/(1 + e^{-z}) \tag{3.4}$$

The logistic model consists of a vector β in a N-dimensional feature space. A point x in the feature space is projected onto this vector to convert it into a real number z in the range -∞ to +∞ [48].

$$z = \alpha + x_1\beta_1 + \ldots + x_N\beta_N \tag{3.5}$$

This value of z is then mapped onto the range of 0 to 1 using the sigmoid function, which is measure of the probability of class membership.

3.4 Regularization

In Machine Learning and Statistics, a common task is to create a model to fit a set of training data. This model is then used for making predictions to classify new data. However, the model that is created maybe too specific (overfit) or too general (underfit) for the dataset. The degree of fit for any model is measured using the loss function. The loss function can be generally defined as ***L(actual value, predicted value)*** that is applied

during the learning process to learn from mistakes that are done during predictions [52]. Consider a dataset D = { $x_1, x_2 \dots x_n$ } where each x_i represents an attribute of the dataset. Let us consider that the complexity of the model can be controlled by a complexity index k, $1 \leq k \leq k_{max}$. Consider a modelling task to fit a conditional regression model,

$$y = g(z_k) + e \tag{3.6}$$

Where y is the class variable and z is a subset of size k of the remaining variables in the vector. *Equation 3.6* shows that the class variable y can be predicted using a model and a loss function.

Hence, the ideal model can be defined as the model that minimizes the loss function. This ideal model can be expressed generally as ***Model = argmin*** $\sum$ ***L(actual, predicted(Model))*** [50].

Regularization is a technique that will avoid the problem of overfitting by making the model more general. The principal idea behind regularization is that, models that overfit the data are complex by probably being built using too many attributes. In order to overcome this, a regularization term is introduced to avoid overfitting [52]. This can be expressed mathematically as [50]

$$\textbf{\textit{Model = argmin}} \sum \textbf{\textit{L(actual, predicted(Model))}} + \lambda \textbf{\textit{ R(Model)}}$$

Regularization creates a more general model by minimizing the Residual Sum of Squares (RSS) (loss function) that is added to the L1 or L2 norm function. The L1-norm works by minimizing the sum of the absolute differences (S) between the target value (Y_i) and the estimated values ($f(x_i)$) of the coefficients or attributes in the dataset [53].

$$S= \sum_{i=1}^{n} |y_i - f(x_i)| \tag{3.7}$$

On the other hand, the L2-norm is also known as Least Squares. It works by minimizing the sum of the differences (S) between the target value (Y_i) and the estimated values ($f(x_i)$) of the coefficients of the attributes [52].

$$S= \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{3.8}$$

In the usual linear regression model, given N predictors $x_1 \ldots . x_N$, the response y is predicted by,

$$y = \beta_0 + x_1\beta_1 + \ldots + x_N\beta_N \tag{3.9}$$

Where the model fitting procedure produces the vector of coefficients $\beta = (\beta_1 + \beta_2 + \cdots + \beta_N)$ The regularization thus tries to solve the following equation [50]:

$$\min_{\beta_0, \beta} \frac{1}{N} \sum_{i=1}^{N} w_i \, l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha)/2 \, \|\beta\|_2^2 + \alpha \|\beta\|_1 \right] \tag{3.10}$$

Where, l(y, η) is the negative log-likelihood, β_0 is the intercept, y_i is the response variable, x_i is the i^{th} predictor variable and α is the mixing parameter $0 \leq \alpha \leq 1$ [51].

Regularization techniques can be of three types, namely:

1. ***Ridge:*** It performs shrinkage of the coefficients and avoids overfitting by penalizing large coefficients using the L2-Norm. This method seeks to minimize $RSS(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2$ where λ is the regularization parameter, β is the estimated coefficients and p is the number of coefficients. This method solves *Equation 3.10* by setting α=0 which is the ridge penalty.

2. ***Lasso:*** It performs selection and shrinkage of the coefficients and avoids overfitting by penalizing large coefficients using the L1-Norm. This method seeks to minimize: $RSS(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$, where λ is the tuning parameter, β is the estimated coefficients and p is the number of coefficients. This method solves *Equation 3.10* by setting α=1 which is the lasso penalty.

3. ***Elastic Net:*** It is similar to the lasso technique and performs selection and shrinkage of the coefficients and tries to "retain the big fish" or the most important attributes. It mixes the penalties of Lasso and Ridge and sets $0 \leq \alpha \leq 1$ which is the elastic net mixing parameter in *Equation 3.10 [*54].

3.5 Classification Model Testing

Once the classification models have been created, we have to test the model to check if it is a reasonable representation of the system. For checking the accuracy of a model, the dataset is divided into the train and test sets, where the train set is used for building the model and the test set is used to examine "how well the model has been trained". The model testing methods that we have used in this thesis are:

1) ***Percentage Split:*** The dataset is split N% into train and (100-N)% test sets. This method is also called hold-out.
2) ***Cross-Validation:*** This method involves splitting of the dataset into a number of equally sized groups of observations (folds). The model is then trained on all the folds except one. This one left-out fold is used for testing the model. This process is repeated until all the folds have acted as the test set. The performance measures are averaged across all folds to finally evaluate the model.

3) ***Custom testing:*** In this form of testing, the training set is made up of the clean dataset and the test set is made up of the noisy dataset, which is obtained by adding noise of various Signal to Noise Ratios (SNR) to the clean dataset. The train and test datasets are divided into folds. For example, if the train and test datasets are made up of 10 folds each, then 10 files are created. Each of these files are created by taking 9 folds of the train set and the 1 left out test set is obtained from the test set. This is repeated until all of the folds in the test set have been added at least once to any of the 10 files. Once the 10 files are obtained, the model is evaluated on each of these 10 files, by taking 90% as train data(clean dataset) and the remaining 10% as test data (noisy dataset). The final accuracy of the model is calculated by averaging the performance measures that are obtained by running the model on these files.

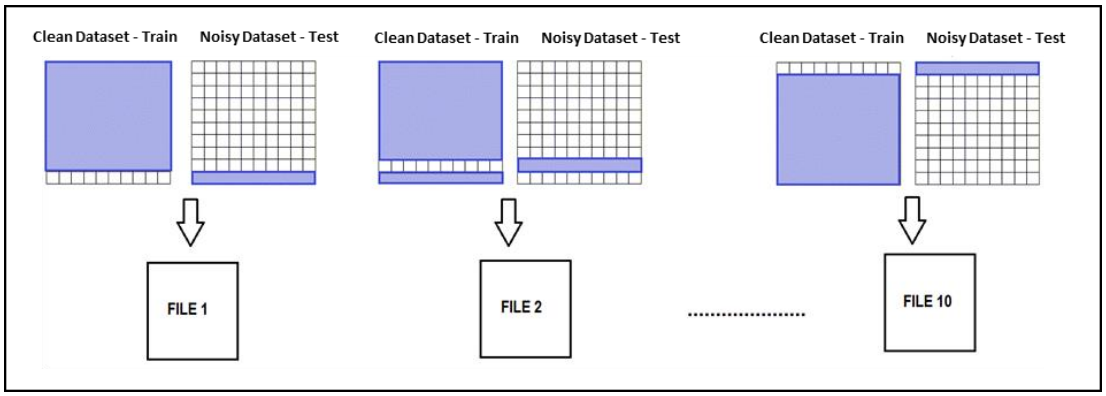

Figure 19: Creation of the 10 files

This type of pessimistic testing that we have used not only validates the accuracy of the system when posed with different levels of noise but also validates the accuracy when tested with new samples that haven't been encountered in the training set.

Chapter 4

4. Experimental Evaluation

4.1 IEC 61850 Merging Unit

The Merging Unit (MU) is the computing power of the grid system and is equipped with Digital Signal Processing (DSP) and communication capabilities for continuous monitoring of the power system. Smart Grid functions like self-healing and asset management require communication between the MU and the transformers. The MU communicates using the IEC 61850 protocol, which allows high-speed Ethernet communication between the various system devices and provides a standardized configuration language and data model [49].

This MU is powerful and can monitor and detect impending faults like PDs. Once PDs are detected, the MU can classify the type of PD and can communicate this information to the Supervisory Control And Data Acquisition (SCADA) system. This makes the MU perfect for anticipating PDs and reporting these impending faults to the operator at the SCADA system, so that remedial action can be taken.

Power system equipment like transformers and circuit breakers are also known as Intelligent Electronic Devices (IEDs) as they are microprocessor-based controllers. The communication between the IEDs and MUs are via Generic Object Oriented Substation Events (GOOSE) messages, as they provide a fast and reliable mechanism by multicasting or broadcasting event data to system devices in a network [49][55].

4.2 Experimental Setup

The data for our experiments has been generated from our laboratory setup in UAE. *Figure 20* shows the experimental setup. The smart MU was implemented using Robin Z530, which is a single board computer with a COM Express carrier board and a data acquisition system.

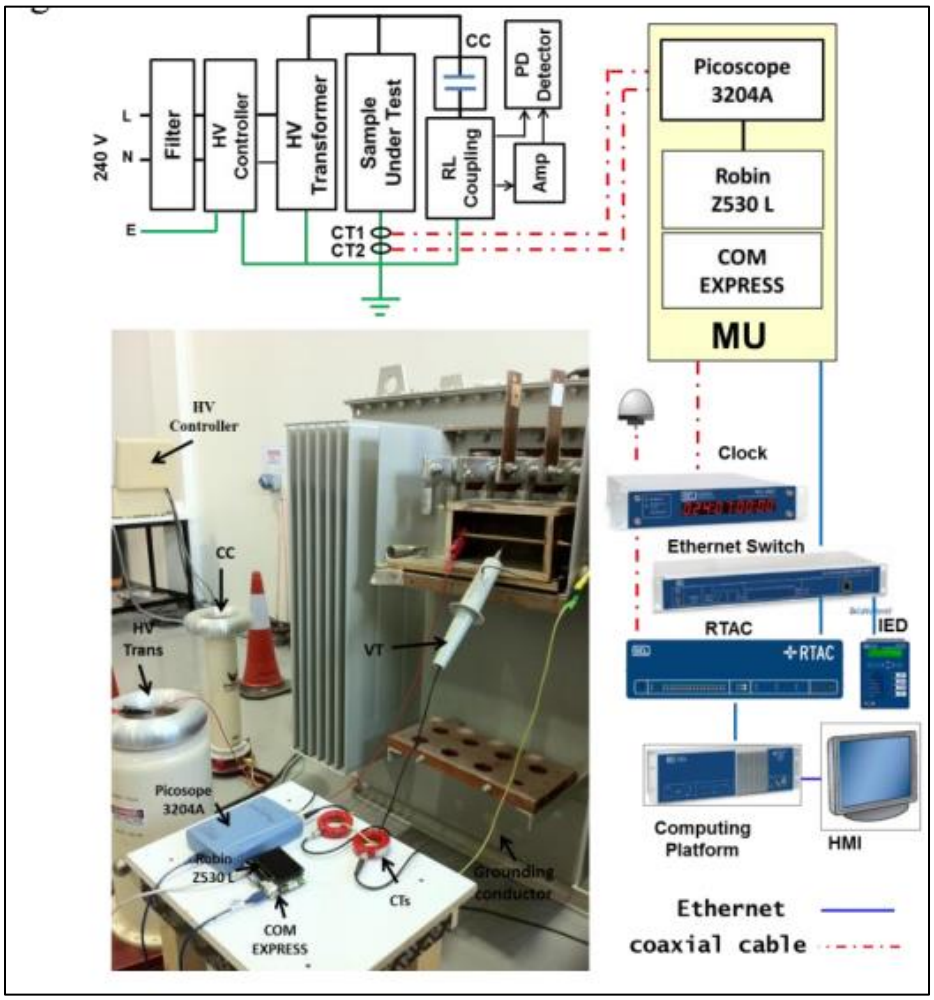

Figure 20: Experimental Setup [49]

The main components of the setup are [48]:

1) ***HV Test Equipment***: Used for generating controllable PDs during insulation breakdown.
2) ***Picoscope:*** It is a data acquisition system with a bandwidth of 60Mhz and sampling rate of 500 MS/s.
3) ***Real-Time Automation Controller (RTAC):*** It can convert data between multiple protocols. It also manages and communicates with the IEDs.
4) ***IEDs:*** The MU is integrated with IEDs to emulate a primary substation that can control, protect and communicate with a power system.
5) ***Ethernet Switch:*** All the IEDs are connected using a 24-Port Ethernet switch.
6) ***Satellite Synchronized Clock:*** The IEDs are synchronized using this clock.
7) ***SCADA-HMI System:*** The MU reports PD activities directly to the SCADA system.

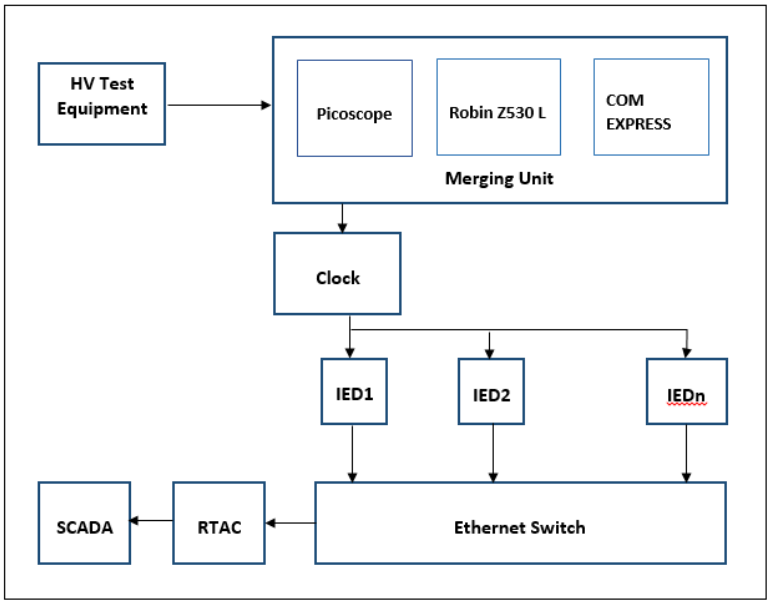

Figure 21: Simplified Block Diagram Showing the Major Components of the Setup

Two wide-bandwidth Current Transformers (CTs), CT-A (200 Hz – 500 MHz) and CT-B (1.0 Hz – 20MHz) and are used for capturing current signals over a wide frequency range (1.0 Hz – 500 MHz).

4.3 **Overview of our PD monitoring system**

The overview of a general PD monitoring system is depicted in ***Error! Reference source not found.***. The PD system is made up of the following main components:

1. Transformer
2. CT sensors
3. MU
4. SCADA system

The PD system uses CT sensors to record the signals from transformers. These raw signals are sent to the computing power of the system, that is the IEC-61850 MU. This MU is made up of:

1. ***Computing unit*** – It is responsible for feature extraction and classification. It can predict the type and severity of the PD activity.
2. ***Storage unit*** – It stores the raw signal and features for a period of time so that they can be accessed by a human operator for further details.
3. ***Communication Interface*** – It acts as an interface between the MU and the SCADA system.

The SCADA system displays the type of PD activity to a human operator at the SCADA who can request for more details like the signal and the features that determined the type of the PD for some period of time.

In this thesis, we will focus on the computing unit of the MU that is responsible for

feature extraction and classification of the PD types.

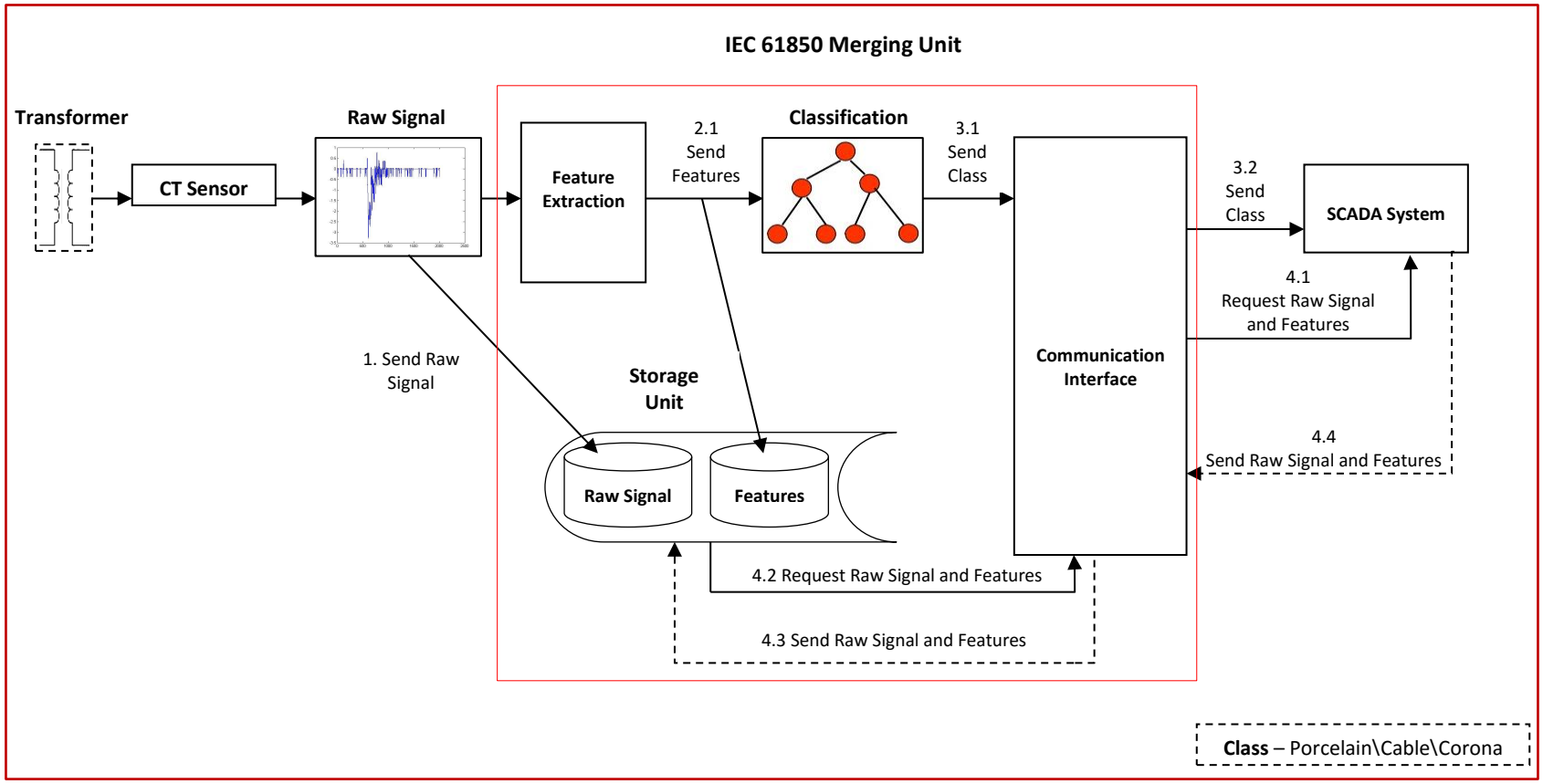

Figure 22: Overview of our PD monitoring system

4.4 Challenges in Monitoring Impending Faults

The objective of the MU is to detect impending faults, as they could potentially lead to a flashover. However, this is quite a difficult task because of the following reasons:

1) The low magnitude and intermittent nature of the PD activity makes it hard to be detected. In addition, the IEDs will not be able to detect these PD activities until the magnitude of the PDs reach full scale.
2) The non-stationary features of the impending faults vary depending on the type of the insulation material and degradation level, sensor types and their locations.
3) The bandwidth of the measuring instruments (CTs) should be compatible with the bandwidth of the MU.

4.5 Description of the Dataset

The data that we obtained from the laboratory is made up of 285 observations, which are equally distributed among three PD classes, namely, Porcelain, Cable and Corona. These three types correspond to the common PDs that can affect the transformer externally. Each PD class refers to a PD activity that may occur in the different parts of the transformer.

1. ***Porcelain:*** This type of PD occurs due to damage to the porcelain bushing.
2. ***Cable:*** It occurs due to faulty cables and cable accessories.
3. ***Corona:*** This occurs due to transient gaseous ionization when the voltage stress exceeds a critical value.

Each of these 285 observations is made up of 2016 samples, which shows the variation of the magnitude of the signal over time. For the purpose of this research, two CTs, CT-A and CT-B monitored the health of the transformer. The data that is received from CT-A (200 Hz-500 MHz) which is the higher frequency sensor, is referred to as Channel A

whereas the data from the lower frequency sensor CT-B (1.0 Hz – 20 MHz), is referred to as Channel B. The sampling rate is 500 Ms/s.

Figure 23 and *Figure 24*, shows the difference in the shape of the signal between the 3 types of the PD classes from CT-A.

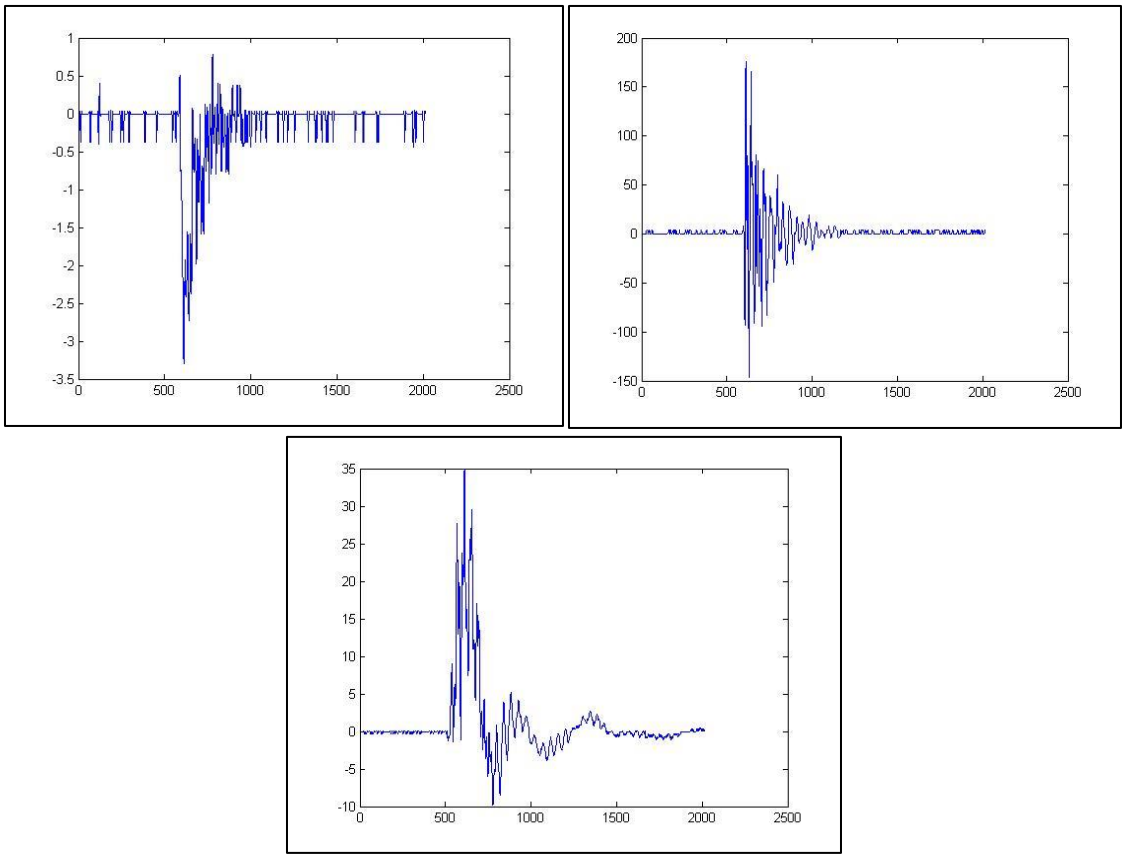

Figure 23: PD activity measured by CT-A. *Clockwise:* a) Porcelain b) Cable 3) Corona

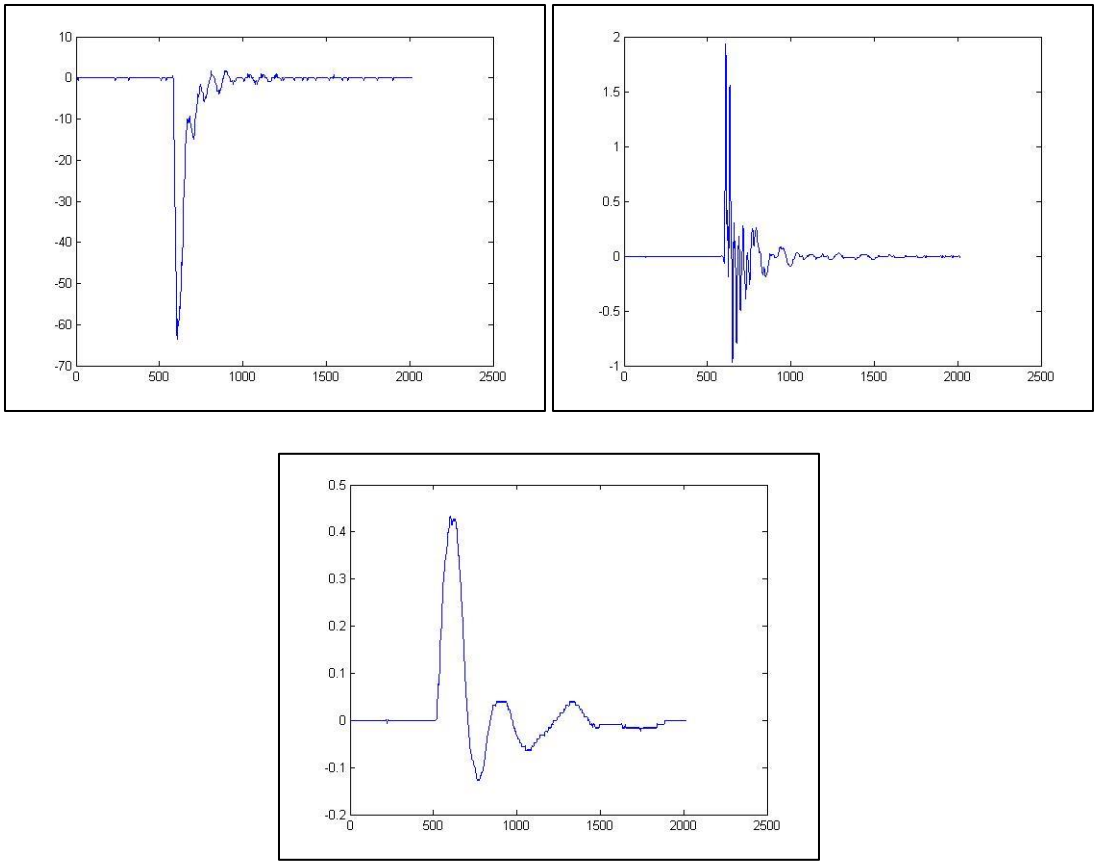

Figure 24: PD activity measured by CT-B. Clockwise: a) Porcelain b) Cable 3) Corona

4.6 Feature Engineering

We processed the signal in the time and frequency domains. We first started with working on the signal in the time domain. From the signal in the time domain, we extracted two types of novel features, namely:

1) ***Simple Features:*** This type of features extracted information about the signal by measuring the

    - Average of the signal
    - Maximum magnitude (highest positive peak)
    - Minimum magnitude (lowest negative peak)

- Total number of positive peaks
- Total number of negative peaks/valleys.

The Simple features extracted these 5 features from the 2016 sampling points of each observation.

2) ***Shape Features:*** On the other hand, unlike the simple features, the shape features were aimed at extracting information about the PD peak in the signal and not the signal itself. A feature vector containing details about the highest 10 positive peaks and lowest 10 negative peaks were extracted.

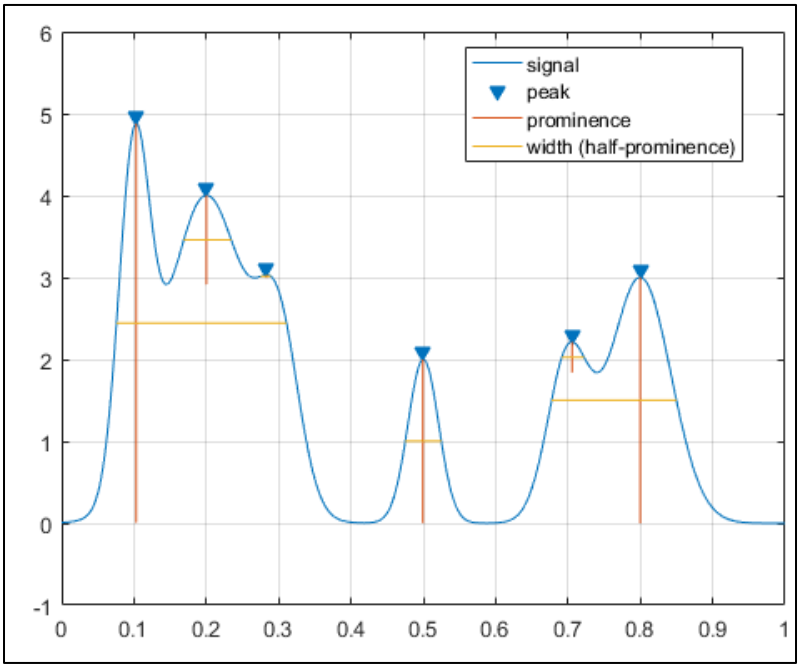

Figure 25: Measurement of peak magnitude, prominence and width in Matlab

Source: https://www.mathworks.com/help/signal/ref/findpeaks.html

The details that were extracted about each peak are

- Location
- Magnitude
- Prominence: This measures how much the peak stands out due to its height and its location relative to other peaks.
- Width: Measure of the width of the peak at half the prominence.

The shape features can be thought of as a variation of pulse shape features. *Figure 25* shows the measurement of the shape features in Matlab. The Shape features extracted these 4 features from the 2016 sampling points of each observation. These 4 features were extracted for the 10 highest peaks (10 x 4) and 10 lowest valleys (10 x 4). Hence, a total of 80 features were used to represent the 2016 sampling points.

The simple combined with the shape features did very well in the absence of noise. Both sets of features gave us almost perfect results when a Cross-Validation of 10-folds was executed on the datasets from both CTs. *Figure 26* and *Figure 27* show the classification accuracies of clean PD signals using simple and shape features, respectively. Five different classification models including Naive Bayes, Bagging, J48, SMO and Multi-perceptron were used to examine the effectiveness of these features in the absence of noise.

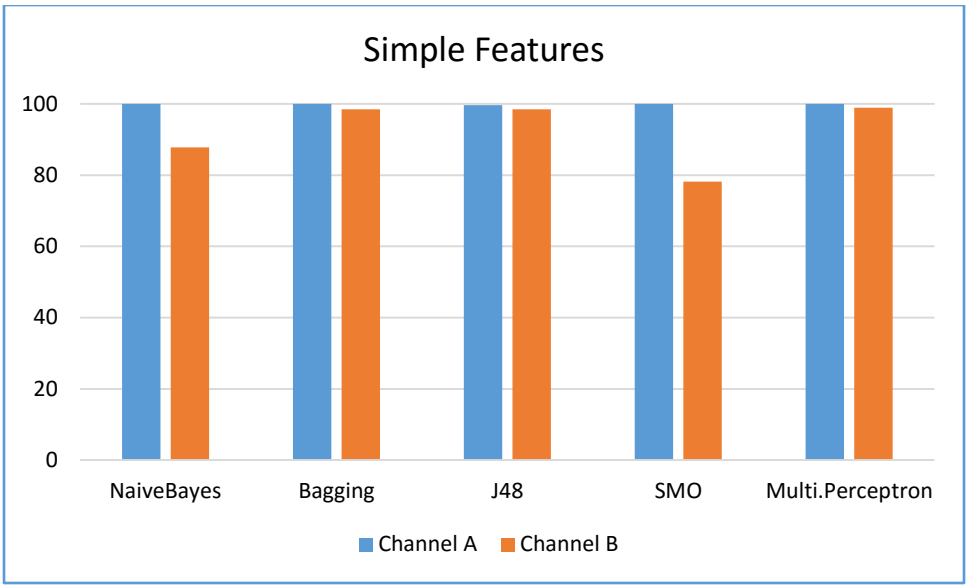

Figure 26: Classification accuracy using simple features extracted from Clean Data

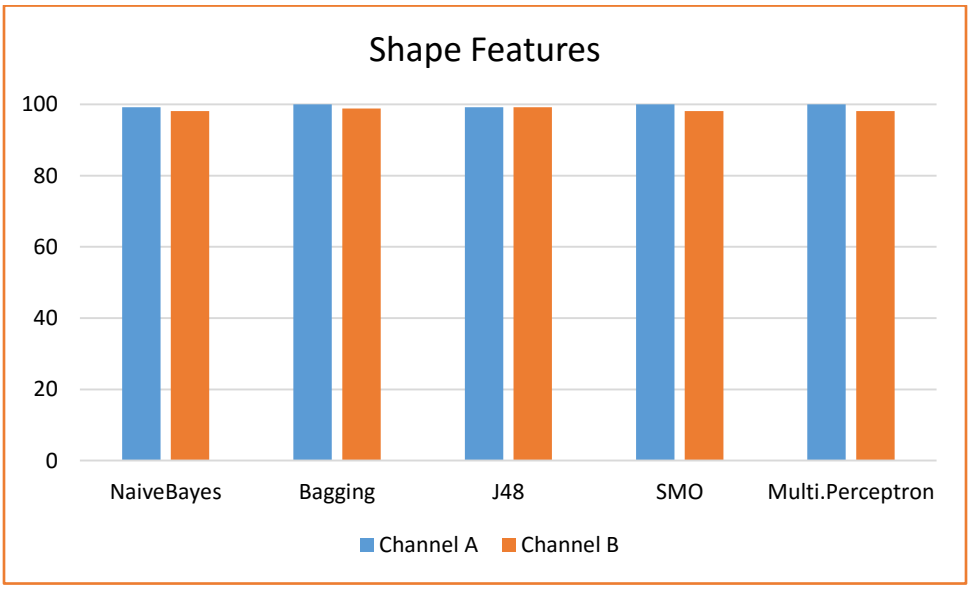

Figure 27: Classification accuracies using shape features extracted from Clean Data

However, when noise of SNR -10 was incorporated into the clean signal, the accuracy dropped drastically. *Figure 29* and *Figure 30* show the accuracy of the simple and shape

features when five classifiers are run on the noisy data. The raw (clean) data was used as the training set and the noisy data of SNR -10dB was used as the testing set.

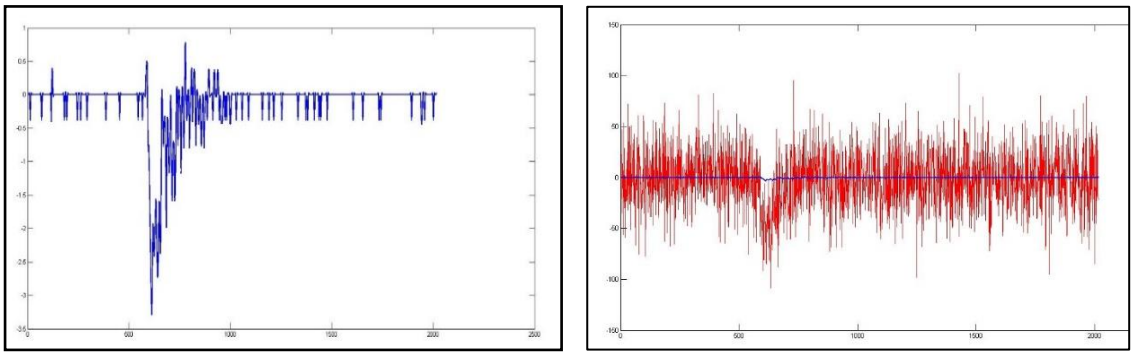

Figure 28: Left: Clean Signal. *Right:* Noisy signal (in red) acquired when SNR of -10dB is added to the clean signal (thin line in blue)

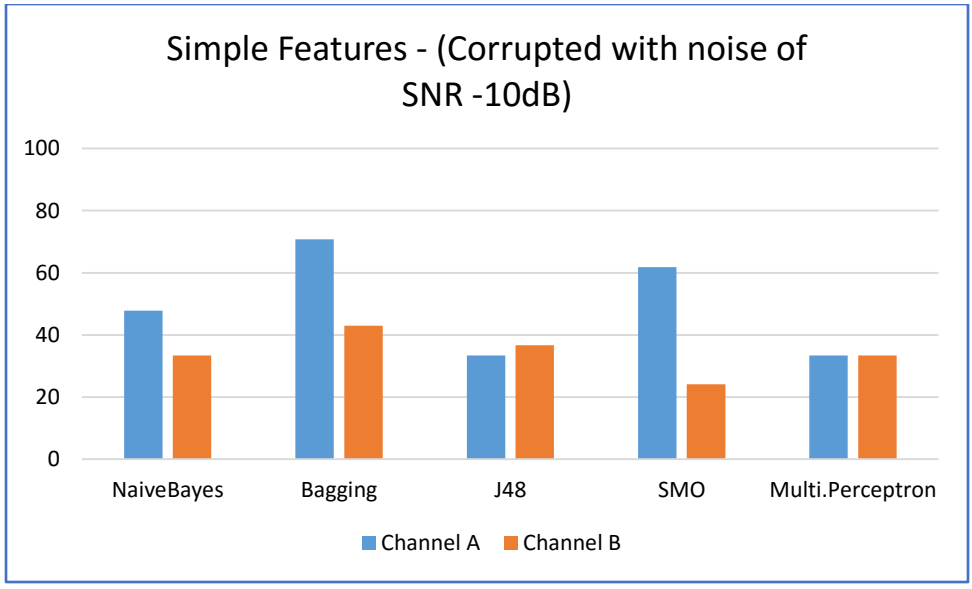

Figure 29: Classification accuracies using simple features extracted from noisy data of SNR =-10dB

Table 1: Confusion matrix for Channel A using Simple Features (SNR=-10dB)

	Porcelain	**Cable**	**Corona**
Porcelain	90	0	0
Cable	18	72	0
Corona	1	62	27

Table 2: Confusion matrix for Channel B using Simple Features (SNR=-10dB)

	Porcelain	**Cable**	**Corona**
Porcelain	90	0	0
Cable	90	0	0
Corona	64	0	26

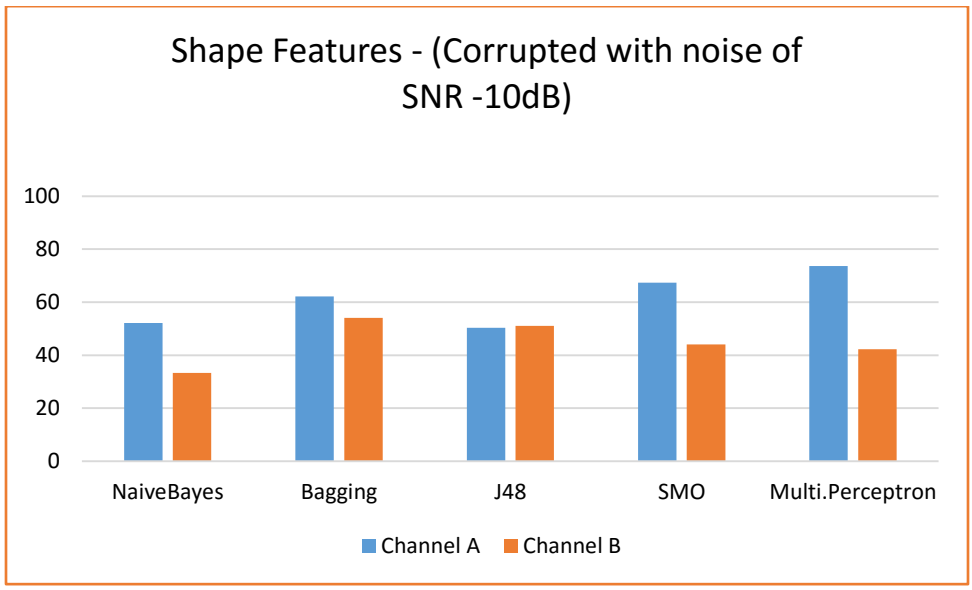

Figure 30: Classification accuracies using shape features extracted from noisy data of SNR= -10dB

Table 3: Confusion matrix for Channel A using Shape Features (SNR=-10dB)

	Porcelain	**Cable**	**Corona**
Porcelain	51	0	39
Cable	0	90	0
Corona	0	63	27

Table 4: Confusion matrix for Channel B using Shape Features (SNR=-10dB)

	Porcelain	**Cable**	**Corona**
Porcelain	72	0	18
Cable	72	0	18
Corona	16	0	74

In the simple features, the 'average of the signal' feature was greatly disturbed due to the addition of noise. This was because of the introduction of various peaks to the signal, which affected the magnitude and the number of peaks and valleys in the signal.

On the other hand, the shape features also suffered as the feature vector is completely made up of peak-related features. From *Figure 28* we can see that the average of the noisy signal is vastly different from the clean signal. We can also see that several peaks have been introduced in the noisy signal. This affected all the peak-related features in both Simple and Shape features. As the classification accuracy deteriorated in the case of both types of features, it is confirmed that the signal in the time domain is highly susceptible to noise. As the time domain was not robust enough to withstand noise, we moved to processing the signal in the frequency domain.

The signal in the time domain was converted into the frequency domain by applying FFT. Four statistical parameters: *Mean, Variance, Skewness and Kurtosis* were extracted from the signal. This method gave better results than when the signal (corrupted with noise) was processed in the time domain.

Table 5 and Table 7 depict the results of applying Custom Testing on the Statistical Features that were extracted from the signal in the frequency domain. For the Custom Testing, each of the 10 files was made up of 90% raw data and 10% noisy data for different SNR levels.

Table 5: Classification accuracies of PD signals from Channel A using statistical features

SNR (db)	SVM	SMO	KNN	RF	Bagging	OneR	JRip	J48
20	36.3	100	100	100	100	100	97.78	99.26
10	37.41	100	100	100	100	100	97.78	98.52
-10	33.33	76.3	69.26	51.85	51.85	51.85	60.37	75.93
-20	33.33	66.67	66.67	33.33	33.33	33.33	33.33	42.96
-30	33.33	43.33	42.96	33.33	33.33	33.33	33.33	33.33
-40	33.33	33.7	34.44	33.33	33.33	33.33	33.33	33.33

Table 6: Confusion matrix for Channel A using Statistical Features (SNR=-10dB)

	Porcelain	**Cable**	**Corona**
Porcelain	9	0	0
Cable	0	9	0
Corona	0	9	0

Table 7: Classification accuracies of PD signals from Channel B using statistical features

SNR (dB)	SVM	SMO	KNN	RF	Bagging	OneR	JRip	J48
20	77.04	90	100	100	100	100	99.26	98.52
10	76.3	83.7	100	100	100	87.04	98.89	97.78
-10	33.33	67.04	51.48	77.78	77.78	52.96	72.22	53.33
-20	33.33	58.52	33.33	33.33	33.33	19.63	52.96	33.33
-30	33.33	66.67	33.33	33.33	33.33	9.63	30	33.33
-40	33.33	33.33	33.33	33.33	33.33	21.48	33.33	33.33

Table 8: Confusion matrix for Channel B using Statistical Features (SNR=-10dB)

	Porcelain	**Cable**	**Corona**
Porcelain	9	0	0
Cable	9	0	0
Corona	9	0	0

4.7 Existing Work

We will compare our results with the system proposed by Ramy et. al [39]. This paper describes the detection of 3 types of PD activities, namely: *Sharp, Void and Surface* using an acoustic sensor.

The data that is received from the acoustic sensor is made up of 2500 attributes and is the signal in the time domain. The captured PD time-domain signals are first transformed to the frequency domain using FFT. From the frequency spectrum that is obtained, the lower frequency components are selected as a feature vector. This feature vector is made

up of 250 points from the lower frequency band of the signal. This paper achieved very high accuracy, even in the presence of very high noise ratios, making it highly robust to noise. *Figure 31* shows the selection of the lower frequency components that are used as the feature vector.

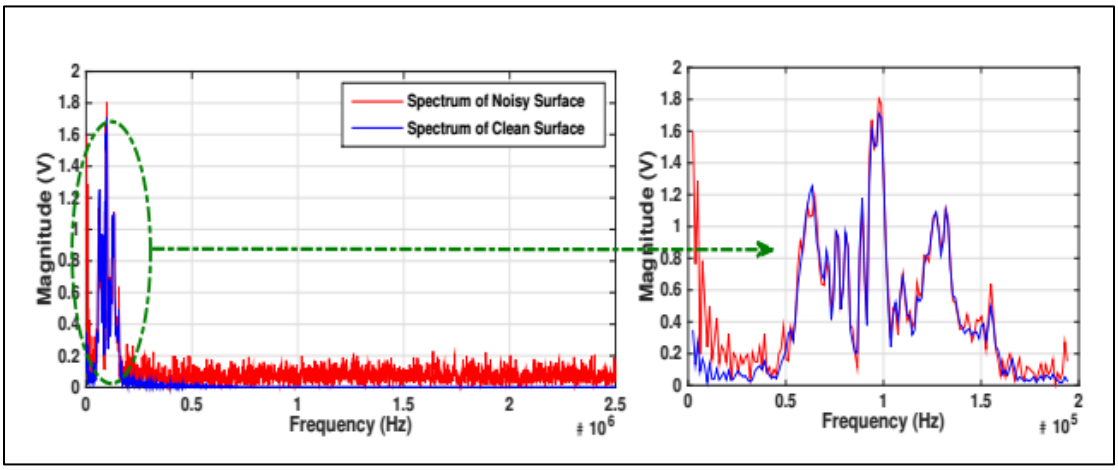

Figure 31: Selection of lower frequency components that are used as the feature vector. [39]

The results we get from this system will be compared with the results that we get from our system for this research. The classification accuracy obtained on the PD signals corrupted with white noise is depicted in *Table 9*.

Table 9 : Classification accuracy of acoustic PD signals (existing work) corrupted with white noise using Custom Testing

SNR (dB)	RF	KNN	SVM	NB	J48	JRIP	OneR
20	99.26	94.81	98.89	98.89	98.15	97.04	87.78
10	99.26	95.56	98.89	63.7	97.78	93.33	89.26
-10	55.93	95.93	86.67	33.33	59.26	60	78.15
-20	33.33	47.04	33.33	33.33	37.04	38.89	52.96
-30	33.33	33.33	33.33	33.33	34.44	33.7	34.81
-40	33.33	33.33	33.33	32.59	34.07	34.07	33.7

Table 10: Confusion matrix of PD Acoustic Signals (SNR=-10dB)

	Porcelain	**Cable**	**Corona**
Porcelain	8	0	1
Cable	1	7	1
Corona	0	3	6

By applying the methodology of the this system on our data, we achieve the accuracies depicted by Table 11 and Table 13 for Channel A and Channel B respectively.

Table 11: Classification accuracy of PD signals from Channel A corrupted with white noise using Custom Testing

SNR (dB)	SVM	SMO	KNN	RF	Baggin	OneR	JRip
20	85.56	100	38.15	100	100	96.67	78.52
10	93.33	100	38.52	100	100	100	89.26
-10	80	68.52	33.33	56.67	82.96	75.93	71.48
-20	49.26	66.67	33.33	18.89	45.56	63.33	48.89
-30	33.33	42.96	33.33	45.93	61.85	37.04	34.44
-40	33.33	33.33	33.33	30.37	44.44	34.07	33.33

Table 12: Confusion matrix of Channel A using the methodology described by the existing work (SNR=-10dB)

	Porcelain	Cable	Corona
Porcelain	9	0	0
Cable	0	9	0
Corona	3	2	4

Table 13: Classification accuracy of PD signals from Channel B corrupted with white noise using Custom Testing

SNR (dB)	SVM	SMO	KNN	RF	Baggin	OneR	JRip
20	99.26	90.37	87.41	95.93	93.33	91.11	99.63
10	85.19	90.37	33.33	87.04	90.37	89.63	98.52
-10	51.48	76.3	33.33	51.85	77.41	73.7	83.7
-20	28.15	48.52	33.33	66.67	53.33	46.3	60.37
-30	32.96	33.33	33.33	52.96	38.89	19.63	49.63
-40	33.33	33.33	33.33	19.63	34.07	16.3	38.52

Table 14: Confusion matrix of Channel B using the methodology described by the existing work (SNR=-10dB)

	Porcelain	**Cable**	**Corona**
Porcelain	0	0	9
Cable	0	8	1
Corona	0	1	8

Our aim is to beat the results that we achieve from using this methodology described by this system, even at higher noise levels.

4.8 **Regularization**

We applied Lasso Regularization on the data using the GLMNET library that is available in the R language [63]. The Lasso Regularization model is run using the Logistic Regression classifier that is pre-defined in the GLMNET library. Lasso can be applied on *Equation 3.10* by setting α to 1. The equation is reduced to

$$\min_{\beta_0, \beta} \frac{1}{N} \sum_{i=1}^{N} w_i \, l(y_i, \beta_0 + \beta^T x_i) + \lambda \, \|\beta\|_1 \tag{4.1}$$

The best-fit for λ is obtained by performing a cross-validation of the model. The ideal model is built using this best-fit λ for predicting the class of the observations. The best-fit for λ is obtained by performing cross-validation on 15 observations (5 observations from each class). These observations are not a part of the Custom Testing and remain

unseen for further processing.

The custom testing is performed on the remaining 270 observations. Consider File-1 (out of 10 files) of the Custom testing, where 90% of the data is clean and 10% of the data is corrupted with SNR +20 noise. The number of attributes has been decreased from 2016 to 22. The best-fit for lambda ($\lambda = 0.004$) was chosen from 1000 lambda values by the model after 10 folds of cross-validation. The selected coefficients (non-zero coefficients) is depicted in

Figure 32. The axis on the top shows the number of non-zero coefficients at each level of lambda.

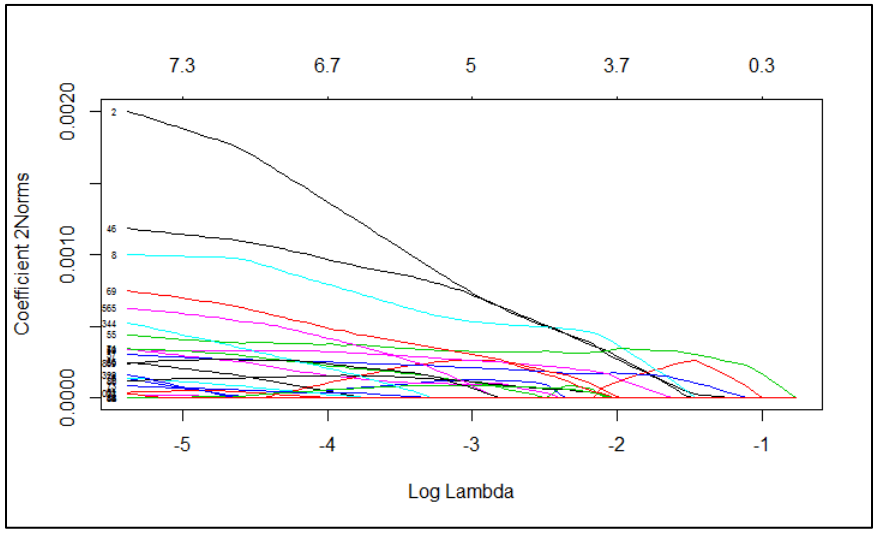

Figure 32: Non-zero coefficients chosen by the model plotted against the log of the lambda values

Figure 33 shows the path of the coefficient against the L1 – norm of the whole coefficient vector as λ varies.

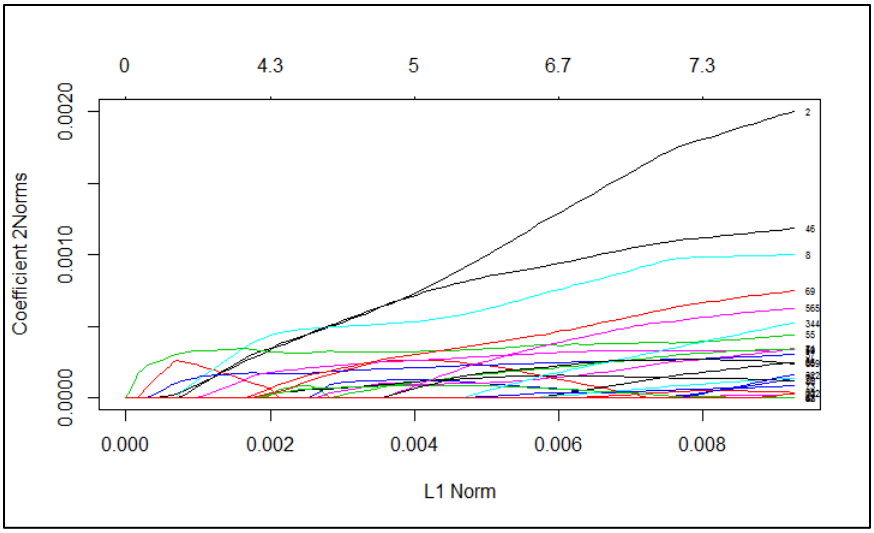

Figure 33: Non-zero coefficients chosen by the model plotted against the L1 Norm of the coefficients

A summary of the GLMNET path is depicted by *Figure 34*.

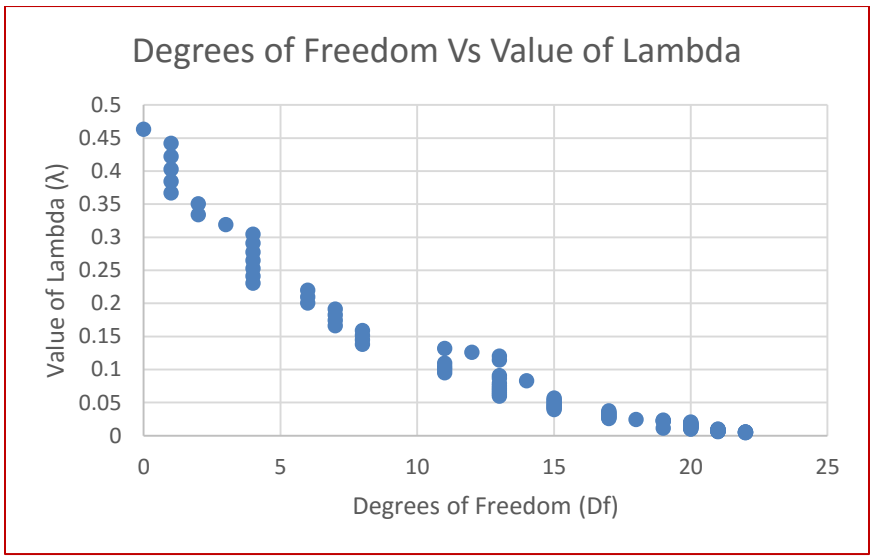

Figure 34: GLMNET path at λ =0.004

Where, Df (Degrees of freedom) represents the number of non-zero coefficients at the current value of λ. *Figure 35* represents the cross-validation curve, which is represented

by the red dotted line. The upper and lower standard deviation curves along with the λ sequence are shown like error bars around the red-dotted line.

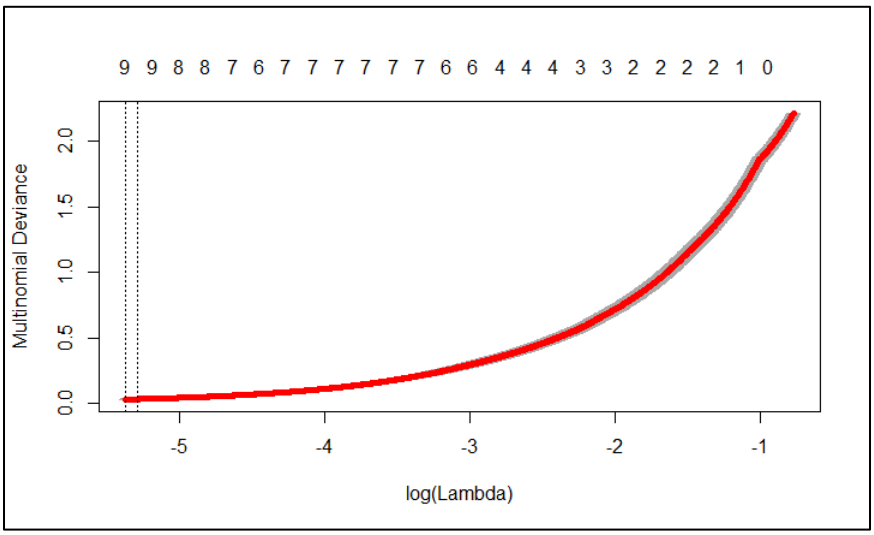

Figure 35: Values of lambda.min and lambda.1se plotted against deviance.

There are two vertical lines that depict two values of lambda. These values are:

1. lambda.min – Value of lambda that gives minimum mean cross-validation error.
2. lambda.1se – Value of lambda, which gives the most regularized model, such that the error is within one standard error of the minimum.

Generally, the lambda.min for each dataset is identified by cross-validation of the model. In the case of custom-testing, the lamba.min is taken as the average of lambda.min values that are found for each of the ten files that are used in the custom testing.

4.9 Results

In this section, we will present the results of applying Regularization on our dataset for both channels.

The results are reported for two settings of the Lasso Regularization.

1. ***Setting-1:*** The dataset is not standardized for Lasso.
2. ***Setting-2:*** The dataset is standardized for Lasso.

The results that are reported in Table 15 and Table 17 are obtained by running Custom Testing that is described in *Section 3.5*. The training data is corrupted with little noise of SNR = 0dB as opposed to using raw data. This type of training data in which the raw signal is corrupted with some noise helped to achieve high accuracies using Lasso. The clean signal and the signal corrupted with minimal noise of SNR = 0dB are shown in *Figure 36*.

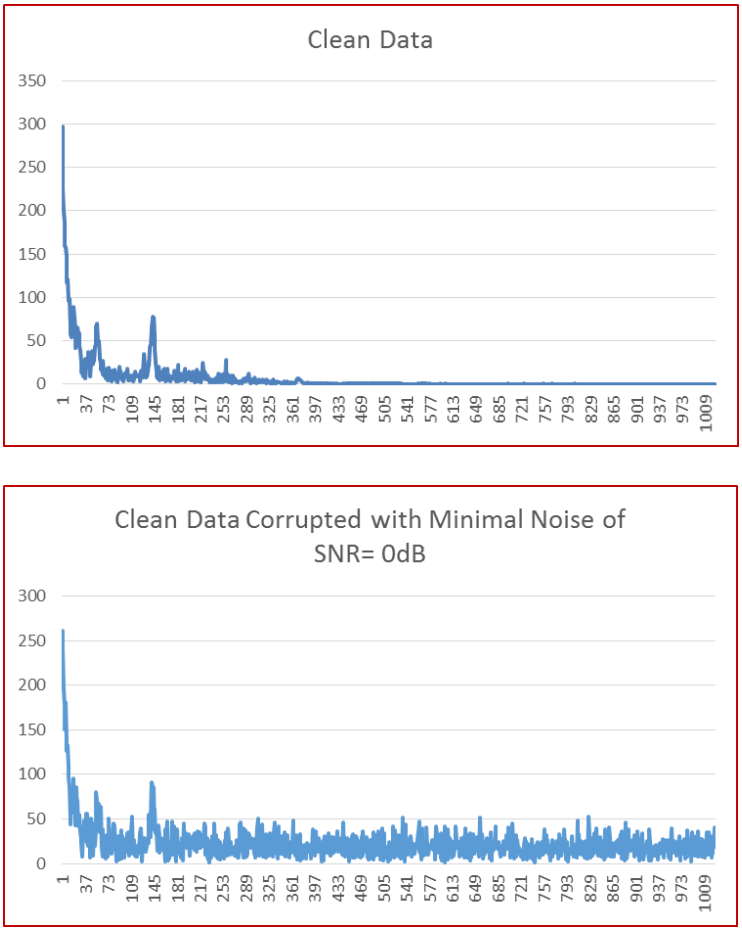

Figure 36: Top: Clean Signal. *Bottom:* Signal corrupted with SNR 0 noise.

Regularization was applied on Channel A and Channel B under the two different settings, as described previously. The values of the best fit for lambda for each setting is indicated in Table 15 and Table 17 for Channels A and B respectively.

Table 15: Classification accuracies of applying Lasso on Channel A

SNR (dB)	Standardization = FALSE λ = 15.971 Setting-1	Standardization = TRUE λ = 0.0049 Setting-2
20	100	100
10	100	100
-10	98.88	99.25
-20	75.92	82.59
-30	68.51	65.92
-40	46.66	40

Table 16: Confusion matrix of Channel A using Regularization (SNR=-20dB)

	Porcelain	**Cable**	**Corona**
Porcelain	9	0	0
Cable	0	6	3
Corona	0	1	8

From Table 15, we can see that Setting-1 did better than Setting-2. We can infer from these values, that for Channel A we need not standardize the data.

Table 17: Classification accuracies of applying Lasso on Channel B

SNR (dB)	Standardization = FALSE λ = 182.49 **Setting-1**	Standardization = TRUE λ = 0.004 **Setting-2**
20	66.29	67.03
10	66.29	66.29
-10	61.11	60.37
-20	57	71.85
-30	60	44.44
-40	65.55	33.33

Table 18: Confusion matrix of Channel B using Regularization (SNR=-10dB)

	Porcelain	**Cable**	**Corona**
Porcelain	4	2	3
Cable	0	9	0
Corona	9	0	0

From Table 17, we can see that Setting-2 does better than Setting-1. Hence, in the case of Channel B, we would have to standardize the data to get better classification accuracies.

4.9.1 Comparison of the results

In this section, we will compare the accuracy of the best performing Lasso setting with the average of the classification accuracies of all the classifiers used by the state-of-the-art system through each noise level.

As the SVM classifier did not do well constantly, two averages of the classifiers were calculated. The first average was calculated over all the classifiers whereas, the second

average was computed over all the classifiers except SVM.

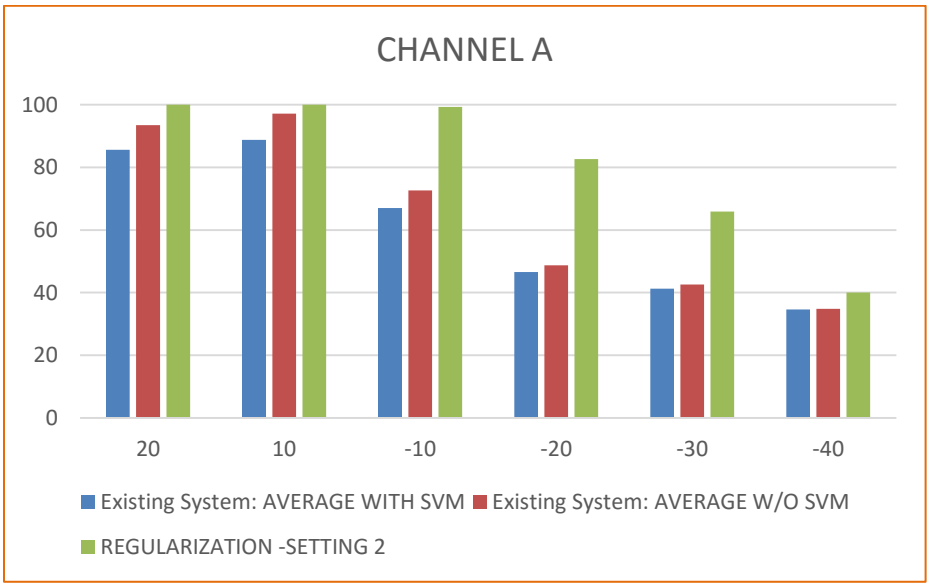

Figure 37: Comparison of results of Channel A

From Figure 37***Error! Reference source not found.***, we can infer that Lasso constantly does better than the methodology followed by the state-of-the-art system, across all levels of noise.

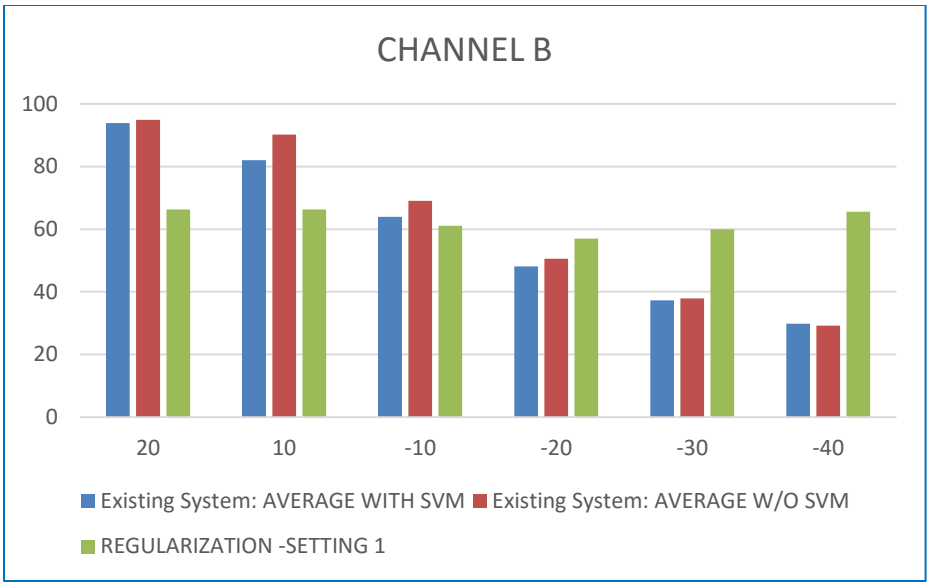

Figure 38: Comparison of results of Channel B

From Figure 38, we can see that the Lasso either performs better than the existing system for Channel A.

Overall, Lasso performed much better on Channel A which is the data from the sensor with the higher bandwidth, when compared with Channel B which is the data from the sensor with the lower bandwidth. This shows that it is better to use Channel A for achieving higher classification accuracies. It is interesting to note that Regularization performs better than the existing system for higher levels of noise in the case of Channel B. Hence, Regularization performs well when it has sufficient information from the signal, as in the case of Channel A.

Chapter 5

5. Conclusion and Future Work

This chapter provides a summary of the thesis based on the experimental results that we have achieved using the different methodologies along with the future work for this research.

5.1 Conclusion

PD location detection and type classification are well-known areas of research. Nowadays, PD research focuses on enhancing the usability of Smart Grids by evaluating the health of HV equipment. In this thesis, we addressed the classification problem between three types of PD activities that occur in power transformers, which are one of the costliest and mostly found components in any power grid.

The data from two sensors with different bandwidths were compared and it was inferred that the data from the higher frequency sensor achieved higher classifier accuracy when compared to the cheaper lower frequency sensor.

Simple and Shape features from the signal in the time domain and statistical features from the frequency domains were extracted. However, these features collapsed when the signal was corrupted with even small levels of noise. As we wanted to create a more robust PD classification system, we applied Lasso Regularization to create a simple classification model that could achieve high classification accuracies even when a very pessimistic form of testing was applied. This simple Lasso model did not require tuning for different levels of noise, as the regularization parameter λ is fixed across all noise levels. We also inferred that Lasso achieved higher classification accuracies than the existing system especially when the data was corrupted with higher levels of noise.

We contributed to PD research mainly in the following ways:

1. Proposed simple features for processing the signal in the time domain.
2. Proposed shape features for processing the signal in the time domain.
3. Proposed Lasso Regularization for processing the signal in the frequency domain.

Other contributions:

1. Applied statistical features for processing the signal in the frequency domain.
2. Used data collected from CT sensors and compared the results between two sensors of different bandwidths.

5.2 Future Work

The work done as part of this thesis will be added on to construct a complete PD monitoring system. This system would be capable of detecting the severity of the PD type in addition to classifying PD types, which was the work of this thesis. The research can be made more profound by considering other types of real noise such as random noise and discrete spectral interferences. Other techniques to get better classification accuracies from lower frequency sensors can also be investigated.

Research can also be done to check if Lasso can be applied to other types of PD signals (acoustic PD, electrical PD etc.) and for other signal types like Electroencephalogram (EEG) signals.

References

[1] Hu, J., Zhu, J., & Platt, G. (2011, August). Smart grid—the next generation electricity grid with power flow optimization and high power quality. In *Electrical Machines and Systems (ICEMS), 2011 International Conference on* (pp. 1-6). IEEE.

[2] Lauby, Mark, John Moura, and Eric Rollison. "Reliability considerations from the integration of smart grid." *Innovative Smart Grid Technologies (ISGT), 2012 IEEE PES*. IEEE, 2012.

[3] Khaitan, Siddhartha Kumar, James D. McCalley, and Chen Ching Liu, eds. Cyber physical systems approach to smart electric power grid. Springer, 2015.

[4] Hossain, M. R., Oo, A. M., & Ali, A. S. (2013). Smart Grid. In *Smart Grids* (pp. 23-44). Springer London.

[5] Wei, C. (2010, March). A conceptual framework for smart grid. In *2010 Asia-Pacific Power and Energy Engineering Conference* (pp. 1-4). IEEE.

[6] Chudnovsky, Bella H. Electrical Power Transmission and Distribution: Aging and Life Extension Techniques. CRC Press, 2012.

[7] Lai, K. X., Phung, B. T., & Blackburn, T. R. (2010). Application of data mining on partial discharge part I: Predictive modelling classification. *IEEE Transactions on Dielectrics and Electrical Insulation*, *17*(3), 846-854.

[8] Babnik, T., Aggarwal, R., & Moore, P. (2007). Data mining on a transformer partial discharge data using the self-organizing map. *IEEE Transactions on Dielectrics and Electrical Insulation*, *14*(2), 444-452.

[9] Neto, J. M., Zhang, Y., Jaber, A., Zhu, M., Judd, M., Atkinson, R., ... & Glover, I. A. (2014, August). Radiometric location of partial discharge sources for the future smart grid. In *General Assembly and Scientific Symposium (URSI GASS), 2014 XXXIth URSI* (pp. 1-4). IEEE.

[10] Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." Wiley interdisciplinary reviews: computational statistics 2.4 (2010): 433-459.

[11] Lee, E. A. (2011). Structure and interpretation of signals and systems. Lee & Seshia.

[12] Muhr, M., Strehl, T., Gulski, E., Feser, K., Gockenbach, E., Hauschild, W., & Lemke, E. (2006). Sensors and sensing used for non-conventional PD detection. *CIGRE, Paris*, 1-7.

[13] Sahoo, N. C., Salama, M. M. A., & Bartnikas, R. (2005). Trends in partial discharge pattern classification: a survey. *Dielectrics and Electrical Insulation, IEEE Transactions on*, *12*(2), 248-264.

[14] Aziz, N. F. A., Hao, L., & Lewin, P. L. (2007). Analysis of partial discharge measurement data using a support vector machine. In *Research and Development, 2007. SCOReD 2007. 5th Student Conference on* (pp. 1-6). IEEE.

[15] Note, Application. "The Fundamentals of Signal Analysis." *Hewlett Packard*: 98-107.

[16] Firuzi, K., Parvin, V., & Vakilian, M. (2015, July). Identification of free conducting particles in transformer oils using PD signals. In *Properties and Applications of Dielectric Materials (ICPADM), 2015 IEEE 11th International Conference on the* (pp. 724-727). IEEE.

[17] Harbaji, M., Shaban, K., & El-Hag, A. (2015). Classification of common partial discharge types in oil-paper insulation system using acoustic signals. *Dielectrics and Electrical Insulation, IEEE Transactions on*, *22*(3), 1674-1683.

[18] Németh, B., Laboncz, S., & Kiss, I. (2009, May). Condition monitoring of power transformers using DGA and fuzzy logic. In 2009 IEEE Electrical Insulation Conference (pp. 373-376). IEEE.

[19] Babnik, T., Aggarwal, R. K., Moore, P. J., & Wang, Z. D. (2003, June). Radio frequency measurement of different discharges. In *Power Tech Conference Proceedings, 2003 IEEE Bologna* (Vol. 3, pp. 5-pp). IEEE.

[20] Duval, M. (1989). Dissolved gas analysis: It can save your transformer. *IEEE Electrical Insulation Magazine*, *5*(6), 22-27.

[21] Darabad, V. P., Vakilian, M., Phung, B. T., & Blackburn, T. R. (2013). An efficient diagnosis method for data mining on single PD pulses of transformer insulation defect models. *Dielectrics and Electrical Insulation, IEEE Transactions on*, *20*(6), 2061-2072.

[22] Hao, L., & Lewin, P. L. (2010). Partial discharge source discrimination using a support vector machine. *IEEE Transactions on Dielectrics and electrical Insulation*, *17*(1), 189-197.

[23] Tokmakci, Mahmut, et al. "Effect of chewing on dental patients with total denture: an experimental study." SpringerPlus 2.1 (2013): 40.

[24] Lai, K. X., Phung, B. T., Blackburn, T. R., & Muhamad, N. A. (2007, December). Classification of partial discharge using PCA and SOM. In 2007 International Power Engineering Conference (IPEC 2007) (pp. 1311-1316). IEEE.

[25] Lai, K. X., B. T. Phung, and T. R. Blackburn. "Application of data mining on partial discharge part I: Predictive modelling classification." IEEE Transactions on Dielectrics and Electrical Insulation 17.3 (2010).

[26] Evagorou, D., Kyprianou, A., Georghiou, G. E., Hunter, J. A., Hao, L., Lewin, P. L., & Stavrou, A. (2010, October). Performance of the Support Vector Machine Partial Discharge classification method to noise contamination using phase synchronous measurements. In Electrical Insulation and Dielectric Phenomena (CEIDP), 2010 Annual Report Conference on (pp. 1-4). IEEE.

[27] Valens, Clemens. "A really friendly guide to wavelets." *ed. Clemens Valens* (1999).

[28] Zhang, Hao, et al. "A novel wavelet transform technique for on-line partial discharge measurements. 1. WT de-noising algorithm." *IEEE Transactions on Dielectrics and Electrical Insulation* 14.1 (2007).

[29] Bhasme, N. R., and Bhushan Salokhe. "PARTIAL DISCHARGE MEASUREMENT AS A DIAGNOSTIC TOOL FOR CURRENT TRANSFORMER." *International Journal of Advances in Engineering & Technology* 8.3 (2015): 347.

[30] Soomro, Irfan Ali, and Md Nor Ramdon. "Study on different techniques of partial discharge (PD) detection in power transformers winding: Simulation between paper and EPOXY resin usingg UHF method." *International Journal of Conceptions on Electrical and Electronics Engineering*.

[31] Tipler, Paul A., and Gene Mosca. Physics for scientists and engineers. Macmillan, 2007.

[32] Obralic, A., R. Plath, and W. Kalkner. "PULSE SEQUENCE ANALYSIS OF PD-DATA MEASURED ON OUTER BOUNDARY SURFACES PAPER TITLE." (2009).

[33] Haykin, Simon S. Neural networks: a comprehensive foundation. Tsinghua University Press, 2001.

[34] Raymond, Wong Jee Keen, Hazlee Azil Illias, and Hazlie Mokhlis. "Partial discharge classifications: Review of recent progress." *Measurement* 68 (2015): 164-181.

[35] Demirhan, Ayşe, and İnan Güler. "Combining stationary wavelet transform and self-organizing maps for brain MR image segmentation." Engineering Applications of Artificial Intelligence 24.2 (2011): 358-367.

[36] Leung, K. Ming. "Naive bayesian classifier." Polytechnic University Department of Computer Science/Finance and Risk Engineering (2007).

[37] Chakravorti, Sivaji, Debangshu Dey, and Biswendu Chatterjee. "Recent trends in the condition monitoring of transformers." *Power Systems Springer-Verlag: London, UK* (2013).

[38] Satish, L., and B. Nazneen. "Wavelet-based denoising of partial discharge signals buried in excessive noise and interference." *IEEE Transactions on Dielectrics and Electrical Insulation* 10.2 (2003): 354-367.

[39] Hussein, Ramy, Khaled Bashir Shaban, and Ayman H. El-Hag. "Robust Feature Extraction and Classification of Acoustic Partial Discharge Signals Corrupted With Noise." *IEEE Transactions on Instrumentation and Measurement* 66.3 (2017): 405-413.

[40] Tahir, Dr Nabeel. "International Journal Of Computer Science And Security (Ijcss)."

[41] Quinlan, J. R. (1996, August). Bagging, boosting, and C4. 5. In AAAI/IAAI, Vol. 1 (pp. 725-730).

[42] Ruck, Dennis W., Steven K. Rogers, and Matthew Kabrisky. "Feature selection using a multilayer perceptron." Journal of Neural Network Computing 2.2 (1990): 40-48.

[43] Platt, John. "Sequential minimal optimization: A fast algorithm for training support vector machines." (1998).

[44] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." (2007): 3-24.

[45] Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot. "Variable selection using random forests." *Pattern Recognition Letters* 31.14 (2010): 2225-2236.

[46] Hindle, Abram, et al. "Automatic classication of large changes into maintenance categories." *Program Comprehension, 2009. ICPC'09. IEEE 17th International Conference on*. IEEE, 2009.

[47] Dreiseitl, Stephan, and Lucila Ohno-Machado. "Logistic regression and artificial neural network classification models: a methodology review." *Journal of biomedical informatics* 35.5 (2002): 352-359.

[48] Subasi, Abdulhamit, and Ergun Ercelebi. "Classification of EEG signals using neural network and logistic regression." Computer methods and programs in biomedicine 78.2 (2005): 87-99.

[49] Gaouda, Ahmed M., et al. "A Smart IEC 61850 Merging Unit for Impending Fault Detection in Transformers." *IEEE Transactions on Smart Grid* (2016).

[50] Friedman, Jerome H. "Fast sparse regression and classification." International Journal of Forecasting 28.3 (2012): 722-738.

[51] Hastie, Trevor, and Junyang Qian. "Glmnet vignette." (2014).

[52] Ng, Andrew Y. "Feature selection, L 1 vs. L 2 regularization, and rotational invariance." Proceedings of the twenty-first international conference on Machine learning. ACM, 2004.

[53] Tibshirani, Robert. "Regression shrinkage and selection via the lasso: a retrospective." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73.3 (2011): 273-282.

[54] Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67.2 (2005): 301-320.

[55] Zidan, Aboelsood, et al. "Fault Detection, Isolation, and Service Restoration in Distribution Systems: State-of-the-Art and Future Trends." *IEEE Transactions on Smart Grid* (2016).

[56] Pattanadech, N., & Nimsanong, P. (2014, October). Partial discharge classification using principal component analysis combined with self-organizing map. In *TENCON 2014-2014 IEEE Region 10 Conference* (pp. 1-5). IEEE.

[57] Pattanadech, N., & Nimsanong, P. (2014, October). Effect of noise signals on partial discharge classification models. In *TENCON 2014-2014 IEEE Region 10 Conference* (pp. 1-5). IEEE.

[58] Li, W., Luo, J., Xia, R., Zhao, J., Meng, S., Jiang, Y., & Min, H. (2010, March). Partial discharge pattern analysis based on 2-D wavelet transform. In *2010 Asia-Pacific Power and Energy Engineering Conference* (pp. 1-3). IEEE

[59] Klueter, T., Wulff, J., & Jenau, F. (2013, May). Time domain analysis of partial discharges at DC voltage in air and insulation oil. In *Environment and Electrical Engineering (EEEIC), 2013 12th International Conference on* (pp. 308-312). IEEE.

[60] Hao, L., Lewin, P. L., & Dodd, S. J. (2006, June). Comparison of support vector machine based partial discharge identification parameters. In *Conference Record of the 2006 IEEE International Symposium on Electrical Insulation* (pp. 110-113). IEEE.

[61] Thayoob, Y. H. M., Ahmed, S. K., Piau, C. C., Ping, C. Y., & Balasubramaniam, Y. (2015, October). Characterization of Phase Resolved Partial Discharge waveforms from instrument transformer using statistical signal processing technique. In 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA) (pp. 355-360). IEEE.

[62] Wu, Min, et al. "An overview of state-of-the-art partial discharge analysis techniques for condition monitoring." *IEEE electrical insulation magazine* 31.6 (2015): 22-35.

[63] Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. "glmnet: Lasso and elastic-net regularized generalized linear models." R package version 1.4 (2009).