UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

Electronics and Computer Science

An Approach Towards Plant Electrical Signal Based External Stimuli Monitoring System

by

Shre Kumar Chatterjee

Thesis for the degree of Doctor of Philosophy

March 2017

Declaration of Authorship

I, Shre Kumar Chatterjee, declare that the thesis entitled *An Approach Towards Plant Electrical Signal Based External Stimuli Monitoring System* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research.

I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at the University,
- Where any part of this thesis has been previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated,
- Where I have consulted the published work of others, this is always clearly attributed,
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work,
- I have acknowledged all main sources of help,
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ……..

Date: …….

University of Southampton

ABSTRACT

Faculty of Physical Sciences and Engineering
Electronics and Computer Science

Doctor of Philosophy

An Approach Towards Plant Electrical Signal Based External Stimuli Monitoring System

by

Shre Kumar Chatterjee

Plants have sensing mechanisms which are employed to monitor their environment for optimal growth. This sensing mechanism can be observed by the change in behaviour in plants like *Mimosa pudica* (*Touch Me Not*) which closes its leaves when touched or *Dionaea muscipula* (*Venus Flytrap*) which closes its trap when an insect gets in it. It has been established that plants produce an electrical signal response to stimuli that is used to control various physiological phenomena within the plant. If such electrical signals are extracted and analysed, information about the external stimuli which caused the electrical signal may be found. If such an analysis is successful, then plants can be used as a living multiple stimuli sensor.

This work explores the possibility of extracting information from the plant electrical signal response to the external stimuli which caused the plant to produce such a signal. Initially, the plant was treated as a black box system and a simple input (light pulse as stimulus) – output (electrical signal response) system was modelled through *system identification* techniques. Thereafter, an inverse system was modelled for input (electrical signal response) – output (light pulse as stimulus) to find out if there exists, within the plant's electrical signals, adequate information about the *time* of application and the *intensity* of the applied stimulus.

Next, *classification* methods were employed to find out if there was adequate information, within the raw plant electrical signal response, about the *type* of stimulus applied to the plants. More complex stimuli such as *Sulphuric acid*, *Ozone* and *Sodium chloride* solutions were applied to the plants to find out if the plant electrical signal response could be used to classify these stimuli in a binary classification scenario. Discriminant analysis based

classifiers were employed along with simple statistical features which produced classification accuracy of around 70%.

A *decision tree* based classification strategy was then explored, using discriminant analysis classifiers and statistical features, in a multiclass classification strategy with the aim of enhancing classification accuracy. This exploration involved more datasets which enabled a prospective study (separate data held out) to be carried out to see the results in a more realistic scenario. The decision tree based classification system produced an accuracy of around 90% for both retrospective and prospective studies. In this work, both raw and filtered signals were used, of which the raw signals produced marginally better results than the filtered ones.

Lastly, curve fitting coefficients were explored for classification of stimuli by fitting four different curves to raw plant electrical signals. Classification accuracy of around 90% was achieved during the retrospective study by using polynomial curve fit coefficients. This enabled features to be extracted from the entire duration of the time series rather than small segments of it, in order to see if classification accuracy improved.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Om-ajnana-timirandhasya jnananjana-salakaya

caksur unmilitam yena tasmai sri-gurave namah

I offer my respectful obeisance's unto my spiritual master, who has opened my eyes, which were blinded by the darkness of ignorance, with the torchlight of knowledge.

Leaving a well-settled life in industry to pursue a PhD has been one of the boldest decisions of my life. Had it not been for my supervisor Prof Koushik Maharatna, I would perhaps not have realised how important that decision was going to be. I am grateful to Prof Maharatna for spending so many days over critical analysis of my research. I have really enjoyed my journey so far. I am also indebted to Dr Srinandan Dasmahapatra for his valuable suggestions, advice and guidance.

I am thankful to Dr Saptarshi Das for always being available for discussions and continuous support. His belief in me and motivation was a breath of fresh air on so many occasions. I would also like to thank my friends and colleagues Dr Dwaipayan Biswas, Dr Sanmitra Ghosh, Dr. Valentina Bono, Dr Taihai Chen, Dr Tristan Aubrey-Jones and Nawfal Al Firas for always being there for me. A heartfelt gratitude also goes towards Dr. Obaid Malik for having the time for discussions and suggestions.

I would like to thank my parents, wife, son and in-laws for supporting and understanding me throughout these four years. I would not have come this far without their support. And a special thanks to my brother for first instilling the dream of pursuing a PhD, in me.

Finally, a wholehearted thanks to Dr Aldo Faisal (external examiner) and Prof Mark Zwolinski (internal examiner) for valuable suggestions which made this thesis better.

Abbreviations

A

A/D, 63

ABA, 16, 17, 21, 28

ADP, 18

Ag, 25, 26

AgCl, 25, 26

AP, 6, 13, 14, 22, 23, 24, 28, 29, 30, 31, 38

ARMAX, 48, 55, 56, 69, 70

ARX, 48, 54, 55, 56, 57, 58, 69, 70

ATPase, 18, 33

B

BJ, 48, 56, 69, 70, 74, 77

BSS, 39

C

Ca^{2+}, 4, 17, 18, 21, 22, 32

CCCP, 31

CCD, 27

Cl^{-}, 20, 28

CO, 2, 17

CO_2, 4, 16, 34

D

DAQ, 63, 66, 99, 100

db3, 37, 38

DFA, 105, 106, 107, 108, 124, 138, 149, 162, 184, 185, 186, 187, 188, 189

DNP, 30

E

EMG, 62, 99, 151

ESN, 10

F

FCCP, 31

FDR, 108, 109, 140, 141, 142, 156, 158

H

H^{+}, 18, 20, 31

H_2SO_4, 4, 6, 7, 29, 97, 99, 100, 101, 125, 127, 128, 129, 131, 132, 134, 135, 136, 137, 139, 144, 145, 146, 147, 156, 157, 158, 159, 162, 181, 182, 185, 186, 187, 188

I

ICA, 39, 41

IIR, 9, 151, 152, 153, 154

K

K^{+}, 4, 16, 17, 18, 19, 21, 22, 28, 45, 79

KCl, 25

L

LDA, 110, 117, 124, 125, 127, 129, 131, 139, 143, 145, 146, 149, 155, 156, 160, 161, 183

LED, 29, 64, 66, 80, 100

LEP, 6, 22

LOOCV, 113, 134, 135, 137, 141, 144, 145, 155, 156

LSE, 48, 51, 53, 54, 59, 69

LTI, 47

N

NaCl, 2, 3, 4, 6, 7, 21, 97, 99, 100, 101, 121, 123, 125, 126, 127, 128, 129, 131, 132, 134, 135, 136, 137, 139, 144, 145, 146, 147, 156, 157, 158, 159, 162, 181, 182, 185, 186, 187, 188

NLARX, 48, 57, 58, 62, 70, 93

NLHW, 44, 48, 61, 70, 71, 74, 77, 78, 93, 181

O

P

Q

R

S

V

W

1 Introduction

Our natural environment is one of the most important aspects of our lives, wherever we are in the world. The basic reason behind this being the supply of food, water and oxygen, which are required for our sustenance. Two of the main constituents of our environment are the *flora* – plants and trees which provide us with oxygen and fruits, vital components required for human beings, and *fauna* – the animals. Plants, trees and forests also provide us with cleaner air by removing some harmful components from the atmosphere such as *ozone* (O_3), *sulphur dioxide* (SO_2) and *nitrogen dioxide* (NO_2) which in quantity may cause adverse neurological, cardio-vascular and pulmonary health effects in human beings [1].

As a precaution, many institutions across the globe constantly monitor our environment for pollutants which may harm us and the flora/fauna, directly or indirectly. Such monitoring may present us with valuable real-time information that can be used to control or even prevent any short- or long-term damage to our environment and thereby improve our healthy maintenance. The monitoring systems employed usually sense one or more environmental parameters and may cover a particular geographical area. However, such monitoring over large geographical areas could be quite expensive and complex infrastructure required [2]–[4].

Therefore, not many countries may be enthusiastic in pursuing it as others, thereby increasing the chances of long-term damage to their local environment as well as the overall global environment.

Given such a scenario, if a monitoring system can be developed which is cost-effective and can be implemented on a large scale, it becomes worthy of being implemented by every nation for the justified cause of detecting harmful environmental pollutants which affect the global quality of life. Such a holistic system seems possible when plant electrophysiology is considered, i.e. electrical signals generated by plants in response to external stimuli. These electrical signals, which may have embedded signatures capturing the essence of the stimuli affecting them, may be used as means of sensing the environment in which they grow. Plants cover around one third of Earth's land mass [5] in the form of forests, and are globally distributed. Plants are also affected by the same environmental pollutants as other living things sharing the same natural environment. Therefore, if electrical signals from a plant could be used to monitor a certain area around that particular plant, then their abundantly

distributed presence could be used to monitor a larger area at the same time. The viability of such an approach seems immense, but several questions rise along with such optimism.

This work aims to answer a few such questions, taking one step closer to realising such a holistic monitoring system. This may help us take preventative measures in a timely fashion against any harm to us or our natural environment, from environmental pollutants.

Today, the world is witnessing an increased level of environmental pollution which is causing harm to our planet and as a result, us, on a global scale. Among all the types of pollution – air, water, soil, noise and light – *air pollution* is considered the most harmful to our environment. Increasing air pollution is one of the possible reasons linked to lung and bladder cancer [6], asthma [7], various allergies [8] and breathing-related problems [9]. The main components of air pollution are harmful gases such as *tropospheric ozone* (O_3), *sulphur dioxide* (SO_2), *carbon monoxide* (CO) and *oxides of nitrogen* (NO_x), which are emitted by factories and modern vehicles. This is depicted in Figure 1.1 [10].

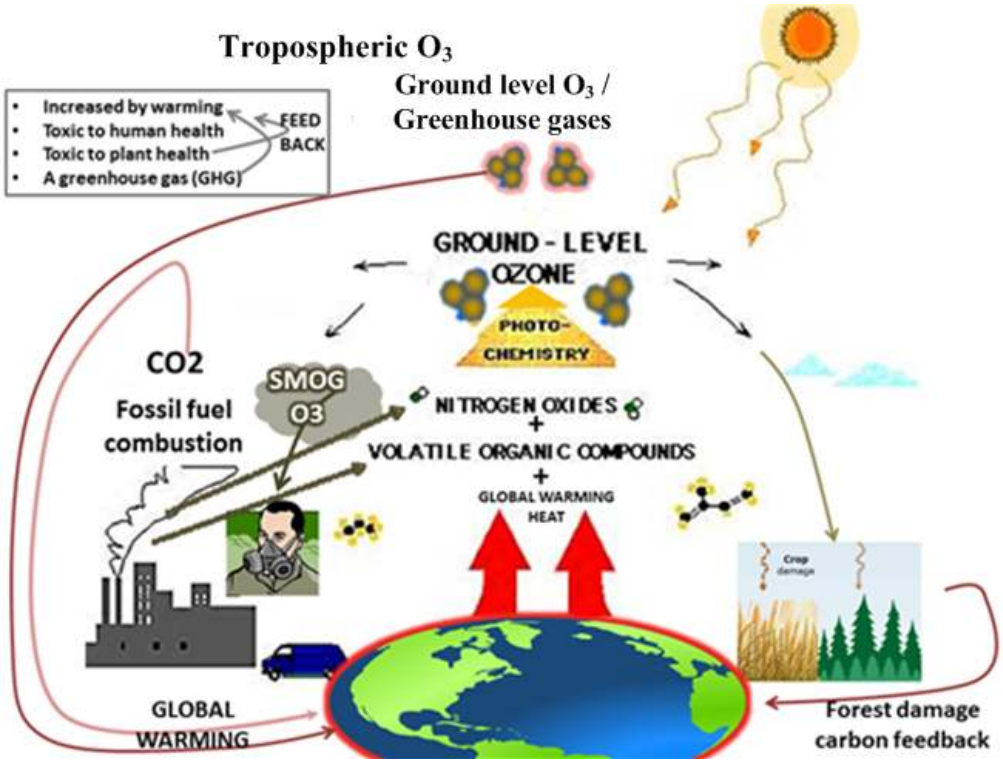


Figure 1.1: Typical constituents of air pollution

Ground level and tropospheric ozone (O_3) is becoming a serious pollutant, affecting health and crops globally [11]. Similarly salt (NaCl) pollution of soil and freshwater is another concern as it has the potential to change the bio-diversity of any region, if the usage of salt is not monitored properly [12]. When large quantities of SO_2 and NO_x are emitted into the

atmosphere from the combustion of fossil fuels, they undergo a chemical reaction to form *acid rain* which affects crops and living organisms, as well as structures which are not immune to acids (such as steel/iron, brick structures) [13].

This work explores the possibility of detecting these three pollutants, easily replicated under laboratory conditions, from the electrical response generated by plants. For that purpose, a set of experiments was conducted, using these three pollutants as external stimuli, on different species of plant, and their respective electrical responses were analysed. The three pollutants and their effects are discussed briefly below.

1.1 Ground level Ozone (O_3) pollution

One of the major components of air pollution today is tropospheric and ground level ozone (O_3) [11], which is the product of volatile organic compounds (VOCs) and nitrogen oxides (NO_x), aided by a rise in ambient temperature. Increasing ground level O_3, which is approximately 1.66 times heavier than air [14], is causing serious damage to crops, forest cover and human health, and has been a topic of research in different parts of the world [11]. O_3 is found in varying concentrations in the troposphere, which extends from the earth's surface up to 10-15 km elevation (dependant on latitude and time of the year), and the stratosphere, which extends from the troposphere up to 45-55 km elevation. In the stratosphere, O_2 molecules interact with ultra-violet radiation from the sun to form O_3. This layer of O_3 prevents harmful radiation from the sun from affecting humans, wildlife and plants [15]. The tropospheric O_3 is thus a secondary pollutant, contributed from multiple sources. It has also been suggested that precursors of O_3 found in the USA were emitted in Asia [15], thereby indicating that the pollutant may not be necessarily be locally dervied. Monitoring of tropospheric O_3 has been implemented in different parts of the world for its perceived adverse health impact [11], [14]–[20].

1.2 Soil pollution due to sodium chloride (NaCl)

Salinity of the soil is a major concern, especially for agriculture, as it leads to a reduction in crop yield [12]. Apart from affecting crops and vegetation on an immediate basis, there is a long-term effect of degradation of the soil, which is considered irreversible [12]. The quantity of croplands, estimated to be around one-third of the total current area, is expected to be increased with a rise in global climate change [12]. One of the major components of salinity in the soil is sodium chloride (NaCl). When NaCl permeates the soil, it can be harmful to

both plants and organisms such as snails, slugs, frogs, newts and earthworms thriving in the soil [21]. It can also alter the salinity of the groundwater table beneath and will thus increase the salinity of drinking water, which will harm human and animal health too [22].

If the salt reaches nearby freshwater streams, ponds, and lakes, it will not only affect the aquatic life [21] living in such waterbodies but also affect the animals and human beings exposed to such water resources. Apart from being a health hazard and damaging to the ecosystem, increased NaCl concentrations can also be corrosive to vehicles and infrastructure such as bridges. The NaCl input to the soil arises from several sources including water softeners, septic/sewage effluent and de-icing of roads/highways [22]. The amount and scope of usage of common salt as a de-icer for roads during the winter months across the globe makes it a major contributor to soil pollution [22]–[32].

1.3 Acid rain – sulphuric acid (H_2SO_4)

The admixture of wet and dry material deposited from the atmosphere with abnormal amounts of sulphuric and nitric acid constitutes *acid rain*. The acid deposition is formed when large quantities of SO_2 and NO_x are emitted to the atmosphere from the combustion of fossil fuels [33], [34]. These pollutants undergo chemical reactions in the atmosphere and form sulphuric acid (H_2SO_4) and nitric acids (HNO_3) which come down as acid rain. Several studies on the constituents of the acid rain in different parts of the world and their impact on different types of plants have been published [35]–[40]. Acid rain alters the chemical composition of the soil by leaching the base cations (such as Ca^{2+}, Mg^{2+}, K^+ and Na^+) with SO_4^{2-} and NO_3^- [13]. This affects not only the plants which thrive on land but also the microbes which are present in the soil [13].

Other obvious pollutants such as carbon monoxide, carbon dioxide (CO_2), nitrogen oxide, etc., affect the atmosphere, and these are also monitored and researched globally. However, this work focused on the three pollutants O_3, NaCl and H_2SO_4.

1.4 Monitoring the environment over a large geographical area

The need and scope of environmental sensors is thus apparent, so that monitoring multiple parameters of the environment (pollutants) which are of interest to us can be undertaken. A network of such sensors would give us information about the environmental pollutants over a large geographical area. Such a combination of sensors is called an *Environmental Sensor Network* (ESN) and is a useful tool to monitor an area of interest, from a remote location [4].

One of the major challenges of ESN is power management because often the sensor nodes in the network are stand-alone and will have to survive without frequent replacement of batteries [4]. Therefore, the sensor needs to be designed in such a way that the operation it carries out consumes limited power. It may also be designed to carry out discrete/intermittent rather than continuous monitoring, thereby reducing power consumption [4].

There has been research into different applications of ESNs to enable monitoring of large geographical areas, such as [41] whose authors propose a gas-based sensor system network, which they argue is better than chemical sensors, for monitoring O_3 and NO_2 in Tokyo. Sensor networks have also been used in agriculture, where data is required for monitoring soil water availability, leaf temperature, plant water status, insect/disease/weed infestations, sunlight intensity, etc. [2]. Such sensors also aid in *precision agriculture* – defined as a method to optimize agricultural production by tailoring soil/crop management techniques to individual fields [42]–[45], by providing real-time information about soil, water availability, plant health and environment. This, as suggested by many researchers and by users of such methods, boosts farm production and crop yields and optimizes costs [42]–[45]. Similarly, ESNs have been also employed for monitoring salinity in rivers [46], which helps managing salinity in waterbodies by complying with set standards.

Different environmental parameters need monitoring and ESNs already established help in achieving the required monitoring. However, any project which involve ESNs needs a lot of planning, manpower and financial resources to enable effective monitoring of the environment [4]. The possibility of monitoring the environment at the same time as monitoring crops, over a large geographical area seems to hold potential. Such a system may be implemented for two reasons – knowing what the ambient environmental parameters are in real-time and how it is affecting the plants/crops/forests, etc. However, to be implemented widely, such a system also needs to be economical. That is, it needs to be cheap to employ, can be left alone for long periods of time without replacement of batteries, and can withstand harsh environmental conditions without itself affecting any of the parameters which it monitors [2], [4]. Various methods are already employed by the developed, as well as by many developing, nations to monitor these factors individually [2], [4]. However, such monitoring systems could be quite expensive and not all administrations or governments may be able to afford them. On the other hand, if a reliable sensory mechanism is available, which can monitor these factors across a large geographical area and is cheap to employ, then it will

be instrumental in saving precious lives and property, and provide a means of universal data acquisition system at low cost.

This work proposes to use plants as a monitoring system for the three parameters of focus: O_3, NaCl and H_2SO_4. Plants, available everywhere freely, can act as a cheap and versatile sensor system if we know how they react to certain stimuli. Such reaction can be sensed, categorized and then relayed through a wireless system to a nearby data processing station, thereby enabling the monitoring of a large geographical area.

1.5 The Plant – environmental stimuli reaction mechanism (Electrical Signals)

Studies have shown that plants react to their environment by producing an electrical signal, which may control various physiological processes [47]. Although plants lack a complex nervous system like animals, they nonetheless produce electrical pulses called the *Action Potential* (AP) for non-harmful stimuli [47]. Similar electrical impulses for harmful stimuli such as burning or cutting, are called the *Variation Potential* (VP). The mechanism to transport both these electrical signals occurs through different cells in the plant body. In addition to these signals, plants also produce a sub-threshold response locally, called the *Local Electrical Potential* (LEP) [47]. Chapter 2 gives a detailed overview of these signals, from a cellular level.

Thus if APs and VPs can be collected from plants, and why they were generated established, then this would be a step closer to a plant-based monitoring system.

1.6 Thesis organization

Chapter 2 presents a literature review covering:

- plant electrophysiology, plant structures and how they support electrical signal response to any stimuli, various transport systems acting within the plant body, types and mechanisms of electrical signals generated by plants.
- all measurement techniques available to extract electrical signals from plants.
- types of stimuli which have been used so far to generate electrical signal response in plants.
- signal processing steps employed so far to analyse plant electrical signals.

Chapter 3 presents the extraction of information from the raw plant electrical signal about an external light stimulus, by using black-box modelling on experimental data. The data

acquisition setup is described in detail. The results show that there is sufficient information within the extracted plant electrical signals about the *magnitude* and *on-off time* of the stimulus applied.

Chapter 4 addresses the question of whether there is enough information about the *type* of stimulus applied (involving more complex stimulus) embedded within the raw plant electrical signals? Under laboratory conditions, plant electrical signals were obtained for Ozone (O_3), Sulphuric Acid (H_2SO_4) and Sodium Chloride (NaCl) as external stimuli. A mapping of such a stimuli-electrical signal was the first of its kind, filling a gap in our knowledge. Also, if a small window of the *time series* of the raw signal contains enough information about the stimuli, then a sensor system could be designed where the decision time about the environmental stimuli would be small, since any segment of the incoming plant electrical response would indicate the stimuli affecting it. The binary classification methodology is presented in detail. Using simple *linear* classifiers, accuracies of around 71% was achieved.

Chapter 5 addresses whether raw signals could be used in a multiclass classification setting, in order to extract information about the type of stimuli applied. Effort is made to improve classification accuracy and generalization. A decision tree-based multiclass classification scheme was designed, that explores five different types of classifier and 15 features. The results were evaluated for independent datasets (*prospective study*), around 90% classification accuracy being achieved. Chapter 5 then explores whether stochastic parts of the plant electrical signals holds more information about the type of stimuli, and thus improve classification accuracy. The stochastic parts were isolated by appropriate filtering. The best classification accuracies were found to be ~93% (retrospective study) and ~89% (prospective study).

Chapter 5 also presents a comparative analysis of multiclass classification (employing decision trees) of externally applied stimuli using both deterministic (raw) and stochastic (filtered) parts of the plant electrical signal.

Chapter 6 explores feature extraction from the entire duration of the raw plant electrical signal response, to obtain as much information as possible. Four different types of fitted curves and their coefficients are considered. Results of around 90% classification accuracy were obtained on a retrospective study.

1.7 Original contributions to scientific knowledge

To our knowledge, this is the first exploration of how the electrical signal response from plants could be used for detecting external stimuli. More specifically, the objectives of this work are:

Objective 1: To show a mathematical relationship can be established, in the form of input-output models, between raw plant electrical signals and a basic stimulus such as light pulses. This relationship should be able to quantify the strength and time of application of the stimulus.

In Chapter 3, we show that there is enough information within the raw plant electrical signal response about the time of application as well as amplitude of incident light pulses, by using black-box modelling techniques. Using this method, it is established that there exists a clear relationship between the incident light pulses (stimulus) and the resulting electrical signal responses. Three best models are found from a broad range of standard system identification models which are explored.

Objective 2: To show that the raw plant electrical signals can be used to distinguish between more complex stimuli using simple binary classification techniques.

In Chapter 4, we apply binary classification methods and show that there is enough information within the raw plant electrical signal response, about the type of stimuli applied to the plants under laboratory settings. The classification accuracy achieved is around 70%. Since this was a prototype exploration, simple binary classification techniques are used.

Objective 3: Following the success of a step-by-step approach to binary-classify complex stimuli from raw plant electrical signals, to explore a way to successfully classify applied stimuli from raw plant electrical signal response in a multi-class classification setting.

Also explore whether the classification results can be improved by focussing on the stochastic part of the time series.

In Chapter 5, we design a custom decision tree for multiclass classification and show that external stimuli can be classified with an average accuracy of 90%. The results are validated by prospective study. We present the features and classifiers which will produce consistent good results for both retrospective and prospective study.

In Chapter 5, we also show that multi-class classification is possible from only the stochastic part of the response, obtained using appropriate filtering. We also show that raw electrical signals produce marginally better classification results than filtered electrical signals.

Objective 4: Explore the possibility of using the entire time series (rather than windowing) for extracting features for classification of the stimuli applied to the plants.

In Chapter 6, we show four different types of curves fitted to extract the coefficients, which are used as features for classification with accuracies around 98% in retrospective study. A prospective study is also done to validate the decision tree structure.

These four objectives are the original contribution to scientific knowledge and have been disseminated through relevant publications.

1.8 Publications

Published

1. S. K. Chatterjee, S. Ghosh, S. Das, V. Manzella, A. Vitaletti, E. Masi, L. Santopolo, S. Mancuso, and K. Maharatna. (2014). Forward and Inverse Modelling Approaches for Prediction of Light Stimulus from Electrophysiological Response in Plants, *Elsevier Measurement*, 53, pp. 101-116.

2. S. K. Chatterjee, S. Das, K. Maharatna, E. Masi, L. Santopolo, S. Mancuso, and A. Vitaletti. (2015). Exploring strategies for classification of external stimuli using statistical features of the plant electrical response, *Journal of The Royal Society Interface*, 12(104), p. 20141225.

3. S. Das, B. J. Ajiwibawa, S. K. Chatterjee, S. Ghosh, K. Maharatna, S. Dasmahapatra, A. Vitaletti, E. Masi, and S. Mancuso. (2015). Drift removal in plant electrical signals via IIR filtering using wavelet energy, *Elsevier Computers and Electronics in Agriculture*, 118, pp. 15-23.

4. S. K. Chatterjee, S. Das, K. Maharatna, E. Masi, L. Santopolo, I. Colzi, S. Mancuso and A. Vitaletti. A Decision Tree Based Classification Strategy to Detect External Chemical Stimuli from Plant Electrical Response, *Elsevier Sensors & Actuators: B. Chemical*, In Press (2017)

In draft

S. K. Chatterjee, K. Maharatna. Feature Extraction Using Curve Fitting on Plant Electrical Signal Response for Classification of External Stimuli.

1.9 Limitations of this work

The work presented here is a part of an EU-funded project (PLEASED, grant: 296582) involving five partners, one of whom was responsible for carrying out the experiments and providing the data. Hence the limitation with availability of data influenced some of the choices and justification on methodologies used.

The long term objective of this consortium is to build a stand-alone sensor node which should be able to accurately detect, process and map the electrical signal from the plant to the stimulus affecting it. This sensor node may thereafter be connected to a *Wireless Sensor Network* (WSN) to realize the goal of a holistic monitoring system for the surrounding environment for factors such as harmful environmental gases, wildfires, pesticide poisoning, etc. Thus the work presented in this dissertation on classification results achieved, features proposed, classifiers used etc. forms crucial information for other project partners to use towards the goal of designing and implementing a plant electrical signal response based ESN.

2 Literature review

2.1 Introduction

It is scientifically established that a plant senses its environment for optimal growth and there is a mechanism of electrical signal response to the surrounding environment. This chapter reviews how such electrical signals are generated, how such signals are recorded, and what type of signal processing has been undertaken on such signals.

To realize the goal of using plants as biosensors that can monitor the environment, the first step is to understand how the electrical signals are generated in a plant when an external stimulus is applied to it. An understanding is needed of how such electrical signals could be processed with standard signal processing techniques, in order to extract appropriate information from it. This chapter surveys the literature on the mechanisms behind the generation of electrical signals in plants, established measurement techniques of such electrical signals, effects of stimuli applied to plants on the electrical signals, and signal processing methods applied to electrical signals. It is useful to know the kinds of stimulus that have been used on plants and the nature of the electrical signals generated, and what kind of information has been extracted from such signals. Through this review, the gaps in our knowledge are identified, some of which will be addressed through research presented in subsequent chapters.

2.2 Discovery of electrical signal generation in plants

In 1872, Burdon-Sanderson [48] and then in 1888, Charles Darwin, demonstrated the generation of electrical responses in insectivorous plants [49]. Six years after this, Darwin provided evidence for the propagation of chemical signals (e.g. hormones) in plants. This evidence compelled other researchers to focus mainly on chemical signals in plants. The common belief which arose was that animals had *neuro-electrical* signals (electrical phenomena generated by the nervous system) and plants had both chemical and electrical signals [50]. However, in the early 20^{th} century, through discovery of animal hormones and with the emerging field of endocrinology, more evidence surfaced which suggested that animals also had chemical signals [50]. This gave a rise to the concept that both animals and plants generate some form of electrical and chemical signals as a response to external stimulus. In 1924, J. C. Bose isolated the vascular bundles of a *Fern* to show that physiological events, such as those present in animal nerves, triggered excitation, which

travelled as electrical signals [51]. Much later, in 1973, Barbara Pickard published a review of many previous publications related to generation of *Action Potentials* (AP) or electrical pulses in plants [52].

This review included works such as Burdon-Sanderson's demonstration of plants like *Dionaea muscipula* (Venus flytrap) using AP to control its rapid leaf movements. This work by Burdon-Sanderson was extremely important as it elaborated on many important aspects of plant electrical signalling and included characteristics of the electrical signal response (i.e. the AP in plants) such as the rise times, rate of propagation, and duration. Barbara Pickard's review raised several important questions such as: Is phloem transport in plants regulated by AP? Does AP release plant hormones or short-range control substances (to control different physiological aspects within the plant body)? and several other such questions.

It was observed that an electrical signal is generated from an external stimulus as a part of response not only in *lower plants* (a collective term to describe some of the oldest organisms on earth such as mosses, liverworts, lichens, fungi and algae) such as *Chara corallina*, which belongs to the family of *Characeae,* and in sensitive (to mechanical stimulation) *higher plants* such as *Mimosa pudica* and *Venus flytrap*, but these electrical signals also play an important role in *non-sensitive* higher plants such as *Cucumis sativus* (cucumber) [47]. In 1984, an important discovery by Schroeder established that ion channels, similar to those which exist in the animal body, are present in plants as well [53]. This led to further studies to find the nature of ion channels, which aid in generating electrical signals in plants and to verify whether the plant ion channels are in any way similar to those found in animals [53]. Various experiments showed that ion channels in plants, activated by bioelectrical activities due to external stimulus, may be different to those found in animals [54].

This important historical background, coupled with an understanding of the mechanism behind the electrical signal generation and conduction in plants, will enable better use to be made of plants as environmental sensors.

2.3 Environmental sensing mechanism in plants

Plants sense their surrounding environment and try to use it to their advantage. *Sunflower* plants are one of the best examples, because they show a sense for direction of light by bending themselves to maximise photosynthesis. It has also been known for many years that plants are capable of sensing the climate and temperature accurately by flowering at different

times consistently. They also show *sleep movements* by changing orientation of their leaves, by maximising absorption of light during the day and by minimizing heat loss during the night, thereby showing a sense of day and night [47]. They also adjust their height, leaf distribution, etc. according to their growing condition, i.e. a difference in growth when placed in full sunshine outside or in the shade.

Plants also show sensitivity to mechanical stimulus. Insectivorous plants such as the *Venus Flytrap* or *Sundew*, which usually live in nitrogen- and mineral-depleted regions, assimilate the necessary nutrients by capturing and digesting insects [47], [55]. These insects get attracted to the plant's trap due to its colour or shape or smell, and get stuck in them. The interesting feature of the traps are, the more effort the insect puts in getting out (providing mechanical stimulus), the more the trap closes [47]. This is possible due to the sensing by the plant through its hair-like tentacle tips in the traps, which sends out some form of information carrying signal [47]. In the *Sundew*, such information causes the tentacles to bend and push the insect towards the centre of the leaf. Research has shown that, although information carrying signals do not get transmitted to the neighbouring tentacles, a slower hormonal signal causes the tentacles to wrap around the insect and make a secure trap around it. The secretory cells, situated nearby, then produce digestive enzymes to digest the trapped food [47], [55].

Thus, there is a neat sensing mechanism which helps higher plants like sunflower maximise photosynthetic abilities and insectivorous plants to capture their food in a proactive manner. This sensing mechanism, which is a closed-loop external stimulation and response mechanism, is adopted by the plant to adapt to varying natural conditions and optimize its growth.

To gain some understanding of the theory behind such sensitivity of plants to various stimuli, studies were conducted on the *Characean algae,* such as *Chara* and *Nitella* [55], which are considered as ancestors of all higher plants. APs were induced in *Characean* cells by various stimuli such as sudden changes in temperature/pressure, ultra-violet radiation, mechanical stimuli, odorants, and even a depolarizing current [47], [55].

Just as in animal nerve cells, APs induced in *Characean* cells were independent of the stimulus strength [55]. There was also a refractory period, in which a second AP was not generated by the application of a stimulus [47], [55]. The AP propagated in both directions from the point of impact, i.e. where the plasma membrane was first depolarized. The

propagation rate depended on the type of medium and was found to be faster than sponges but slower than animal nerves [47], [55]. An AP can only be transmitted if there is a sufficient flow of ionic current through, and sufficient to depolarize, the next membrane, although below the threshold. If the adjacent plasma membrane becomes depolarized to a value above the threshold, a new AP will be generated [47], [55].

So, behind some of the easily observed characteristics of the higher plants such as *Sunflower*, *Venus flytrap*, and through some experiments conducted on lower plants such as *Chara* and *Nitella*, some information about Action Potential is forthcoming that explains the possible reaction of plants to external stimuli.

2.3.1 Plant sensing structure

This section discusses the basic structures of higher plants and how they support generation and transmission of electrical signals that carry information about the stimulus which caused the signal.

2.3.1.1 Epidermis

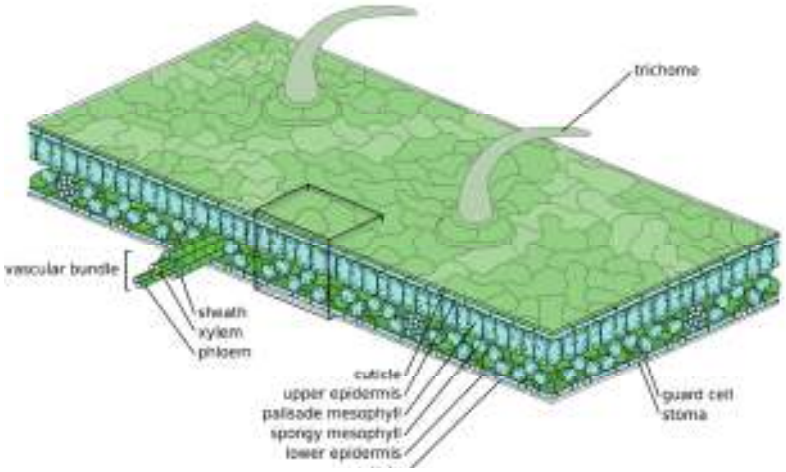


Figure 2.1: Plant epidermis (image source: wikipedia)

The outermost cell layer of the primary plant body is called the *epidermis*. The epidermis is the main component of the dermal tissue system of leaves, stems, roots, flowers, fruits, and seeds [56]. It is usually a transparent layer as its cells lack chloroplasts, except for the *guard cells*. The cells of the epidermis are structurally and functionally variable.

Epidermal cells are tightly linked to each other and together provide the mechanical strength required as well as protection to the plant. The epidermal cell walls of the plant organs above the ground contain *cutin*, and are covered with *cuticle*. This cuticle helps in reducing water loss to the atmosphere [56].

2.3.1.2 Stomatal cells

On the surfaces of leaves and stems are small pores called *stomata* (Figure 2.2), which control the exchange of gases, water vapour and CO_2 between the interior of the leaf and the atmosphere [57]. The stomata are bound by *guard cells* which control the opening and closing of the pores. In this capacity they make major contributions to the ability of the plant to control its water relationship with the atmosphere and gain carbon [56], [57]. The number of stomatal pores on the epidermis and their aperture determine the amount of gas exchange occurring between the plant body and the atmosphere. Environmental signals such as light intensity, CO_2 concentration and different plant hormones, control stomatal aperture and development [57]. The term *stomatal conductance* is defined as the rate of gas exchange between the plant and the atmosphere. The stomatal conductance is higher for more stomatal apertures being open, thereby increasing transpiration (loss of water) and photosynthesis [58]. Stomatal conductance depends on density, size and degree of opening of stomata [59]. The relationship between elevated O_3 and CO_2 and stomatal conductance, density, index (ratio of stomata to epidermal cells), length of guard cells, epidermal cell size and numbers with respect to crown position and leaf types, of two silver birch clones have been studied [60]. This study provides an important insight about the sensing mechanism in plants whereby the change in environment is reflected in some form of change in physiological condition of the plants. This study also reported the relationship between stomatal characteristics and leaf spot disease, caused by *Pyrenopeziza betulicola* (a type of fungus), which again provides an invaluable link between what is affecting the plant and the changes to its physiological conditions, thereby pointing to some form of sensing mechanism.

2.3.1.3 Guard cells

Guard cells, usually found in two different shapes (Figure 2.2), are located in the epidermis of the leaves and surround the stomata. They are exposed externally (to the atmosphere) as well as internally (i.e. interior of the leaf) [61]. Guard cells are known to be sensitive to light, CO_2 and temperature fluctuations. The phytohormones *abscisic acid* or ABA (playing a regulatory role in different physiological processes), and *auxin* (plant growth regulator, promoting cell elongation), as well as *reactive oxygen species* (ROS), also affect guard cells [61]. Guard cells are known to help regulate stomatal apertures through influx or efflux of potassium (K^+) ions. When K^+ ions are pumped out of the guard cells into the neighbouring cells, the water molecules follow the concentration gradient (neighbouring cells have higher concentration of solutes now) and move out of the guard cells as well. This causes the guard cells to *deflate*

and close the stomata, thereby controlling rates of CO_2 uptake and water loss and hence influencing photosynthesis and the water content status of the plant [62]–[64].

Guard-cell volume increases due to the increased uptake of ions such as K^+ along with water molecules. Higher concentrations of K^+ in the guard cell wall have been found as a result of switching light conditions (darkness/light) or as a hormonal response, e.g. ABA [61]. The ABA starts cystolic calcium (Ca^{2+}) increase in the guard cells, which activates two types of anion channels – *slow* activating sustained (S type) and *rapid* transient (R type). Both encourage release of anions (negatively charged ions) from the guard cells, thus causing depolarization. This in turn deactivates the inward rectifying K^+ channels and activates outward rectifying K^+ channels. Through this mechanism, there is a loss of turgor in guard cells and the stomata closes [63], [65]–[67].

Figure 2.2: Two different types of stomata in plants [57]

2.3.2 Generation of electric potential through ion-flow in plant cells

Understanding the ionic conduction and various transport mechanisms within plants will help an understanding how different stimuli generate the resulting electrical output signals from the plants. To begin with, the different routes of transport of water, nutrients and ions between various parts of the plant cells and organs need to be addressed.

In Figure 2.3, three transportation routes can be seen – *transmembrane route* is across the cell membrane pathway, *apoplastic pathway* is the path through the cell walls, and *symplastic pathway* is the path through the tiny opening in the cytosol called the *plasmodesma* [68]. The plant tissues have two compartments on either side of the plasma membranes. The cytoplasmic or symplastic compartment is on the inner side of the plasma membrane and the

apoplastic compartment is on the outer side of the plasma membrane. The plasmodesma acts as a tiny channel which connects the cytoplasm of plant cells to form a *symplastic continuum* [68].

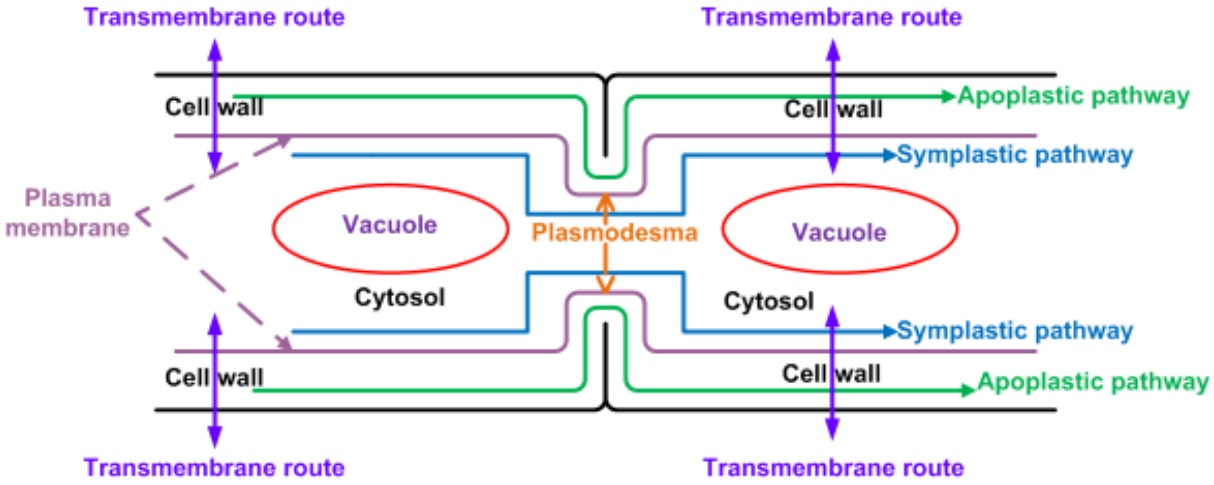


Figure 2.3: Connecting pathways between plant cells

There are three important points for understanding the ionic movement across plant cell membranes. These are [69]:

- The plasma membrane of a plant cell is selective in its permeability to ions and metabolites,
- Proteins are required to transport inorganic ions (K^+, Ca^{2+}, Na^+, Cl^- etc.), organic solutes (e.g. sugars) and protons (H^+) across the plasma membrane,
- Enzymes, such as *Adenosine Triphosphatase* (ATPase) which help decompose ATP into *Adenosine Diphosphate* (ADP) – an essential organic compound required during metabolism, along with channel proteins and co-transporters which are present in the plasma membrane, all help transport important ions and solutes required by the cell.

By looking into the ionic movement mechanism of a typical guard cell, the opening and closing of stomata can be understood further.

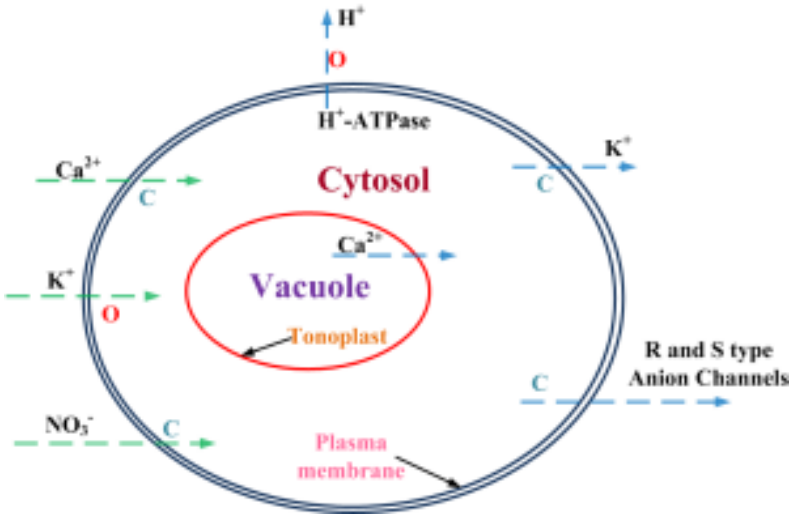


Figure 2.4: Ion Channels and Transporters in Guard Cells [63]

In Figure 2.4, different ion channels within the guard cells of plants can be seen. These channels either contribute to the closing (marked by C) or opening of the stomata (marked by O). When there is a reduction in the stomatal aperture, the R(apid) and S(low) anion channels are responsible for Cl^- and NO_3^- efflux from the guard cells [63].

Once the channels are opened, selective pores on the plasma membrane are formed through which ions can move without inducing any change of the proteins. The ions then move with maximum velocity through the open channels (approx. 10^7 ions per second) [64]. Since ion channels are involved in controlling the membrane potential, the transduction of signals in plants is very similar to the transduction mechanism in animals. The channels are also involved in the uptake of ions from the soil, secretion of ions into the xylem sap, etc. [70].

Depending on the opening or closing of the ion channels, also known as *ion channel gating*, the plasma membrane can either *depolarize*, *repolarize* or *hyperpolarize*. During membrane depolarization (membrane potential turning positive), K^+ loss occurs through outwardly rectifying K^+ channels which are activated by the membrane depolarization event. This influx and efflux of K^+ ions occur to balance the plasma membrane back to its resting potential. The difference in the membrane potential (between the two sides of the membrane) occurs due to active transport (i.e. ions moving against its concentration gradient with the expenditure of energy) or passive diffusion of ions (i.e. ions moving from high concentration towards lower concentration without any expenditure of energy) [71]. When membranes hyperpolarize, i.e.

membrane potential turns more negative, K^+ uptake starts passively, via inwardly rectifying K^+ channels.

2.3.2.1 Active and passive transport

Plant cells take in solutes and water molecules through their membranes (plasma membrane and tonoplast) through the combination of *active* and *passive* transport.

Passive transport is described as the mechanism in which a solute moves down its electrochemical potential gradient (i.e. without any expenditure of energy), whereas in active transport a solute moves up against its electrochemical potential gradient (i.e. with some energy expenditure) [72].

The energy expended during active transport may either be obtained from the hydrolysis of ATP or *pyrophosphate* (a high-energy polymer of phosphate), or obtained from the movement of a co-transported solute. It can also be taken from a coupled solute going down its electrochemical gradient. Coupling between solutes moving in opposite directions (the downhill movement of one solute to the uphill movement of another) is also a common phenomenon in membrane transport [72].

The plasma membrane and the tonoplast contain H^+ (proton) pumps which move H^+ ions across these membranes by using energy released from hydrolysis of ATP or pyrophosphate. A positive charge is pumped out of the cytoplasm of the cell, thereby establishing a large membrane voltage (inside negative, outside more positive) along with a steep pH gradient. A very large electrochemical gradient can set in, around 400 mV for H^+ ions. It is suggested that the proton pump in plasma membranes moves one H^+ ion per hydrolyzed ATP [72].

This H^+ electrochemical gradient helps in driving coupled active movement, called *secondary active* transport, of other solutes across the plasma membranes. Secondary transport indicates the reliance on an already established gradient of another ion. For example, the uptake of a Cl^- ion is coupled to the influx of two H^+ ions. Cl^- ions must involve *active transport* across the membrane as the membrane potential is usually more negative than the equilibrium potential for Cl^-. Coupling of each Cl^- ion to two H^+ ions results in a net positive charge transfer of +1 into the cell. This is energetically more favorable than a zero or a negative net charge [72].

Along with H^+, even Na^+ ion gradients drive secondary transport in plant cells. This was found in *Chara* cells as well as in some higher plants. Na^+-driven transport may have evolved

in plants due to the surrounding alkaline conditions (high external pH). In such cases the H^+ gradient may not have been sufficient to drive the ionic transport by itself [72].

Recent studies [73]–[75] have also pointed out the role of Calcium (Ca^{2+}) channels in initiating a large number of processes in higher plants, such as light- and hormone-regulated growth and development, regulation of gas exchange, formation of buds, etc. Intracellular organelles store Ca^{2+} which is about 10^4 times the concentration of free Ca^{2+} found in the *cytosol*. Thus when the Ca^{2+} channels in the organelle membrane open up, there is a rapid rise in the cytosolic Ca^{2+} concentration [74].

It has been suggested that the influx of Ca^{2+} in the cytosol plays a crucial role in initiating ABA-induced stomatal closure during water stress. An increase in the cytosolic Ca^{2+} concentration may block inward rectifying K^+ channels, thereby inhibiting stomatal opening. The stomatal closure thus may result from the activation of the calcium-stimulated chloride channels [66].

High environmental NaCl concentration has also been reported to elicit an increase in cystolic Ca^{2+} in *Populus euphratica* (Desert Poplar) [76] and *Arabidopsis thaliana* (Thale Cress) [77]. This increase in Ca^{2+} is generated by extracellular influx and release from intracellular stores, e.g. vacuole.

2.3.2.2 Phloem and Xylem transport

The ionic channels in guard and other plant cells help regulate various physiological functions (such as opening and closing of stomata). But how do all these ions get transported to various parts of the plant body as per various requirements of the cells? To answer this question, the transportation system within the plant which consists of the *phloem* and *xylem* needs to be understood.

The phloem and xylem form the important long distance transport system of water and solutes throughout the plant, traversing from the roots through to the leaves. The xylem helps in the transport of water and nutrients from the roots to the shoots, as well as providing structural support, and the phloem helps transport the products of photosynthesis such as sucrose and amino acids, from a source (e.g. mature leaves) to a destination (e.g. roots, growing fruit, shaded leaves) [78]. The phloem consists of several types of cell such as *parenchyma*, *sclerenchyma* and *sieve tubes* (phloem vessels).

So far, ion channels permeating the cell plasma membrane and various transport mechanisms of solutes and water molecules within the plant body have been discussed. But how does this ionic conduction and transport system translate into electrical signal propagation? In other words, what happens when the plants sense an external stimulus? It has been suggested that the electrical potential generated within the plant body due to stimuli (e.g. wounding shocks, insect attacks) assist in rapid communication with other parts of the plant for a suitable response to a similar stimulus. Three forms of electric potential have been identified – *local electrical potentials* (LEP), *action potentials* (AP) and *variation potentials* (VP) – and it is suggested that AP propagates through the phloem and VP propagates through the xylem towards the phloem [79].

2.3.3 Electrical signals in plants - types and features

Plants have the ability to change the internal condition of their cells, tissues and organs from the effect of various environmental stimuli. Excitability in plants occurs due to the high sensitivity of protoplasm and all cell organelles to any natural and electrochemical effects [80]. It has been reported that any type of stimulus causes an initial influx of Ca^{2+} which triggers a Cl^- efflux via anion channels and leads to massive and quick plasma membrane depolarization [25]. Slow repolarization of the plasma membrane up to its resting potential takes place due to activation of K^+ efflux. Due to ion channel gating, this electrical wave (due to influx and efflux of ions) propagates through the sieve tubes.

A wounding/injury induces and sustains an LEP that stops a few millimetres from a dying cell, and creates an AP and VP [54], [81]. All three signals arise from a transient change in the membrane potential of the plant cells (depolarization/repolarization phases), but only VPs and APs make use of the vascular bundles to systematically spread the signals through the entire plant body [54], [81].

Local Electrical Potential (LEP): This is the sub-threshold response of the plant to a difference in environmental stimuli such as soil, water content, temperature, air, humidity. The effect of LEP is local and is not transferred to other parts of the plant [47], [54], [81]. Although it has been theorized that LEP exists at the local site where the change in stimulus is sensed, not much has been published about it, since most of the literature focuses on AP and VP.

Action Potential (AP): This is induced by stimuli which do not do any permanent damage to the plant, e.g. cold, mechanical, electrical stimuli. APs are generated after a certain threshold of stimulus is crossed, and once generated, do not alter in amplitude or shape despite any increase in the strength of the stimulus. The AP is transmitted in all directions through the *plasmodesmata* and travels at constant velocity and amplitude long distances in the plant body through the sensitive *phloem* cell membranes. Following the generation of an AP, the plant cell plasma membrane enters a *refractory period* during which no second impulse will be transmitted [47], [54], [81].

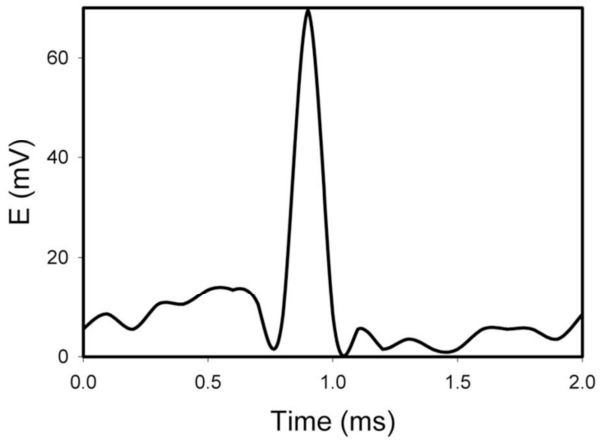


Figure 2.6: AP's in soybean (Glycine max (L) Merrill) induced by changing the direction of white light irradiation [82]

Figure 2.6 shows an example of an AP generated in a soybean plant when light was shone from different directions, a phenomenon known as *phototropism* [82]. Positive or negative phototropism indicates the plant bending towards or turning away from the source of light respectively (e.g. sunflower). Four responses are said to occur during phototropism – receiving the directional light signal, transduction of electrical signal response, transformation of the electrical signal response to a physiological response, and production of a directional growth response [82]. Thus, APs are generated under different environmental and internal influences such as touch (e.g. *mimosa pudica*), changes in light, cold shock, cell expansion during growth, all of which trigger a voltage-dependent depolarization spike in an all-or-nothing manner [83].

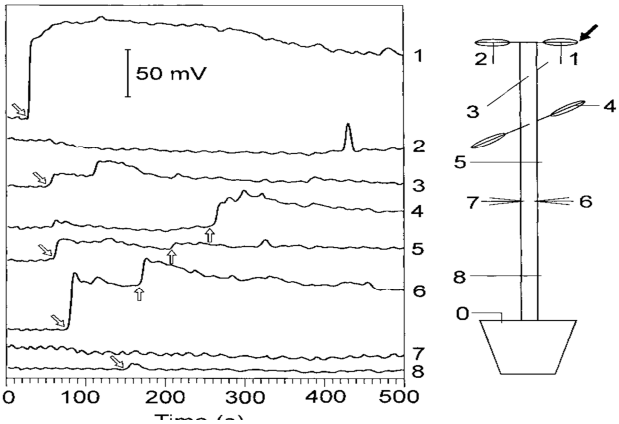


Figure 2.5: Records of electrical potential changes of VP type after thermal stimulation of a leaf of Helianthus annuus [84]

Variation Potential (VP): This signal is generated when the stimulus is damaging or detrimental to the plant such as burning or cutting. VP, or slow wave potentials, gets reduced in magnitude and speed as it travels from

the site of injury (stimulus) and its magnitude and shape varies with the intensity of the stimulus.

Figure 2.5 shows recorded electrical signal response, which may be VP, when the plant was subjected to heat shock (burning) [84]. The transmission of VP depends on xylem tissue tension inside the plant, which appears as local change in a hydraulic pressure wave or chemicals transmitted [80]–[82], [84]. Thus, a VP occurs when a hydraulic pressure wave travels through the xylem after some form of damaging stimulus affects the plant such as organ excision or local burning. During a VP event, depolarization starts with an increase in turgor pressure experienced by the plant cells.

Clearly, VPs and APs are different, between the factors stimulating them, the ionic mechanisms of the depolarization/repolarization phases of the cells, and the transduction pathways of their propagation.

2.4 Measuring electrical signals in plants

The two most common techniques employed to measure electrical signals in plants are [81]

- Intracellular measurement
- Extracellular measurement

Other techniques often used are

- The *Aphid* technique
- Patch-clamp recording
- Non-invasive microelectrode vibrating probe technique
- Non-contact measurement using optical recording [85]

Intracellular measurement can be used to directly record the value of an individual cell plasma membrane potential, while *extracellular* measurement can be used to detect the spatio-temporal total of the depolarization/repolarization process in a large group of plant cells. Extracellular potential measurements are carried out on the surface of higher plants. One of the advantages of this technique is that it can be used to record potential differences over long periods (e.g. several days) as it does not involve any cell altering electrolytes [81].

In higher plants, two types of extracellular measurement can be performed. These are

- Measurements using inserted metal electrodes
- Surface recordings

Measurements using metal electrodes inserted into the plants causes wounding reactions which can conceal reactions to the external stimulus being studied. Therefore such electrodes are required to have thin metal tips made of Platinum (Pt) or Silver/Silver-Chloride (Ag/AgCl)-wires of 0.4 to 1.0 mm in diameter [81]. When inserted into the shoot or leaf vein of the plants, the metal tips of the electrodes come into contact with tissues covering a larger group of cells and can support long-term recording of electrical signal responses [81]. This is shown in Figure 2.7 (1).

The surface recordings are usually done using calomel electrodes, which adopt a suitable ionic solution and connect a salt bridge between the electrode and the plant, and are non-invasive and physically stable [54]. Surface measurement electrodes are usually moistened with 0.1 % (w/v) KCl in agar and then wrapped in cotton to provide appropriate contact with the plant surface. Alternatively, Ag/AgCl pelleted electrodes can also be used for surface recordings since can be connected to the plant surface by means of a conductive aqueous gel (commonly used in ECGs) [81].

Intracellular measurements are performed using glass microelectrodes and this technique can be used to measure potentials over a few hours. This short duration arises from some of the electrolytes employed within the electrode being diffused within the measured cell and changes its original bio-electric condition [81]. If the aim is to observe and study the bioelectrical activity of plants at cellular level, then intracellular measurement techniques may be employed, which usually involves inserting microelectrodes, with a tip diameter lower than 1 μm, into the surroundings of a living cell or into the cell body [54]. The reference electrode, usually made of Ag/AgCl, is typically placed in a bath solution surrounding the cell. The bath solution and its concentration is appropriately chosen according to the conditions of the cells to be studied [54], as shown in Figure 2.7 (4). The electrodes are connected to a high input impedance amplifier and once the amplifier reading is zero with both electrodes being outside the cell, micromanipulators are used to carefully insert one of the microelectrodes into the cytoplasm/vacuole of the cell [81].

An important mathematical model defining the relationship between the intracellular and extracellular measurements of electrical activity in *Vicia faba* L. has been reported in [86].

In order to record electrical signals with high velocities, it is essential to measure the signals in the phloem cells of the plant. The phloem offers a very low resistive channel through its

sieve pores, thus enabling signals to be transmitted over long distances [81]. However, these phloem cells are located inside the plant body making insertion of microelectrodes difficult. Often, the microelectrode tip is not correctly inserted into the phloem, as revealed by microscopic checks after an experiment with dyes inserted into cells [81]. Thus the *Aphid technique* was created in which an aphid (small sap sucking insects, also called *plant lice*, members of the superfamily *Aphidoidea*), is allowed to settle down on a mature leaf overnight (Figure 2.7 (2)). The following day, a laser pulse is used to sever the aphid from its stylet which it used to puncture the sap in the phloem as shown in Figure 2.7 (3). Due to high pressure in the phloem, the stylet exudes sieve-tube sap. To this exuding sap, the microelectrodes are attached. The stylet acts as an effective salt bridge between the cytoplasm and the microelectrode [81]. The aphid technique has been used to study electrical signalling in poplar shoots which were stimulated by flame as well as cold shock [87].

Patch-clamp recording can be used to measure ionic currents from a single cell to a group of cells, yielding knowledge about ionic channels in plants that has been growing for many years [54]. The patch-clamp technique involves sealing glass capillaries with tip diameters of around 1 µm to the surface of protoplast or vacuole membranes and then bursting open a tiny hole in the membrane with a short pulse of around 1 Volt (applied for a few milliseconds). Through this newly created rupture, the solution within the fine glass capillaries replaces the contents within the vacuole or the cytoplasm. This method is termed *whole-cell recording*. In the *inside out patch* recording mode, the inner surface of the plasma membrane comes into contact with the solution in the glass capillary tube. Through these modes, movement of ions can be tracked using their electronic charges, constituting the electrical signal in the form of current. Opening and closing of ion channels are indicated by a square wave of current values across the patch [88].

The *non-invasive microelectrode vibrating probe technique* comprises a reference electrode and an ion selectivity microelectrode, which includes a glass pipette, an Ag/AgCl wire, an electrolyte, and a liquid ion exchanger. The microelectrode vibrates and determines the voltages between two positions, usually a few microns apart, and is dependent on the ionic concentration gradient. The concentration gradient can be determined by the voltage gradient using a curve of pre-calibrated voltage versus concentration of the microelectrode [54]. The non-invasive microelectrode vibrating probe can be used to measure the ionic or molecular activity without invading the cell, in contrast to the patch clamp technique.

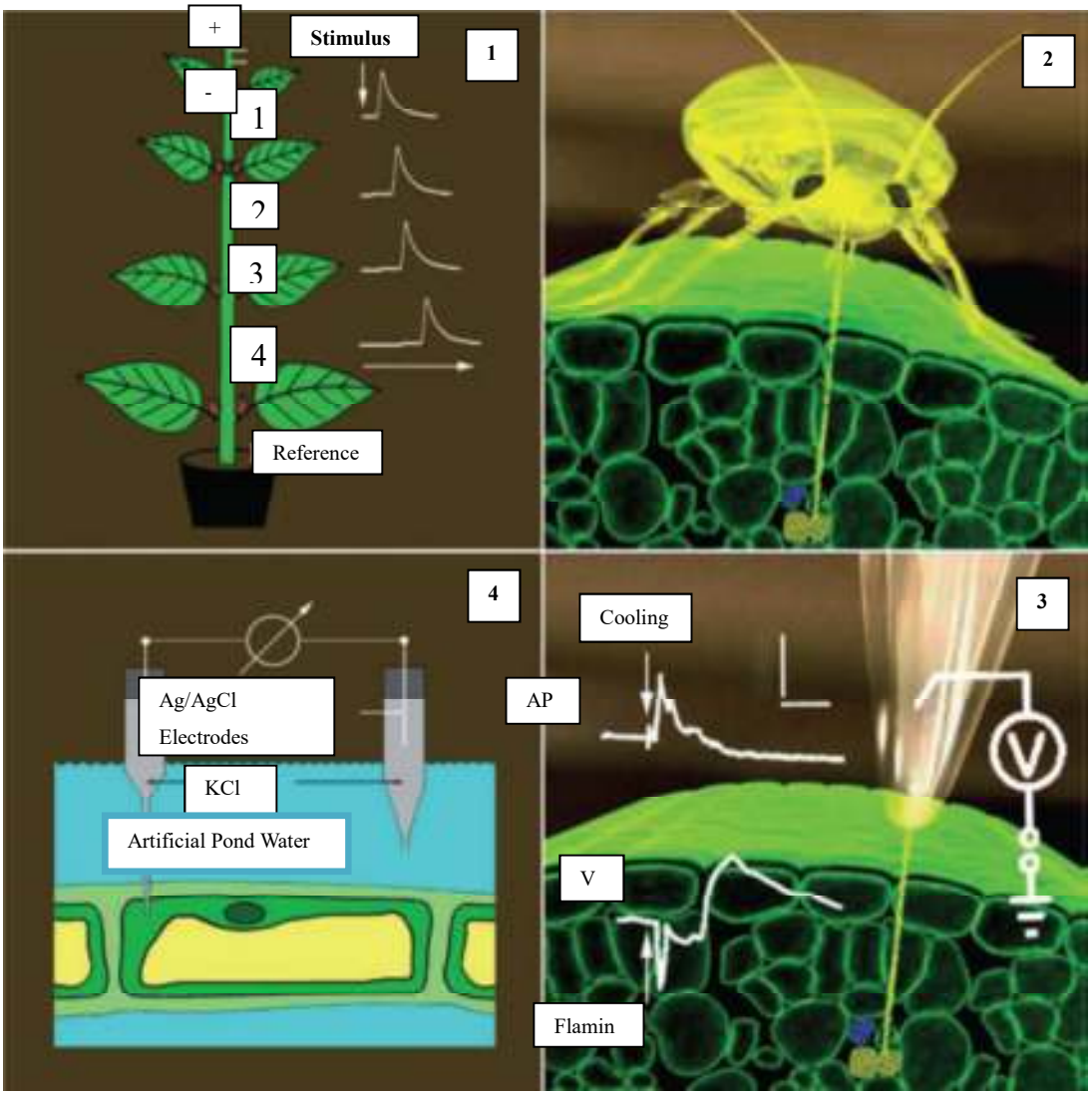


Figure 2.7: Measurement techniques for electrical signals in plants [81]

Optical measurements of animal bioelectrical activity, using voltage sensitive dye (VSD), has been increasingly used and are now being considered for studying plant electrophysiology [89]. The methodology behind the optical measurement system is to capture and analyse the optical signals emitted by the VSD which is bound to the plant cells. A charge coupled device (CCD) camera with a high spatial and millisecond resolution captures the optical signals which are then converted from slow electrical signals in the plants [89] [85].

2.5 Effects of stimuli on plant electrophysiological signals

Environmental stimuli such as changes in light or mechanical wounding, can induce electrical signals at any site of the plant's *symplastic continuum* (the connected pathways between symplast, which is the inner side of the plasma membrane of cells).

This section reviews some stimulus specific electrical signals produced by the plants. It provides an overview of the kind of experiments conducted on plants to observe the generation of electrical signal response to various stimuli which, as shown later, will identify the gap in knowledge.

2.4.1 Light as a stimulus

One of the fundamental requirements of plants to survive and grow is light, which is used for photosynthesis. So there is some form of fundamental sensing mechanism within plants directed towards light, which is used to control various physiological processes.

Light-induced generation of APs has been observed in liverwort (*Conocephalum conicum*) when it was shaded, and its *thallus* cells were found to hyperpolarize, whereas they depolarized upon re-illumination [81]. APs can be transmitted via plasmodesmata to other cells of the symplast [81]. Simply switching the light on/off does not induce any AP [82] (only change in direction of the light induces the generation of AP), however liverwort displayed AP when it was subjected to shade [81] (blocking of light, i.e. subjected to a change from continuous light to continuous dark). It seems to be ambiguous as to whether a simple on-off phenomenon or a changing direction of illumination encourages an electrophysiological response.

Experiments have shown that guard cells are highly responsive to light, and light induced membrane potential changes could be measured [61]. The cells were found switching from driving K^+ uptake in the light, to K^+ efflux in the dark. These were interpreted through membrane potential changes, thus enabling guard cells to increase their cytoplasmic K^+ concentration in the light and lower it in the dark. Hence an osmotic motor is established which drives stomatal movement. The K^+ concentration of the apoplast bordering the guard cell walls changes in response to hormones, such as ABA, for differing conditions of light. The concentration increased from 3 mM in light to 10 mM within 20 min during dark conditions. The concentration of Cl^-, increased transiently from 1 to 8 mM, during dark conditions. However, after 1 hour, the concentration returned to its original value. It was found that the level of K^+ and Cl^- tends to increase in the dark [61], [90]. Stomatal opening is stimulated by K^+ as well as Cl^-. The dark period induced alterations in K^+ and Cl^- concentrations are thus in opposition to the stomatal closure [54], [81], [82], [84], [91]–[93]. Wavelengths of 450±50, 670 and 730 nm have been used to induce APs in three week old soybean plants [92]. A similar result using ultraviolet (UV) and blue light radiation has been

reported [94]. Electrical signals from woody plants such as avocado, lemon, olive and blueberry, have had microelectrodes inserted deep into the trunk of the trees to study generation and conduction of AP/VP using light intensity and water availability [91], [95]. LED lighting has been used to generate electrical signals in *Sansevieria*, where changes in the electrical signals are related to the changes in the rate of photosynthesis [96].

All these studies point to the importance of light as a stimulus for plants, and how the internal sensing mechanism may take place by ionic conduction for different wavelengths of light. This is critical for determining the optimum light requirement for plants and a need-based sensor that can be employed in greenhouses. By need-based sensor is meant an automated lighting system that will provide adequate amounts of light based on the requirements of the plants.

2.4.2 Chemical as stimulus

Understanding the effects of various chemicals in generating electrical signal responses in plants is extremely important for designing a plant based sensing system. This would allow acid rain or uncontrolled use of fertilizers or pesticides for irrigation to be monitored over a large geographical area cheaply, and promote effective and timely countermeasures when the plants respond to a sudden rise in growth inhibiting chemicals in their surroundings. One approach is to measure the response of plants to the spray of commonly-used chemicals, addressed here.

Spraying a soybean plant with an aqueous solution of H_2SO_4 with pH in the range of 5.0 to 5.6 did not induce any APs [80]. However, spraying the plant with 0.1 ml solution or depositing 10µl drops of aqueous solution of H_2SO_4 or HNO_3 (pH ranging from 0 to 4.9) on leaves induced APs. The propagation speed of the APs was 55 ± 5 cm/s. These APs are shown in Figure 2.8.

The duration of a single AP was 0.2 s after treatment by HNO_3, and 0.02 s after H_2SO_4. The APs were also generated if the pH value of the soil was acidic, without spraying the plant directly. When the phloem cells are stimulated at any point, the change in transmembrane potential in cells creates waves of depolarization (AP), which affect the adjoining cell plasma membranes, initially at their resting potentials [80]. Hence, during stimulation of the phloem, an AP is propagated over the whole length of the plasma membrane and all along the phloem with a constant potential. Propagation of each of these impulses is then followed by an

absolute refractory period during which the fibre cannot transmit a second pulse [80]. This study is important from the perspective of acid rain as two of the main constituents to produce this are HNO_3 and H_2SO_4. Pure water has a pH of around 7.0 which is considered neutral, whereas natural unpolluted rainwater actually has a pH of about 5.6 which is considered acidic. This acidity is attributed to the presence of CO_2, NO and SO_2 found in the troposphere. In urban settlements, especially in the developed and industrial nations, rain water gets polluted and its pH value is somewhere around 3.0 or lower, because increased concentration of NO and SO_2 combines with water molecules to form a high concentration of acids in the rain. Since plants that were exposed to a pH value between 0 and 4.9 showed a measurable electrical signal response, this is an important contribution in understanding that plant electrical signals can be used to sense acid rain [97]–[99].

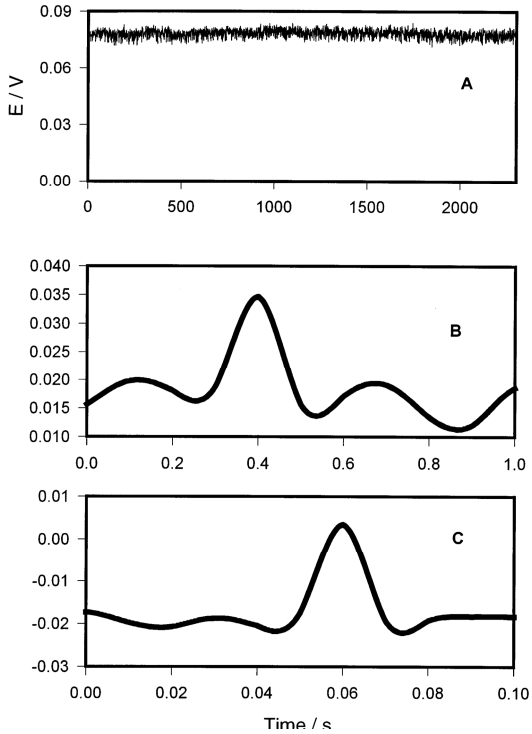


Figure 2.8: AP from soybean plant. (A) After single spray with 0.1 ml of H_2SO_4 (pH: 5.0); (B) After single spray with 0.1 ml of HNO_3 (pH: 3.0); (C) After single spray with 0.1 ml of H_2SO_4 (pH: 3.0) [80]

The high sensitivity of protoplasm and all cell organelles to any natural environmental and chemical stimuli is suggested as the reason for this excitability [80]. Addition to the soybean plant soil of an aqueous solution of *2, 4-Dinitrophenol* (DNP), a highly toxic pesticide, induced fast APs [100].

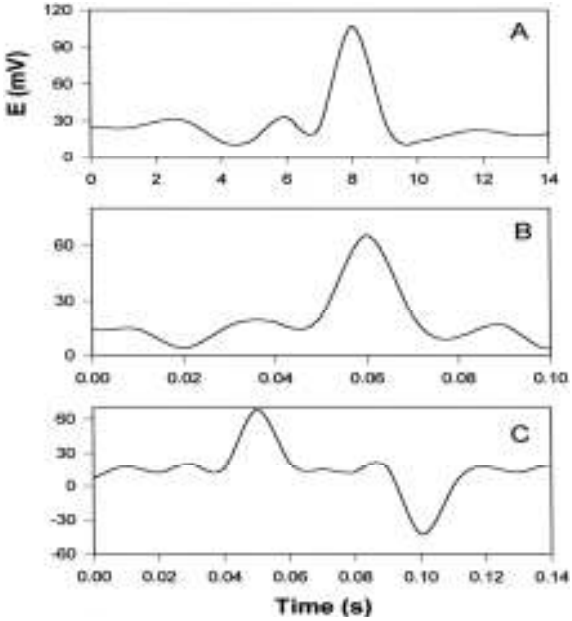


Figure 2.9: Durations of AP in soybean plant, after application of DNP [100]

Environmental exposure to DNP usually occurs when it mixes into water bodies. The AP was identified as fast spikes when measured between two electrodes and was reportedly propagating with a velocity of 1 m/sec

after 24 hours of applying the chemical. It reaches 2 m/sec during the second day after the application.

In Figure 2.9, note the duration of the AP varies from around 3 s (A) to 20 ms (B, C), after 24 hours of application. The amplitude was around 60 mV. A very similar result has been reported, where 3 week old soybean seedlings (having 5-6 well-developed leaves) were treated with a solution of *Carbonylcyanide 4-trifluoromethoxyphenylhydrazone* (FCCP) added to the soil [101]. FCCP is an uncoupler (agents which lower the degree of coupling between ions) which separates the flow of electrons and the pumping of H^+ ions for ATP synthesis. Thus, FCCP eliminates the usage of the electron transfer energy for ATP synthesis. A treatment of the soil by FCCP triggered AP in the soybean plants. These APs had an amplitude of 60 mV and maximum propagation speed of 10 m/s within the first 20 hours of application of the chemical. The duration varied from 2 s to 0.002 s. After 100 hours of application of FCCP, the duration was around 0.0003 s and propagation speed was 40 m/s [101].

Adding an aqueous solution of another insecticide *pentachlorophenol* (PCP) to the soil also induced APs in soybean plants [102]. The amplitude of the AP was 60 mV, with the VP reducing from 75/80 mV to 0 mV within 48 hours of recording. Similarly, use of *Carbonyl cyanide 3-chlorophenylhydrazone* (CCCP) on soybean plants to study the generation of AP and VP, measured the maximum amplitude of AP, duration and speed of propagation [103]. Using chloroform and electrical impulses, the electrophysiological responses were studied in *Goeppertia bachemiana* and *Donax canniformis* plants for their style movements [104].

Salt stress in plants, transmitted to various parts of the plant body via propagating calcium signalling waves [105], have not yet been studied for the nature of the resulting AP/VP. This is a research gap which is addressed in this work by conducting experiments with a salt solution and measuring the plant's electrical response.

2.4.3 Heat and Cold shock

Experimental study on *Aloe vera* plants with thermal shock were carried out [106], where the AP, on being induced due to cold (Figure 2.10) and heat (Figure 2.11), was travelling with a propagation speed ranging between 67 and 132 m/s. These signals travelled along all the leaves [106]. Figure 2.11 shows that *Channel 1* recorded the electrical measurement from the leaf where heat stress was applied. *Channel 2* recorded the response from another leaf where

no heat stress was applied. The AP propagated with constant speed and amplitude along all the leaves of the plant. Heat stress in plants produces *heat stress proteins* (*hsp*) which help the

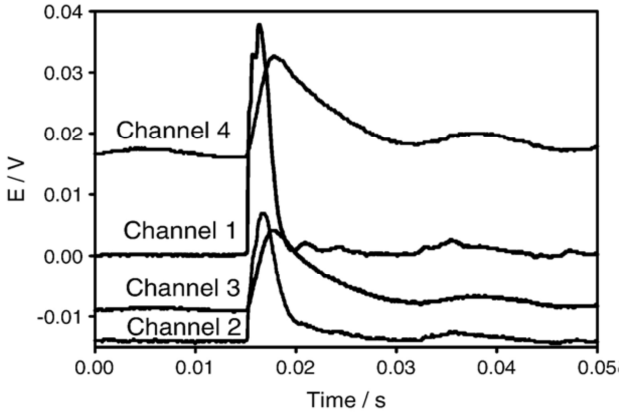


Figure 2.10: Electrical response of Aloe Vera plant due to cold shock [106]

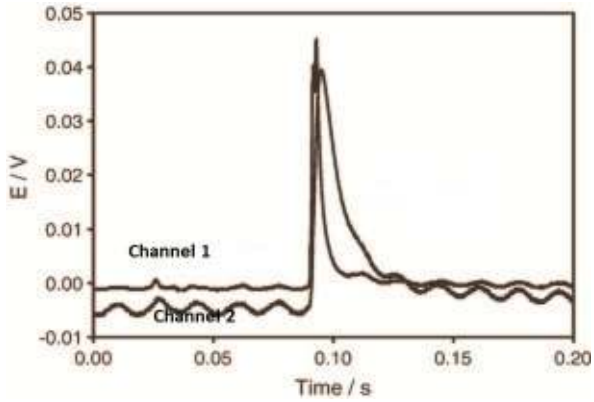


Figure 2.11: Electrical response of Aloe vera plant due to flaming [106]

plants tolerate extreme temperatures. Ca^{2+} has also been suggested as playing a key role in producing a response in the plants when subjected to elevated temperatures/heat shock.

The importance of exploring the reaction of plants to thermal shock lies in the fact that every year, there are huge losses in vegetation, and hence economy, due to forest fires. If such plant electrical responses can be extracted and studied in the field (i.e. non-laboratory conditions), an adequate sensing mechanism can be developed which will help authorities monitor and control such forest fires, thereby controlling the losses to precious green cover, animals that live in such vegetation, and money.

2.4.4 Excision (Pruning and Tipping)

Electrical potentials were monitored in five avocado plants with 7-9 cm trunk diameters, 3-5 branches, and with 50-70 leaves. Each tree was kept in a 25 litre inert sandy substrate filled container [107]. Micro-electrodes were inserted into the trunk of each tree using a microdriller to a depth of 0.5-0.75 cm into the xylematic tissue. The reference electrode was inserted into the inert sand medium. Seven such micro-electrodes were inserted into the trunk, below the distal apex of

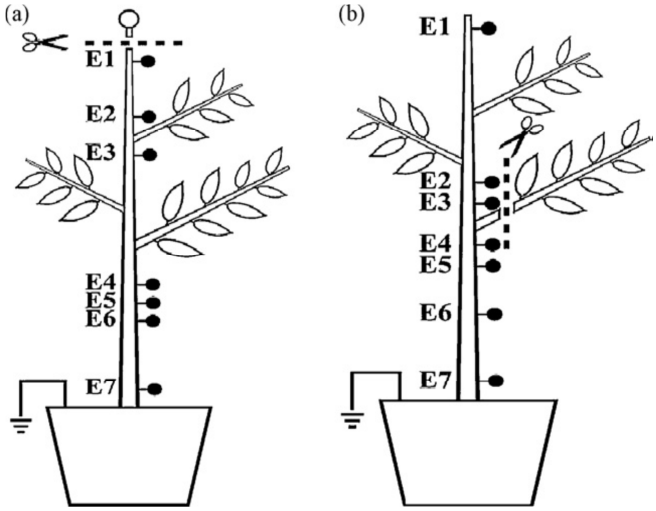


Figure 2.12: Mechanical wounding in Avocado plants – (a) tipping (b) pruning [107]

the tree, which was the tipping site, 2 days before the experiment, Figure 2.12 (a). Again, 2 days before the pruning experiment, the branch closest to the soil was excised. Microelectrodes were placed above and below the pruning site, Figure 2.12 (b), to record the propagation of the response signal of the plant. A waiting period of 60-70 min was allowed for the plant to recover from the initial electrode insertion.

Tipping was carried out 79 s after starting the recording, with a sampling rate of 1 s for a total duration of 200 s. Electrical responses were detected, in the form of VP propagating from the stimulus point, down the trunk to the root. The average linear velocity was 8.7 cm/s. There was a time lag in the response which increased with distance between the stimulation site and the electrodes.

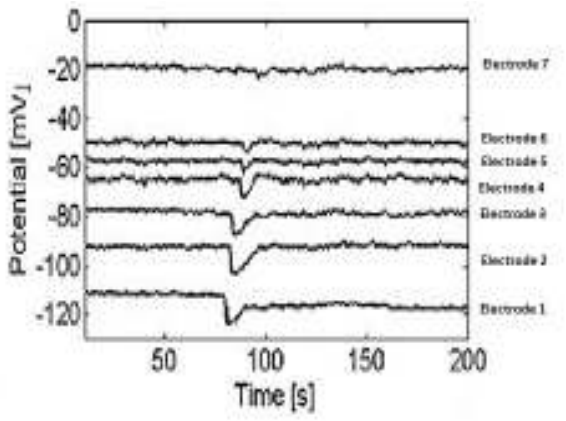


Figure 2.13: Electric potential in Avocado tree(s) due to tipping [107]

It was noted that, with increasing distance from the tipping point, there was a time lag for the signal to be recorded by the electrodes thereby confirming it to be a genuine response from the plant due to the mechanical stimulus. Even the intensity was reported to reduce with increasing distance [107].

Pruning was also carried out 79 s after starting the recording with a sampling rate of 1 s up to a total duration of 200 s. Electrical signals, suggested to be VP, were observed to travel at a linear average velocity of ~20 cm/s towards both upper and lower parts of the stem. Again a linear time lag was noted between the electrodes, thereby suggesting a genuine electrical response of the plant to a wounding mechanical stimulus [107].

It has been suggested that active proton-ATPase pumps (primary transport mechanisms) are not only the source of electrical signal generation within the plant as a result of the tipping/pruning stimulus, but they also inhibit channel opening (passive transport) and are a source of ion carriers (secondary active transport). As the opening and closing of ion channels enable exchange of ions between cytoplasm of the cell and the extracellular environment, a potential difference is created between the inside and outside of the cell [107].

Other recent works on plant electrophysiology include monitoring electrical signals from multiple *Cucumis sativus* (cucumber) plants in a greenhouse as a response to different environmental parameters such as temperature, CO_2, humidity, and light intensity [108], and the detection of water stress in fruit-bearing woody plants by the analysis of electrical signals from *Prunus domestica* (plum) and *Persea americana* (avocado) [109].

2.5 Signal processing techniques

Compared with the work carried out for recording plant electrical signals, the application of signal processing techniques for extracting hidden information within those signals has been limited. The main application of signal processing in understanding the time and frequency characteristics of the recorded signal was carried out in [110]. This presented a comparative study of signals obtained from *Aloe vera* and *Scindpsus Aureus* (Pothos) using all three domains of signal processing method (time, frequency and time-frequency). The time domain plot showed that electrical signals from *Aloe vera* had a high fluctuation in amplitude, with a maximum of 2 mV. This was greater than the maximum amplitude of signals extracted from *Scindpsus Aureus*, for which the value was 190 µV. These values are shown in Figure 2.14 and Figure 2.15. Because the amplitude fluctuation is higher in *Aloe vera*, the standard deviation of the signal is greater compared with *Scindapsus Aureus*. Apart from the maximum amplitudes and standard deviations, other simple statistical parameters such as minimum value, mean, etc. were also computed to see the difference between the electrical signals.

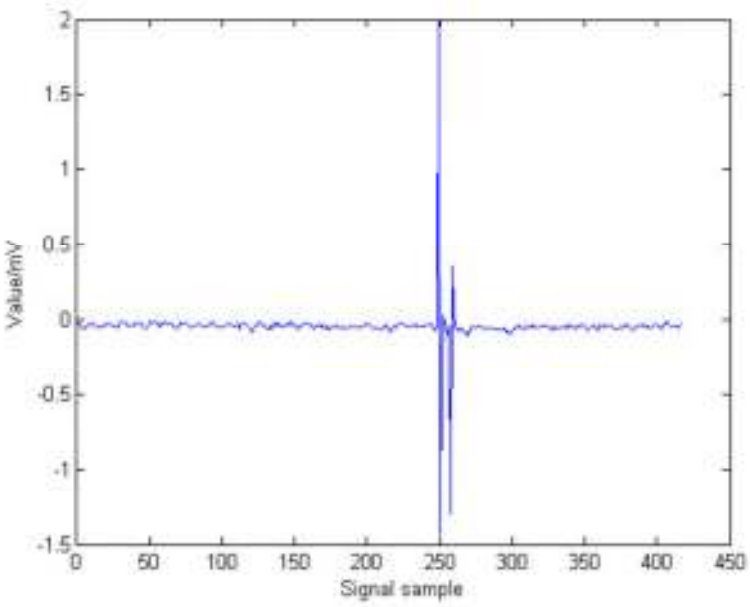


Figure 2.14: Electrical Signal from *Aloe Vera* [110]

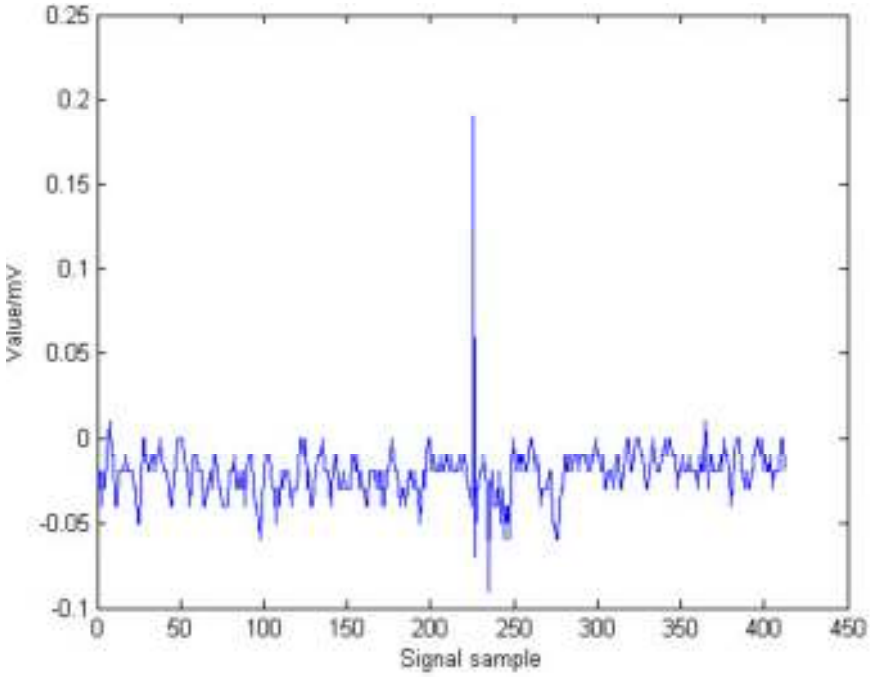


Figure 2.15: Electrical signal from *Scindpsus Aureus* [110]

The time domain will show how the signal varies its amplitude over a period of time, whereas the frequency domain will show the frequency components of the signal.

The power spectrum estimation (energy distribution of a time series in the frequency domain) of the electrical signals obtained from both plants was also reported [110]. The *Fourier transform* was not used for analysis since the plant signals were random (non-stationary).

Power spectrum estimation, involved the following two steps (*Wiener-Khinchin* theorem):

- Estimating the autocorrelation function
- Applying a Fourier transform to the autocorrelation function

The power spectrum estimation from the two plants are shown in Figure 2.16 and Figure 2.17, where the y-axis represents the power and x-axis represents the frequency in Hz.

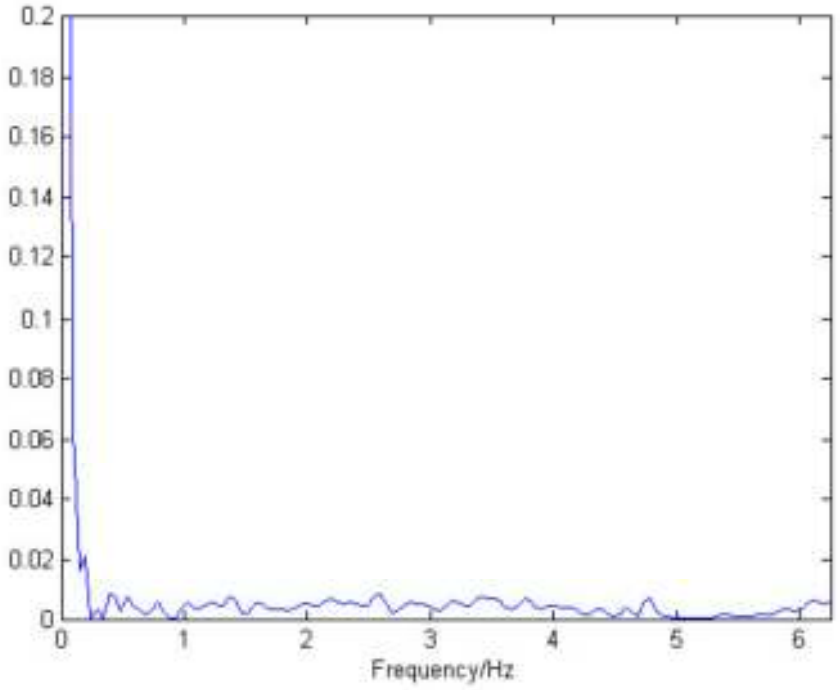


Figure 2.16: Power spectrum of electrical signal from *Aloe vera* [110]

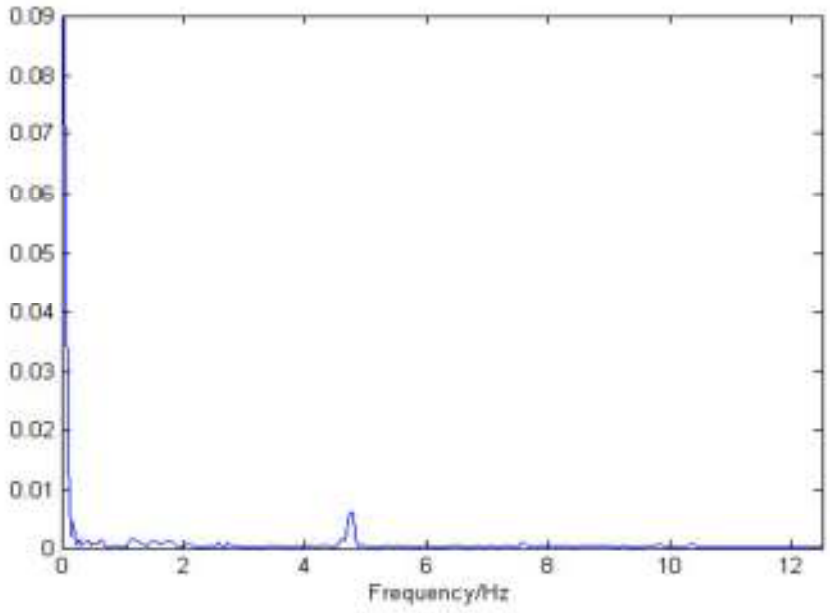


Figure 2.17: Power spectrum of electrical signal from *Scindpsus aureus* [110]

Since a time domain analysis presents accurate information about time (*amplitude* vs. *time*) and frequency domain analysis presents accurate information about frequency components (*signal power* vs. *frequency*), neither can be used to glean any information about frequency components present at a particular time. To obtain such information, a time-frequency

domain methodology is required. This methodology is especially applied to non-stationary signals (whose statistical parameters change over periods of time).

In the time-frequency domain, the analysis involves presenting a one-dimensional time-domain/frequency domain signal in the form of a two-dimensional time-frequency density function. This gives information on different frequency components and the variation of the frequency component with respect to time. *Wavelet analysis*, a time-frequency localization tool, is used to decompose time-series data into a time and frequency domain simultaneously. This provides both the amplitude of the periodic signal at any given time and the variation of the frequency over time. Such analysis has been applied to many fields including EEG and ECG.

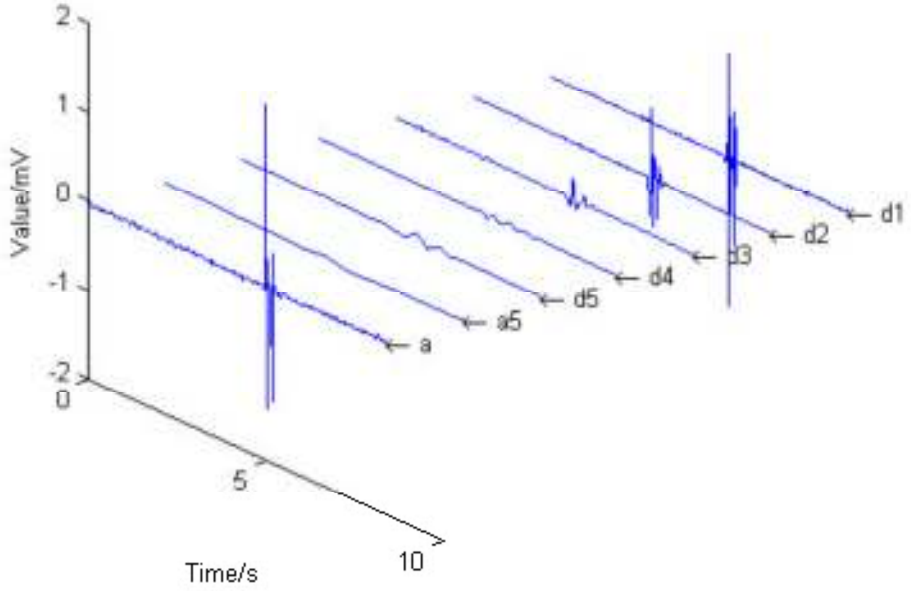


Figure 2.18: Five layer wavelet decomposing of electrical signals from *Aloe vera* [110]

The wavelet analysis, used Daubechies (db3) wavelet with a decomposition scale up to 5 [110]. The results obtained are shown in Figure 2.18 and Figure 2.19. In these figures, *a* represents the original signal, *a5* represents the low frequency components while (*d1*) to (*d5*) represents the high frequency components. Since electrical signals from plants are unstable, time-varying, and contain different frequency components at different time intervals, a time-domain analysis provides better information about the signal characteristics [110].

The *lifting wavelet* was used as a pre-processing step for plant electrical signals obtained after using *heat shock* (burn), *sulphuric acid*, and *irradiation*, on *Aloe vera* plants to study their

transmission speed [111]. Correlation analysis produced the *propagation speed* of the electrical signals for the three different stimuli.

Wavelet analysis was used on electrical signals from six different species of plant in order to study their time-frequency characteristics and identify the occurrence of AP [112]. The *db3* wavelet was used as the basis function and decomposed the electrical signals obtained from six plants into 5 levels. The observed frequency of APs generated in the six different plants and their amplitudes are shown in Table 2.1.

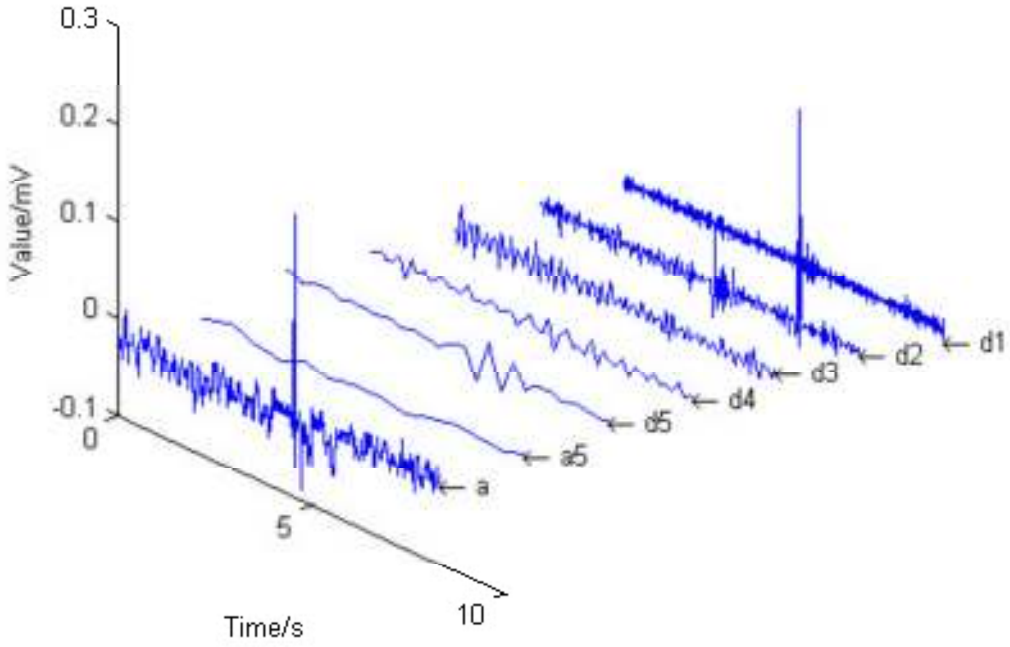


Figure 2.19: Five layer wavelet decomposing of electrical signals from *Scindpsus aureus* [110]

Table 2.1: Wavelet Analysis of electrical signals from plants [112]

Plant species	**Frequency of total observed APs (No. of APs in Time)**	**Time taken for the second AP to appear**	**Amplitude value**
Crassula portulacea	5 times in 180 sec	60 sec	80 µV
Jasminum sambac	2 times in 100 sec	65 sec	190 µV
Aloe vera (chinensis)	2 times in 500 sec	260 sec	310 µV
Scindpsus aureus	2 times in 400 sec	100 sec	550 µV
Catharanthus roseus	3 times in 500 sec	60 sec	590 µV
Celosia cristata	2 times in 100 sec	40 sec	200 µV

Wavelet packet decomposition and employed a BP neural network classifier have been used to classify seven different types of plant species by using their electrical signal responses

[113]. Liu et al. [114] studied water stress related *acoustic* signals emitted by *Broussonetia papyrifera* (paper mulberry) and *Populous* (poplar), and evaluated suitable de-noising techniques using wavelet decomposition and thresholding. Although acoustic signals are not related to plant electrophysiology, it was necessary to review the processing needed to evaluate whether such techniques could be carried out on plant electrical signals.

Electrical signals from a jasmine tree were analysed and de-noised using wavelet soft thresholding method [115]. A Gaussian radial basis function based neural network was employed for forecasting future values of the electrical signals, based on adaptive characteristics of the plants. It was reported that such forecasting could be used for intelligent control systems, with the possibility of using it in agriculture.

2.5.1 Blind Source Separation using Independent Component Analysis

Blind Source Separation (BSS) is the process of separating the source signals from a set of mixed signals, without much prior information on how the mixing occurred or about the sources of the signals themselves. One of the various methods by which a BSS can be achieved is the *Independent Component Analysis* (ICA) technique.

Assuming that the source signals are statistically independent (occurrence of one does not affect the probability of occurrence of the other) and are non-Gaussian (i.e. not normally distributed), ICA can be used to separate a multivariate signal (observation of more than one output variable) into its additive sub-components.

Huang et al. [116] used ICA to separate mixed signals coming from the electrical signals of the *epidermis cells*, *guard cells* and *mesophyll cells*. They tested the method on simulated signals and verified with actual surface-recorded electrical signals from plants, using light/darkness periods as stimuli. The original signal is shown in Figure 2.20, the normalised signal shown in Figure 2.21, and component signals in Figure 2.22. Since individual signals from three different cells were to be detected by using ICA, data was collected from three different sensors fitted to the plant.

By measuring a combined electrical signal using a surface recording system from the leaf of bean plants, individual cell responses to stimulus could be determined by BSS using ICA [116].

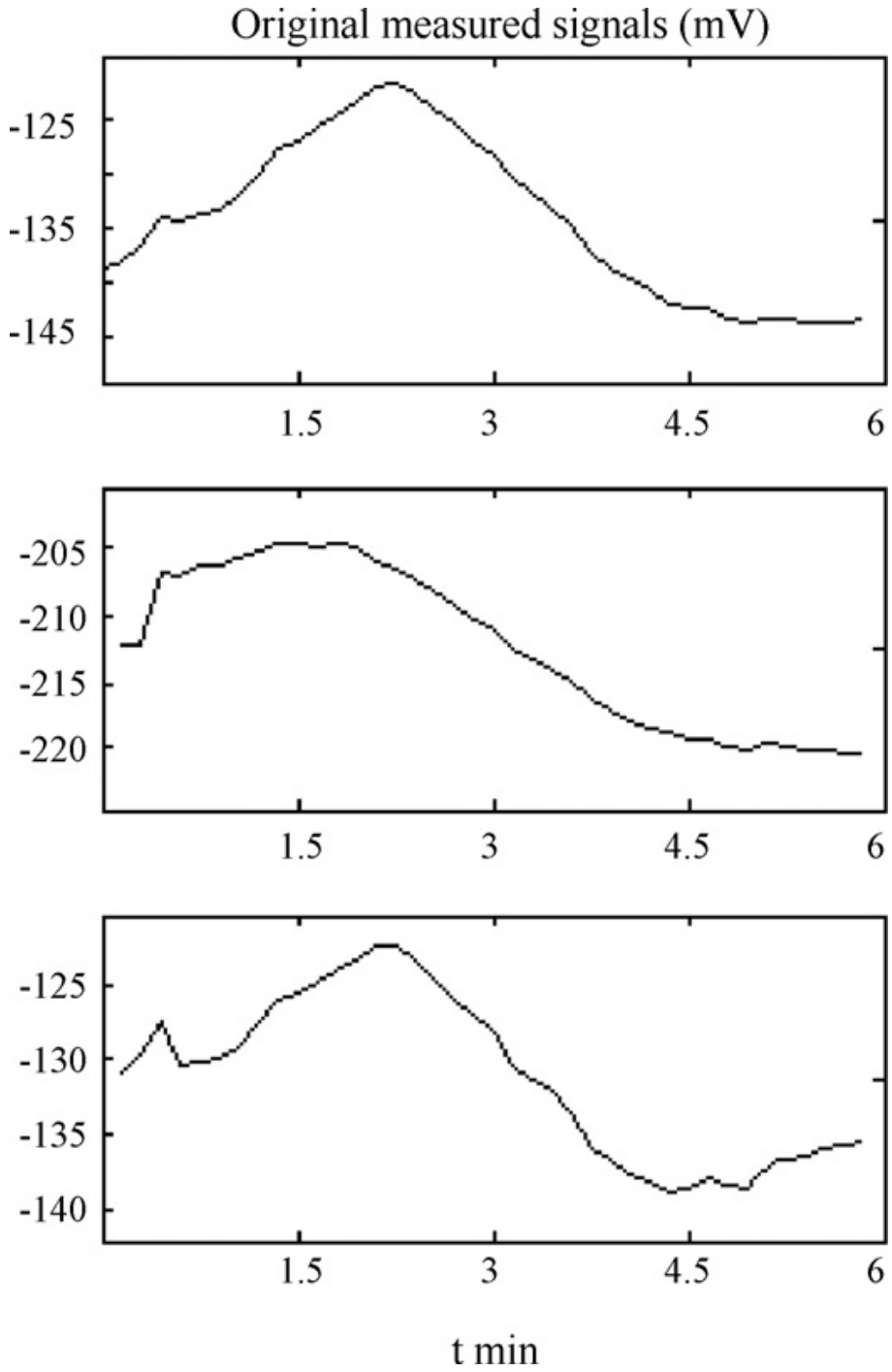


Figure 2.20: Light/Darkness induced electrical signals from three Bean plants [116]

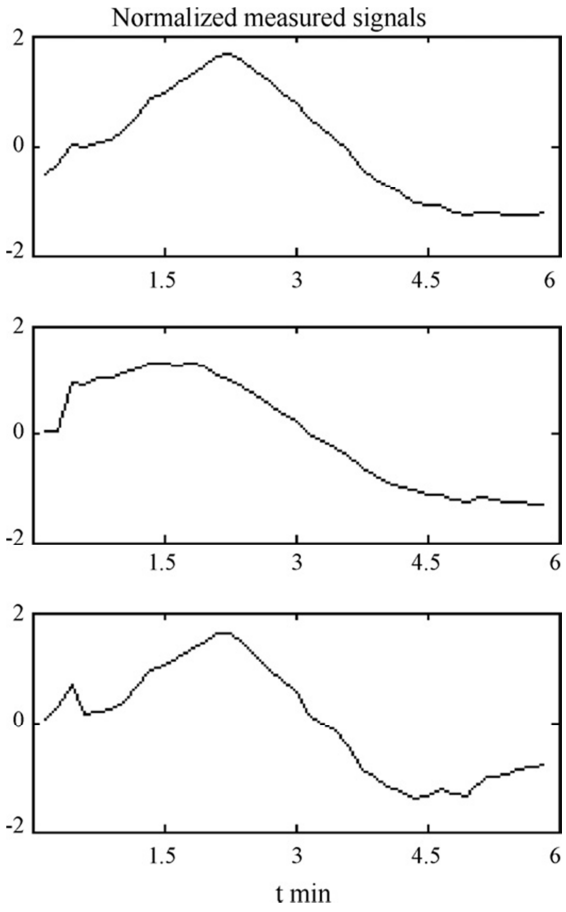


Figure 2.21: Normalized plant electrical signals [116]

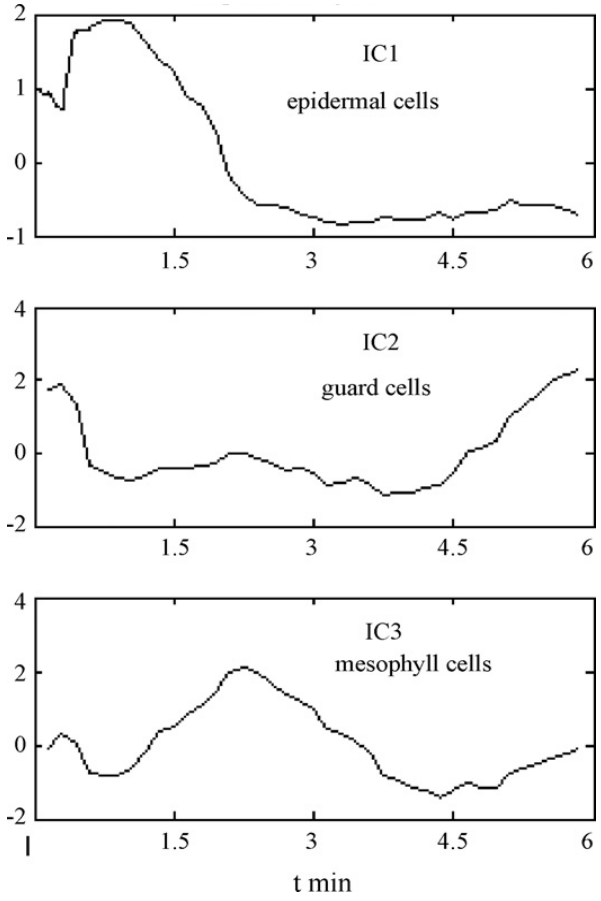


Figure 2.22: Separated signals using ICA [116]

2.6 Summary

This chapter reviewed the electrical signal responses generated by plants to various stimuli, their extraction methods, and what types of signal processing have been employed to process extracted electrical signals. Although some work on processing plant electrical signals has been published, none has attempted to map the electrical signal response of the plants to the external stimuli affecting them.

In this work, an attempt is made to use signal processing and machine learning techniques to identify external stimuli affecting the plants, from the resulting plant electrical signal responses.

3 Forward and Inverse Modelling for Predicting Light Stimulus from Electrophysiological Response in Plants

3.1 Introduction

To find out whether there is sufficient information about any stimulus in the electrical response of a plant, the possibility of detecting light pulses as a basic stimulus was explored. It has been shown that as light pulses are applied on plants, an electrical signal with a trend are generated. The question is whether this electrical signal response can be used to find the following characteristic parameters of the light stimulus:

- switching on (t_{on}) time
- switching off (t_{off}) time, and
- amplitude of the applied light pulse

To detect these parameters of the light pulse, plants were considered as a black-box system with a train of light pulses acting as input and the corresponding electrical response as output. The aim of using the black-box modelling approach, called *system identification*, was to explore a series of linear and non-linear estimators to mathematically extract the properties of the light pulse stimulus from the electrical signal response of the plants.

System identification methodology was adopted to develop a novel *dynamical model* for describing the relationship between light as an environmental stimulus and the electrical response for a *Laurus nobilis* (Bay leaf) plant, which acted as the main dataset for the modelling. Various linear and non-linear models were to be compared using suitable estimators for the main dataset, and then use the parameter settings obtained for the best models (in terms of percentage fit) on 19 more datasets (i.e. 19 different plants, exposed to similar light pulses). This was carried out to see whether the parameters used to find the best models (linear and non-linear) with one dataset produced similar results with other datasets.

Two major classes of system estimator were explored to develop the dynamical models – linear and nonlinear – and their several variants, for establishing a forward and an inverse relationship between the light stimulus and plant's electrical response.

The purpose was to predict the parameters of the input light stimulus (on-off timing and amplitude) from the measured electrical response, by finding a solution to the inverse problem – i.e. where the electrical signal response acts as the input to the black-box system and the light pulse acts as the output.

The best class of models were given by the Nonlinear Hammerstein-Wiener (NLHW) estimator showing reasonable data fitting results exceeding other linear and nonlinear estimators. Consequently, a set of models using variants of NLHW were developed and their accuracy in detecting the on-off timing and intensity of the input light stimulus were compared for multiple plants under a similar experimental scenario.

3.2 Review of light induced plant electrical response models

When plants are grown outdoors, sunlight provides the required energy through photosynthesis. The plants also use the intensity of the available light for sensing and responding to their surrounding environment [117]. Growing plants in sheltered conditions such as a greenhouse requires the use of controlled artificial lighting. Controlled-environment agriculture (CEA) technologies, which includes greenhouse, hydroponics, aquacultures along with vertical farming, provide alternative solutions for crop production. These solutions are especially important in locations with limited daylight or adverse environmental conditions such as drought, flood, storm, and soils with high salt content. They are also useful in cities where there is limited space for growing crops [117].

Indoor farming has to rely on artificial light sources which provide the necessary conditions for plant growth without consuming too much energy. Light-emitting diode (LED) technologies present great potential for providing adequate conditions for optimal plant growth indoors. Among available artificial lighting sources such as metal halide or sodium lamps, LEDs present the maximum Photo-synthetically Activated Radiations (PAR) efficiency of 80-100%.

LEDs, which can emit red, blue, yellow, orange, green and far red, can be combined to provide high radiant *fluence* (radiant energy received per unit area of a surface), or can be used for particular wavelength characteristics due to the LED's narrow-bandwidth light spectrum [117].

To determine the light characteristics required by plants for optimum growth, one study grew lettuce under red LEDs [118]. Similarly in [119], experiments were conducted on lettuce under red light (~670 nm) provided by LEDs and high pressure sodium lamps and reported that the amounts of dry matter gained by lettuce grown using the two different lamp sources was identical [119]. Chang et al. [120] reported that maximum photon utilization efficiency for growth in green alga *Chlamydomonas reinhardtii* were observed under red light provided by LEDs (~674 nm).

Lettuce which were grown under red LEDs were found to have elongated hypocotyls and cotyledons. This is attributed to the efficiency of red light (650-665 nm) which has a significant influence on plant growth as the range of red light wavelengths are exactly in tune with the absorption peak of chlorophylls [121] and *phytochromes*. Higher photosynthetic activity in plants can be triggered with a combination of blue and red light than that obtained using either red or blue light separately.

This effect has been attributed by some to a high content of nitrogen in blue-light-supplemented plants. However, others have attributed this effect to better stomatal opening, thereby making more CO_2 available for photosynthesis. It has been established that stomatal opening is influenced by blue-light photoreceptors [122], and may be directly proportional to an increase in shoot dry matter [123].

If a mathematical relationship can be established between light and growth with an appropriate model, then such a model can be used for precision agriculture where the light requirement by the plant can be understood and best provided for it. This will both save energy (for providing the light) and make it optimal. Although, in this work only white light pulses have so far been chosen as stimulus, in future coloured lights can be addressed for conducting experiments on electrical signal responses.

The mechanisms of light-stimulus induced electrical signal generation in plants, in terms of ionic conduction, are quite well researched. Roelfsema et al. [90] found three physiological states of the guard cells, which are electrically isolated from other plant cells. These states were found to be far-depolarized, depolarized and hyperpolarized. The depolarized guard cells were found to extrude potassium (K^+) ions through the outward rectifying channels, whereas the hyperpolarized cells let in K^+ ions through inward rectifying channels. The guard

cells were reportedly switching from depolarized to hyperpolarized state upon incidence of light and vice versa for a light to darkness transition [90].

In principle, modelling input-output relations of any dynamical system can be done in two ways. The first method involves a mechanistic approach where detailed understanding of the characteristics of physical interactions between the system's components with the input stimuli are exploited, through known laws or equations. The second method considers the system as a black-box and statistically formulates the functional relationship by observing the output responses to input stimuli. While the first approach is desirable as it gives a complete understanding of the internal operation of the system, the second approach, known as *system identification*, is more suitable for developing a working solution when knowledge about the interactions between system components is not complete. To model the electrical response of plants under external stimuli, from the level of existing knowledge, the system identification approach appeared to be the most appropriate.

In order to explain the plant electrical response due to light stimulus, several *mechanistic* models have been proposed by plant scientists. Models proposed in [124]–[127] and [128] explain the generation of AP and VP, while several models describing the underlying generation process of AP and VP are reviewed in [81]. These models essentially describe the generation of different ionic currents across a cell and how these currents lead to a transient depolarization of the membrane potential. The interactions of these currents with the transmembrane voltage is described using either a linear or an ohmic or a nonlinear relationship such as the Goldman-Hodgkin-Katz equation [129]. The stimulus occurs in these models through stimulus-induced calcium current that triggers further ionic currents. None of these models quantify the time course of the stimulus intensity. Thus an attempt to construct a mechanistic model is futile, as the relationship between the electrical responses recorded from the surfaces of plants and that from the cell is not well understood. Furthermore, a surface recording of the stimulus driven electrical response may show traces of AP or VP or a combination of both [130]. Hence it was decided to use a black-box modelling approach to model the relationship between light as stimuli and the corresponding electrical signal response of the plants.

3.3 Modelling approach adopted

Rather than an *a priori* assumption of the relationship between the stimulus and the response arising from the field of plant physiology, here the relationship was inferred from the data (time course of response and stimulus traces) by using black-box modelling. An advantage of this approach over mechanistic modelling is that the same data could be used to construct an inverse relationship between the input and output, which could provide inference of the stimulus from the observed electrical signal response.

The input-output relationship of the plants to be modelled was chosen to be a dynamical rather than a static model, since a dynamical model takes into account the real-time changes of the internal state variables of a system, expressible using one set of ordinary differential equations involving time derivatives. Thus a dynamical model is considered more behavioural in nature than a static algebraic equation-based model. Since the external light stimulus is responsible for the generation of the electrical response, from the point of view of plant physiology, the problem addressed here was essentially the inverse problem where the characteristics of the light stimulus needed to be determined by observing the electrical response signal. Derivation of such an inverse model was challenging since there may exist a number of solutions for the same input-output behaviour.

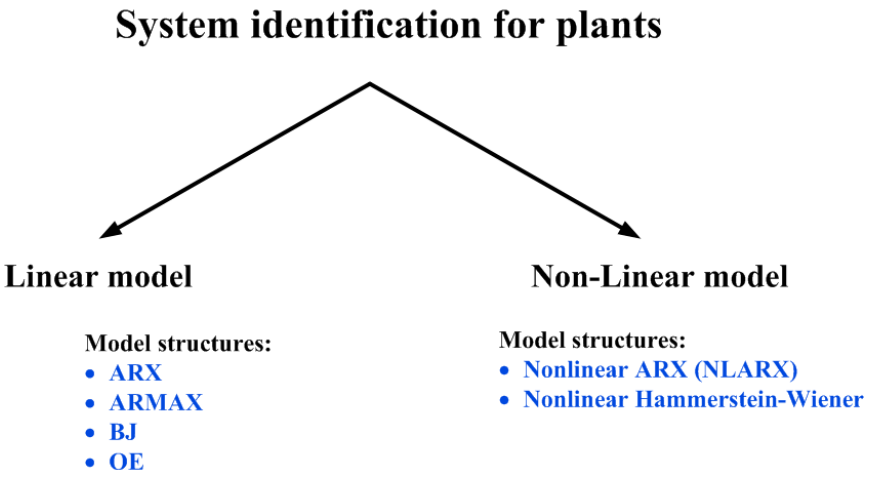


Figure 3.1: Different estimators for black-box modelling

To address these issues, two different classes of system identification techniques were applied: Linear Time Invariant (LTI) discrete time estimators, and nonlinear discrete time estimators, as shown in Figure 3.1. Within each class, the most accurate model structure was chosen, on the basis of obtained accuracies (% fit), for the available input-output data. Four

different linear Least Square Estimator (LSE) based techniques were used within the LTI framework: Auto-Regressive eXogeneous (ARX), Auto-Regressive Moving Average eXogeneous (ARMAX), Box-Jenkins (BJ) and Output-Error (OE), whereas Non-Linear ARX (NLARX) and Non-Linear Hammerstein Wiener (NLHW) [131]–[133] estimators were considered within the nonlinear estimation framework.

Figure 3.2 shows that forward modelling considers the input light stimulus as the input to and the plant electrical response as the output from, the black-box, thereby capturing the physical cause and effect relationship. The inverse model tries to establish a causal relationship as a form of a dynamical model using plant electrical response as input to and light stimulus as the output from the black-box. This gave a working solution for detecting the duration and nature of light inputs by only observing the electrical response.

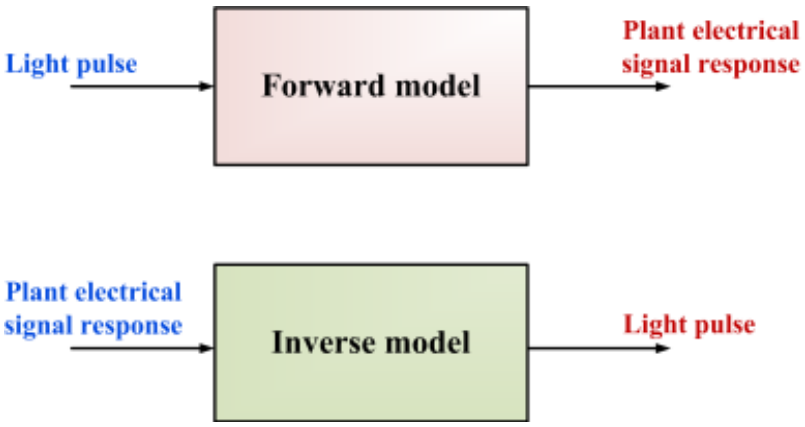


Figure 3.2: Forward and Inverse models using light pulse and plant electrical signal response

Based on percentage of fit, the top three model structures were selected, obtained from exploration with light-induced electrical signal from the *Bay leaf* plant (treated as the main dataset). Thereafter, these three estimator structures were used on data generated by an additional 19 different plants (17 *Zamioculcas zamiifolia* and 2 *Cucumis sativus*) under similar experimental conditions for the forward and inverse modelling, in order to see if the top model structures still gave faithful estimate of the light pulse as stimulus in a generic form.

The chosen models are standard templates given in the System Identification literature [131]–[133]. These models are used in a variety of time-series analyses [131], [134]. The objective here was not to develop a new modelling tool, but to explore the best models which fit the plant electrical signal response data. Three standard templates, Linear and two classes of

Non-linear, were explored for this work. In time series analysis, when the number of parameters of an estimator becomes comparable to the number of data-points, then it becomes an ill-posed problem (large P, small N problem [135]). However, this has not been the issue here because each time series has more than 700,000 points from which the hyper-parameters were tuned. In another sense, for selecting the best models, a low number of datasets were used according to availability. Since this first exploration was to see if the chosen models (out of a pool of already established models and estimators) produced any repeatable results or not, a series of steps were followed, which are described in the following figure.

Figure 3.3: Planned approach to finding the top three models

As can be seen from Figure 3.3, the process started with the 20 Datasets initially available from experiments conducted by co-researchers in Florence and Rome. It was found that the models found from one dataset did not produce much consistency of results for other datasets, which was expected because the morphology of the datasets were not consistent. System identification evaluation will be more robust when the data is repeatable and consistent. However, the data from the initial experiments were found not to be consistent.

The causes of this could be: any variations in the setup for extracting the data from the plants, the ability of the plants chosen to respond to light as a stimulus, ambient conditions, etc. The ambient conditions were kept the same as much as possible, and any variations in temperature and humidity were within tolerance (a minimum of ±10° C is required to initiate a response from the plants [125]).

The rest of the chapter is organized as follows. Section 3.4 provides a theoretical overview of system identification techniques, describing the mathematical preliminaries of various estimators used. Section 3.5 discusses the experiments for obtaining the plant data under light excitation, Section 3.6 discusses the pre-processing steps carried out on the raw plant electrical signals, and Section 3.7 discusses the results of the proposed model structures. Section 3.8 details new experiments which were conducted to gather more data, their results and analysis.

3.4 Theoretical background of forward/inverse dynamical system modelling

System modelling or identification can be viewed as a way of mathematically describing a phenomenon with some physical insight about the system from a measured input and output dataset. It is regarded as a bridge between the application to real problems and the mathematical theories of model abstraction [131]–[133]. The application end of the spectrum of system modelling includes prediction of the input to output characteristics of a model. A forward system model, where the response (output) needs to be related to the cause (inputs), is much easier to develop in theory than developing an inverse model where the inputs are predicted from the observation of output response, as this may result in a one-to-many mapping. Such a forward model will capture the dynamical characteristics of the response caused (due to the excitation) by establishing a physical cause-effect relationship. A similar approach by inverting the input and output may not always indicate physical causation [136] but is capable of predicting the input (the applied light stimulus) by only observing the output (the electrical signal response of the plants), and hence was adopted here.

In principle, once formulated, a forward model could also be inverted to produce an inverse model. However, from the perspective of dynamical system theory, establishing a forward dynamical model (in terms of a transfer function) first and then inverting the poles into zeros (and zeros into poles) to form an inverse model, may not always work. This is due to the

possibility of having zeros in the right half of the s-plane (i.e. non-minimum phase zeros in the forward model), which get converted to poles in the right half s-plane of the inverse model, thereby rendering the system unstable.

Therefore, a different approach was needed for solving the inverse modelling problem. Figure 3.2 shows how this was achieved, by inverting the cause (i.e. light pulse) and effect (i.e. electrical signal response), followed by varying the model structure between these two observed signals in order to match the cause in the best way.

To solve the inverse problem of predicting the characteristics, such as switching on and off time and the amplitude of the light pulse stimulus, from the plant electrical response, two classes of dynamical system parameter estimation technique were adopted: linear and nonlinear methods. The following subsections briefly describe the theoretical backgrounds of each of these system identification techniques after describing the main statistical measures used for measuring the accuracy of model fitting.

The *System Identification Toolbox* of MATLAB [134] was used to develop the input-output linear and nonlinear forward/inverse models.

3.4.1 Least squares estimation for system identification

One of the techniques employed to estimate the parameters of a model is the *Least Square Estimation* (LSE), which helps by minimizing the difference between actual output and predicted output values. In this way, parameters can be selected which produce the closest to actual recorded value of the output. This is best understood by considering equation (3.1)

$$Error^2 = \left(\left(actual\_output\right) - \left(predicted\_output\right)\right)^2 \tag{3.1}$$

The square of the error is taken to avoid any negative values.

If the measured output and input of an unknown system up to time t are $y_t, y_{t-1} \cdots y_{t-n}$ and $u_t, u_{t-1} \cdots u_{t-m}$ respectively, the system can be described by a linear difference equation with coefficients $a_i, i = 1, \cdots, n$ and $b_j, j = 1, \cdots, m$ as shown in equation (3.2)

$$y_t + a_1 y_{t-1} + \cdots + a_n y_{t-n} = b_1 u_{t-1} + \cdots + b_m u_{t-m} \tag{3.2}$$

or,

$$y_t = b_1 u_{t-1} + \cdots + b_m u_{t-m} - a_1 y_{t-1} - \cdots - a_n y_{t-n} \tag{3.3}$$

The estimated system parameter vector (θ) and the measured input-output vector (φ) can be given as equations (3.4) and (3.5) respectively.

$$\theta = \left[a_1 \ldots a_n \; b_1 \ldots b_m \right]^T \tag{3.4}$$

$$\varphi_t = \left[-y_{t-1} \ldots - y_{t-n} \; u_{t-1} \ldots u_{t-m} \right]^T \tag{3.5}$$

Using (3.4) and (3.5), the system model can be developed, incorporating a modelling error (at time *t*) e_t represented by equation (3.6)

$$y_t = f\left(\varphi_t, \theta\right) + e_t, \quad t = 1,2,3,\ldots,N \tag{3.6}$$

We aim to find the certain parameter vector $\hat{\theta}$ which minimizes the least squared error (S) defined as

$$S \triangleq \sum_{t=1}^{N} e_t^2 = \sum_{t=1}^{N} \left(y_t - f\left(\varphi_t, \theta\right) \right)^2 \tag{3.7}$$

Usually, f is linear in the unknown coefficients θ. Hence, equation (3.6) can be rewritten as equation (3.8) and further as equation (3.9) in matrix notation

$$y_t = \varphi_t^T \theta + e_t, \quad t = 1,2,3,\ldots,N \tag{3.8}$$

$$y_t = U\theta + e_t \tag{3.9}$$

Now the minimum error equation can be rewritten as equation (3.10)

$$S = e^T e = \left(y_t^T - \theta^T U^T \right)\left(y_t - U\theta \right) \tag{3.10}$$

or,

$$S = y_t^T y_t - y_t^T U\theta - y_t \theta^T U^T + \theta^T U^T U\theta \tag{3.11}$$

Thus, we obtain:

$$\frac{\partial S}{\partial \theta} = -2U^T y + 2U^T U\theta \tag{3.12}$$

Hence, the parameter vector $\hat{\theta}$ that makes the gradient of S zero is given by equating (3.12) to zero. This results in equation (3.13).

$$\hat{\theta} = \left[U^T U\right]^{-1} U^T y \tag{3.13}$$

Equation (3.13) describes the estimated system parameter for which the minimum of the sum of squared errors over the time interval *t*, where $1 \le t \le N$, is obtained, and hence this method of estimation is called the *least square estimation* or LSE algorithm. It is worth noting that $\hat{\theta}$ is estimated using the measured input and output data and so it is easy to infer that using LSE, a parameterized system model can be developed. This technique is the backbone of all system identification methods [133].

3.4.2 Four linear system models

The accuracy and the efficiency of the system identification process depends on the choice of suitable model structure. In this section, a few model structures (or estimators) are discussed from the perspective of their applicability to the system [133]. These model structures are derived from the realisation that the system can be described by a linear difference equation. A generalized parameterized linear model can be described by the following equation.

$$y(t) = G(q^{-1},\theta)u(t) + H(q^{-1},\theta)e(t) \tag{3.14}$$

where, $y(t)$ (or y_t) and $u(t)$ (or u_t) are output and input to the system respectively, $e(t)$ is a zero mean white Gaussian noise and θ is the parameter vector to be estimated, $G(q^{-1},\theta)$ is the transfer function of the deterministic part (excitation to response) of the system and $H(q^{-1},\theta)$ is the transfer function of the stochastic part (noise to response) of the system. Here q^{-1} denotes the backward shift operator such that $q^{-1}u(t) = u(t-1)$.

Equation (3.14) can be further modified as equation (3.15) which is also known as the equation error type linear LSE.

$$A(q^{-1})y(t) = \frac{B(q^{-1})}{F(q^{-1})}u(t) + \frac{C(q^{-1})}{D(q^{-1})}e(t) \tag{3.15}$$

where, $\{B, F, C, D\}$ are polynomials representing the numerator and denominator of the system and noise model respectively and $\{A\}$ represents the polynomial containing the common set of poles for both the system and the noise model.

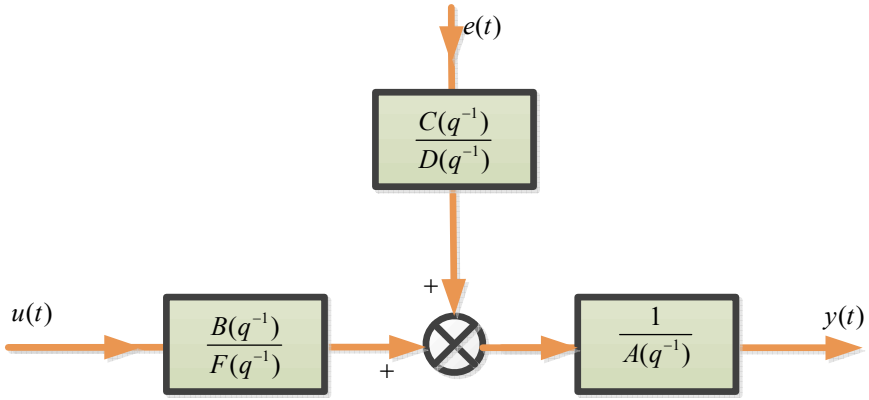


Figure 3.4: Block diagram representation for generalized model structure

Figure 3.4 is the block diagram representation of the generalized model structure in equation (3.15). The generalized LSE can be further customized by considering fewer combinations of the polynomials $\{B, F, C, D\}$ at once, which paves the path towards the choice of a suitable linear estimator for system identification.

3.4.2.1 Auto Regressive eXogenous (ARX) estimator

The basic structure of the ARX estimator is governed by equation (3.16). The *exogenous* term is used for the input signal $u(t)$.

$$A(q^{-1})y(t) = B(q^{-1})u(t) + e(t) \tag{3.16}$$

Expanding equation (3.16), we get

$$\left(y_t + a_1 y_{t-1} + \cdots + a_{na} y_{t-na}\right) = \left(b_1 u_{t-nk} + \cdots b_{nb} u_{t-nb-nk+1}\right) + e_t \tag{3.17}$$

The orders na, nb, nk are the number of poles, zeros and dead time in the system respectively. The *dead time* (or the *delay*) represents the number of samples occurring before the output is affected by an input sample. The main disadvantage of this structure is that the deterministic (system) and the stochastic (noise) dynamics are both modelled with the same set of poles, i.e. the noise is also being processed by the system, which may be unrealistic in many applications.

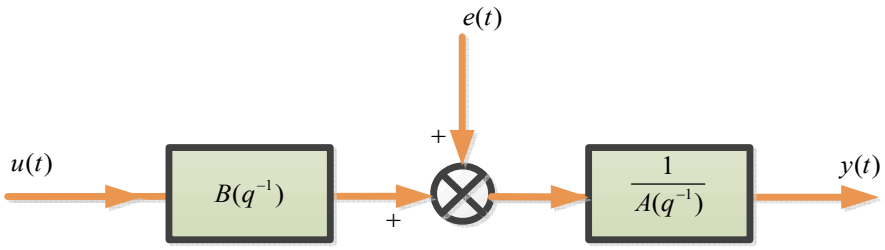


Figure 3.5: Block diagram representation for ARX model structure

3.4.2.2 Auto Regressive Moving Average eXogenous (ARMAX) estimator

The basic structure of the ARMAX estimator is governed by equation (3.18).

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t) \tag{3.18}$$

Expanding (3.18), we get

$$\left(y_t + a_1 y_{t-1} + \cdots + a_{na} y_{t-na}\right) = \left(b_1 u_{t-nk} + \cdots b_{nb} u_{t-nb-nk+1}\right) + \left(c_1 e_{t-1} + \cdots c_{nc} e_{t-nc} + e_t\right) \tag{3.19}$$

The orders na, nb, nk are, just like in ARX structure, the number of poles, zeros and dead time in the system respectively, and nc are the number of coefficients for the error.

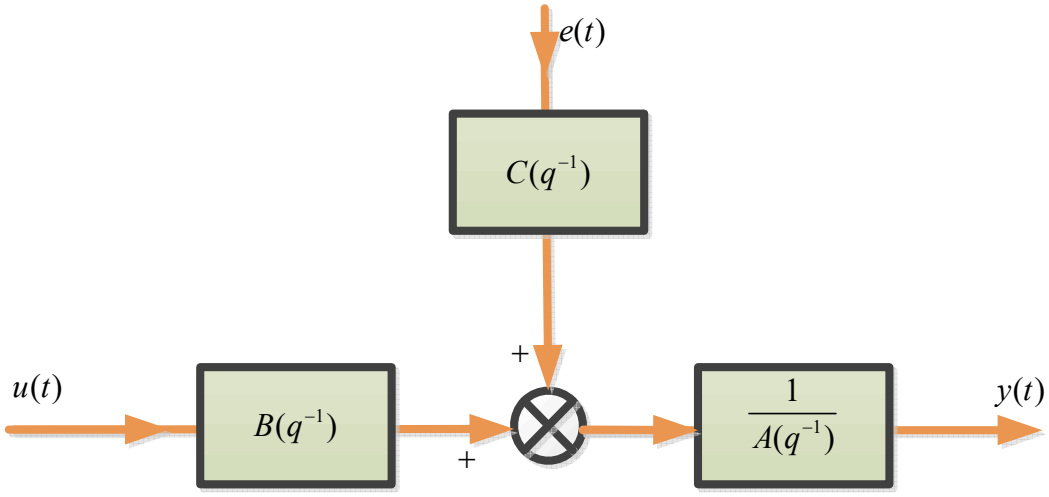


Figure 3.6: Block diagram representation for ARMAX model structure

The major advantage of the ARMAX structure, shown in Figure 3.6, over the ARX structure is that it models the noise dynamics with better flexibility. Although the ARMAX estimates using the same set of poles (given by $A(q^{-1})$), it estimates a different set of zeroes for both the system (given by $B(q^{-1})$) and the noise (given by $C(q^{-1})$) components. Thus ARMAX is useful where the entire system dynamics is dominated by the stochastic component (noise).

3.4.2.3 Box-Jenkins (BJ) estimator

The basic structure of the BJ estimator is governed by the equation (3.20).

$$y(t)=\frac{B(q^{-1})}{F(q^{-1})}u(t)+\frac{C(q^{-1})}{D(q^{-1})}e(t) \tag{3.20}$$

BJ structure allows the estimation of different sets of poles and zeroes for the system and noise component, and are defined by $\{nb, nc, nd, nf\}$ which are the orders of polynomials $\{B, C, D, F\}$ respectively, whereas nk defines the input delay. This structure is especially useful when noise enters the system at a later stage, e.g. measurement noise.

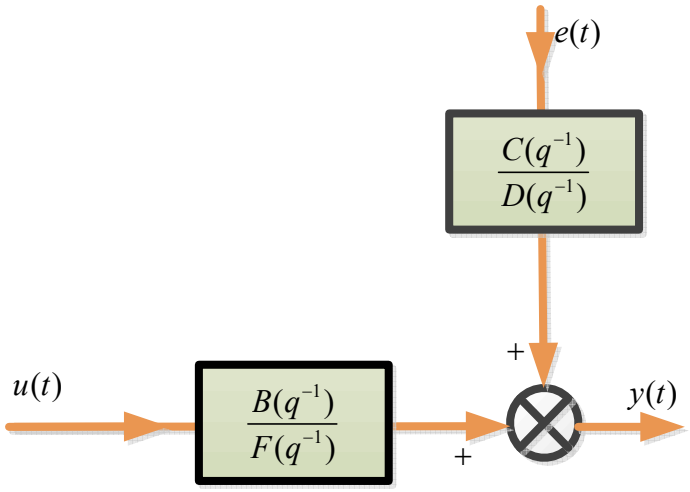


Figure 3.7: Block diagram representation for Box-Jenkins model structure

3.4.2.4 Output-Error (OE) estimator

The OE structure, shown in Figure 3.8, only estimates the poles and the zeroes of the system model, while estimation of the noise model is ignored. The structure is defined by the order of the polynomials B, F represented by nb, nf and the input delay is nk. This structure can be used when the deterministic dynamics dominates the overall system dynamics and the stochastic dynamics has no significant effect.

The estimator has the following structure:

$$y(t)=\frac{B(q^{-1})}{F(q^{-1})}u(t)+e(t) \tag{3.21}$$

For configuring the linear model, only the pole-zero order has been varied, keeping the delay at a fixed value of one. Note that here the order of the poles equals the order of the zeros.

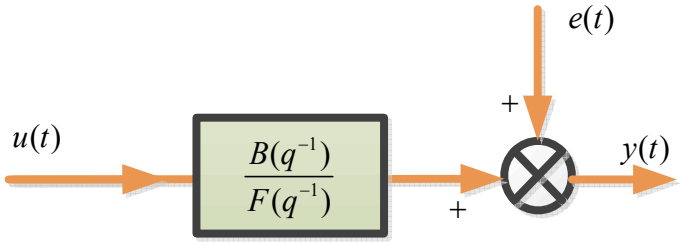


Figure 3.8: Block diagram representation for Output-Error model structure

3.4.3 Two nonlinear system models

Any nonlinear dynamical system whose input and output are $u(t)$ and $y(t)$ respectively, can be expressed as equation (3.22) [134].

$$y(t) = f\left(u(t-1), y(t-1), u(t-2), y(t-2), \cdots\right) \tag{3.22}$$

where, $f(\cdot)$ is a nonlinear function representing any arbitrary nonlinearity. The nonlinear black-box identification is done using two nonlinear model variants and the parameters are found by estimation of the squared error previously discussed. The nonlinear model variants are *Nonlinear-ARX* and *Nonlinear-Hammerstein-Wiener* having the nonlinearity in parallel and series connection with the basic linear blocks, respectively.

3.4.3.1 Nonlinear ARX (NLARX) model

In principle, the NLARX is an extension to the linear ARX estimator.

Equations (3.16) and (3.17) describes the linear ARX structure which implies that the current output is predicted as a weighted sum of past output values and also current and past input values. By rewriting equation (3.17) as a product form, equation (3.23) is obtained.

$$y_t = \left[a_1, a_2, \ldots, a_{na}, b_1, b_2, \ldots, b_{nb}\right] \begin{bmatrix} y_{t-1}, y_{t-2}, \ldots, y_{t-na}, \\ u_t, u_{t-1}, \ldots, u_{t-nb} \end{bmatrix}^T \tag{3.23}$$

where, y_t is the current output and $\{y_{t-1}, \cdots, u_{t-1}, \cdots\}$ are delayed input and output variables called the *regressors*. Instead of a weighted sum as described in equation (3.23), the predicted output y_t can be mapped using a nonlinear mapping function $f(\cdot)$ which gives rise to the NLARX structure. The NLARX is composed of a nonlinear estimator (or simply a nonlinear

mapping function part) and a regressors part, which in turn is the collection of delayed input and output variables. The nonlinear estimator can be further composed of a parallel connection of a linear and a nonlinear function block. Figure 3.9 is the block diagram representation of the NLARX model [134]. The nonlinearity estimator block maps the regressors to the model output using a combination of nonlinear and linear functions.

Various nonlinearity estimators, such as *tree-partition networks*, *wavelet networks*, *sigmoid networks*, *neural networks*, *and custom network* (similar to sigmoid function, where the user can specify the function), can be selected for both input and output blocks. The nonlinearity estimators represent the nonlinear function as a series of nonlinear blocks which can be configured.

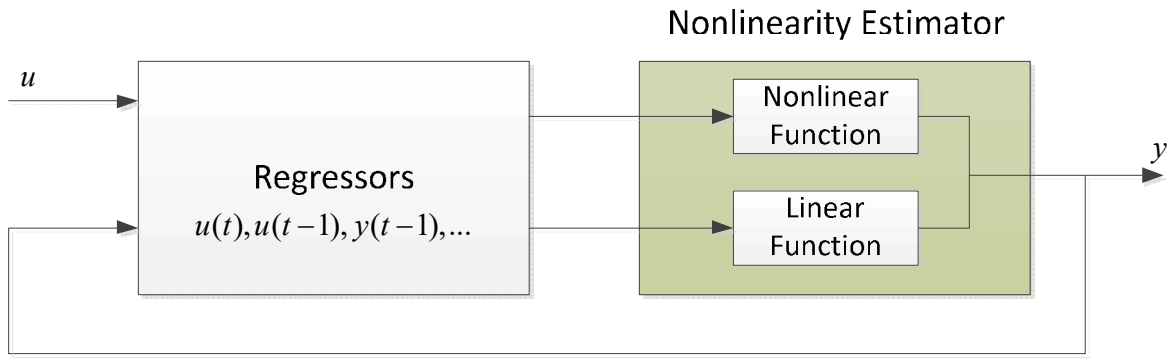


Figure 3.9: Block diagram representation for Nonlinear ARX estimator

The generic output is given by (3.24) and is used to define the nonlinearity block.

$$F(x) = L^T(x-r) + d + f\left(Q(x-r)\right) \tag{3.24}$$

Here, x is an m -dimensional vector of regressors, $L^T(x)+d$ is the output of the linearity block with d being a scalar and $f\left(Q(x-r)\right)$ representing the output of the nonlinear block, r is the mean of the regressors and Q is a $m \times q$ dimensional matrix obtained from principal component analysis of the data.

The following non-linearity estimators for the NLARX model have been explored for our modelling.

3.4.3.1.1 Tree-partition networks nonlinearity estimator

When choosing tree-partition as a non-linear function for estimation, a dyadic partition of the $_X$ -space is carried out. In every partition element P_k , the mapping function F is linear. For any x,

$$F(x) = d + xL + (1, x)C_k \tag{3.25}$$

When tree-partition is chosen as non-linearity, a binary tree nonlinearity estimator with J nodes and 2^{J-1} leaves is initialized. Each node at level $1 < j < J$ has two children and one parent at levels $j+1$ and $j-1$ respectively. The root node at level 1 has two descendants whereas nodes at level J are terminating leaves of the tree and one parent. With each node, one partition element k is associated. The vector of coefficients C_k is computed using the observations on the partition element P_k by the LSE.

3.4.3.1.2 Wavelet network nonlinearity estimator

When using a wavelet network (combination of wavelets and neural networks) nonlinearity estimator, simple wavelet analysis is performed to find the constituent wavelets of the input time series. The structure of the wavelet network nonlinearity estimator is given by equation (3.26), where $\kappa(s)$ is the wavelet function.

$$g(x) = \sum_{k=1}^{n} \alpha_k \kappa\left(\beta_k (x - \gamma_k)\right) \tag{3.26}$$

3.4.3.1.3 Sigmoid network nonlinearity estimator

The sigmoid nonlinearity estimator is described by the structure given in equation (3.27)

$$g(x) = \sum_{k=1}^{n} \alpha_k \kappa\left(\beta_k (x - \gamma_k)\right) \tag{3.27}$$

where $\kappa(s)$ is the sigmoid function and is described as

$$\kappa(s) = \frac{1}{\left(e^s + 1\right)} \tag{3.28}$$

3.4.3.2 Nonlinear Hammerstein-Wiener model

For certain systems where the output varies nonlinearly with its input, the input-output relationship can be broken down to a series of interconnected elements, i.e. the system dynamics is represented by using a linear transfer function and the nonlinearities are captured using nonlinear mappings of the input and output [134]. In the Hammerstein-Wiener model, such an approach has been adopted, where a linear block is connected in series with two static nonlinearities. Figure 3.10 is the block diagram representation of the Hammerstein-Wiener structure [134]. Here, the predicted output y_t is as follows:

$$y_p(t) = h\left(\frac{B(q^{-1})}{F(q^{-1})}(f(u(t)))\right) \tag{3.29}$$

where $f\left(u(t)\right)$ is a nonlinear function transforming the input data. $\frac{B(q^{-1})}{F(q^{-1})}$ is a linear transfer function with $\{B, F\}$ being rational polynomials similar to the output error model and $h(\bullet)$ is another nonlinear function that maps the output of the linear block to the system output.

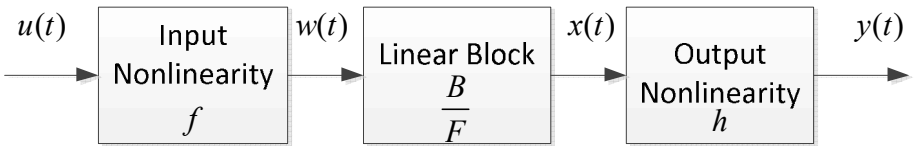


Figure 3.10: Block diagram representation for Hammerstein-Wiener model structure

Here, the input nonlinearity is a static function, implying that the output at a given time depends only on the input at that time. The input nonlinearity can be varied as a *sigmoid network*, *wavelet network*, *saturation*, *dead-zone*, *piecewise linear function*, *one-dimensional polynomial*, or some user-defined custom network [134]. Of these, the sigmoid and wavelet network have already been described above.

3.4.3.2.1 Piecewise linear nonlinearity estimator

A piecewise linear function of inputs with n breakpoints is used as an estimator, where the function is linearly interpolated between the breakpoints.

3.4.3.2.2 Saturation nonlinearity estimator

A saturation type nonlinearity is defined as a nonlinear function as shown in equation (3.30)

$$y = F(x) \tag{3.30}$$

where

$$\left\{\begin{array}{ll} a \le x < b \rightarrow F(x) = x \\ a > x \quad F(x) = a \\ b \le x \quad F(x) = b \end{array}\right\} \tag{3.31}$$

3.4.3.2.3 Deadzone nonlinearity estimator

A deadzone type nonlinearity is defined similarly to (3.30) where the function structure is

$$\left\{\begin{array}{ll} a \le x < b \rightarrow F(x) = 0 \\ a < x \quad F(x) = x - a \\ b \ge x \quad F(x) = x - b \end{array}\right\} \tag{3.32}$$

3.4.3.2.4 One dimensional (single variable) polynomial type nonlinearity estimator

The function is similar to equation (3.30) where

$$F(x) = c_1 x^n + c_2 x^{n-1} + \cdots + c_n x + c_{n+1} \tag{3.33}$$

In (3.33), the polynomial coefficients are represented by $c_1, c_2 \cdots c_{n+1}$ and n is the order of the poles. The input or output nonlinearity can also be ignored. The linear block may be configured by specifying the orders of the numerator $\{B\}$ and denominator $\{F\}$.

Just as the input nonlinearity is static, so the output nonlinearity is a static function as well. The output nonlinearity may be configured in the same way as the input nonlinearity. In all simulations presented in this chapter, the Levenberg-Marquardt search method [134] has been used by the *toolbox* to optimize the parameters of nonlinear models with a criterion for minimizing the determinant of the squared error ($\det\left(e^T e\right)$).

The options which have been varied for configuring the nonlinear models are listed in Table 3.1.

Table 3.1: Nonlinear model configurations

Model	**Estimators**	**Variable Parameters**			
		I/O nonlinearity	**Linear block**		
		No. of units	*poles*	*zeros*	*delay*
NLHW	Piecewise linear	5 to 40 (increment of 5)	1 to 10		1
	Sigmoid network	5 to 40 (increment of 5)	1 to 10		1
	Saturation	N.A	1 to 10		1
	Deadzone	N.A	1 to 10		1
	Wavelet network	Selected by Toolbox	1 to 10		1
	One-dimensional polynomial	1 to 10	Two sets of orders – 5 and 10		1
		Regressors	*No. of units in nonlinear block*		
NLARX	Tree-partition network	1 to 10	Selected by Toolbox		
	Wavelet network	1 to 10	Selected by Toolbox		
	Sigmoid network	1 to 10	Selected by Toolbox		

3.5 Experimental design

This section gives a brief description of the experimental set-up used for the initial data collection of electrophysiological signals of one Bay leaf (*Laurus nobilis*), two cucumber (*Cucumis sativus*) and 17 Zanzibar Gem (*Zamioculcas zamiifolia*) plants. These three plants were chosen because of their availability and to limit the number of species for controlled observation on the variation of responses. A mixture of plants was important to evaluate robust and generic models that can efficiently reflect the electrophysiological behaviour, irrespective of the species.

Approximately, 1 week to 2 year old pot-grown plants of different species were chosen for the experiments. These plants were subjected to a periodic white light stimulus of different pulse widths. The *extracellular* measurement of electrical signal responses of these plants was performed by inserting two metallic Electromyogram (EMG) needle electrodes (Bionen s.a.s.) into the petiole and stem at a distance of approximately 5 cm from each other. The setup is shown in Figure 3.11. The reference electrode was inserted into the plant body, nearer to the soil. A dual instrumentation amplifier EI-1040 [137] with a gain of 1 was chosen to provide the high input impedance without altering the actual amplitude of the

acquired signal. Low frequency and amplitude signals require very high input impedance and very low input bias currents. The EI-1040 provided an input impedance of 10 GΩ and an input bias current of 0.5 nA. National instruments data acquisition device USB 6008 [138] was used for analogue to digital (A/D) conversion at a sampling rate of 1 KHz which was then monitored using LABView 2012 [139] software on a personal computer (PC). This connection to the amplifier and the data acquisition devices (DAQ) is shown in Figure 3.12.

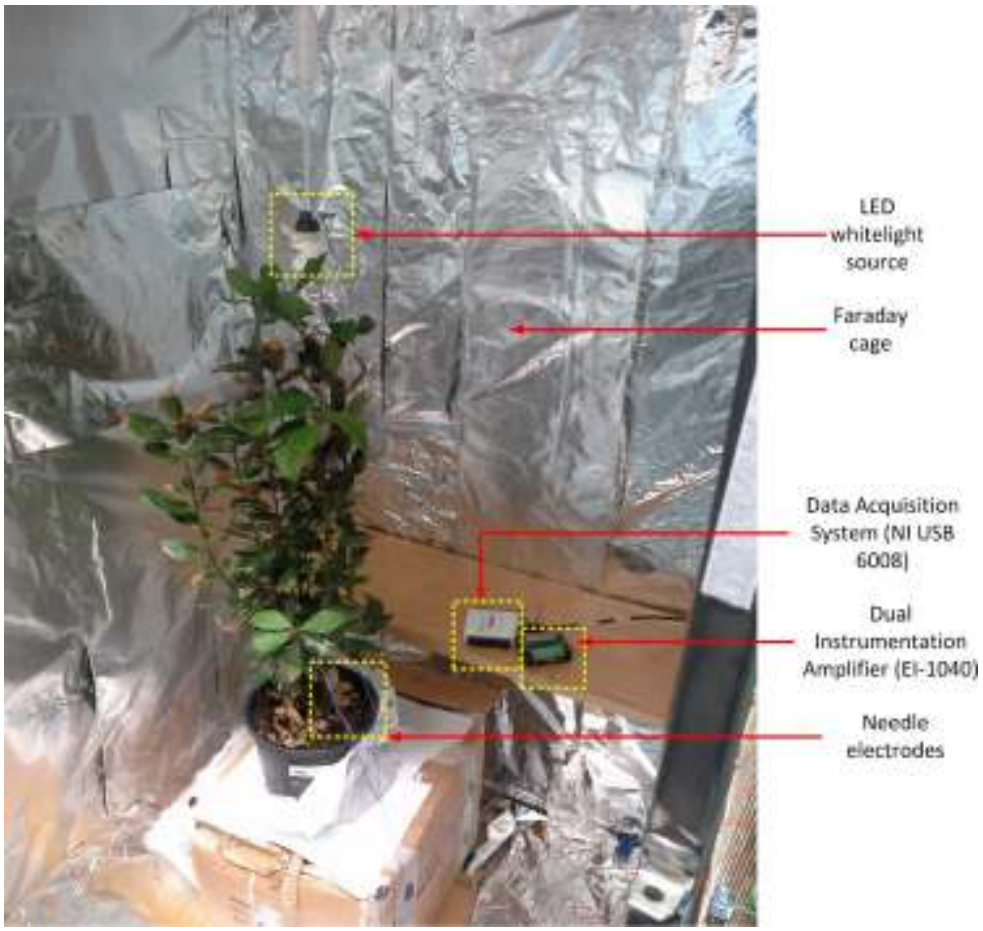


Figure 3.11: Experimental setup (followed in Rome/Florence) to obtain electrical response from a Bay leaf plant when exposed to white-light

As bio-electrical signals are reportedly weaker [140], an external electromagnetic field could induce a lot of noise in it. Therefore this setup (excluding the PC) was placed inside a grounded Faraday cage as shown in Figure 3.11. A digital low-pass filter with a cut-off frequency 1 Hz was provided to eliminate any noise associated with the measurement, as in [110], [112], [114]–[116], [141], [142] where it was reported that plant signals are slow oscillatory signals at a very low frequency. This cut-off frequency was experimentally chosen at this stage, and a more rigorous exploration to find an optimum cut-off frequency will be presented later. A LED light source was used for providing white-light at maximum

brightness. Here, the number of photons used for photosynthesis by plants was used as a basis for measurement of the incident light (in PAR units). Similar experiments and electrophysiological measurements on cucumber plants have been reported previously in [108].

The actual readings from a *quantum meter*, provide the measure of the *photosynthetically activated radiations* (PAR) at different wavelengths. It is never consistent for white light and fluctuates within a tiny bandwidth (for different wavelengths of light composition). Since a quantum meter was not available, the light intensity was measured using a standard LUX meter, and the PAR units calculated from it (using *cool white fluorescent lamp* as reference [143]). The conversion from LUX to PAR units is shown in Table 3.2.

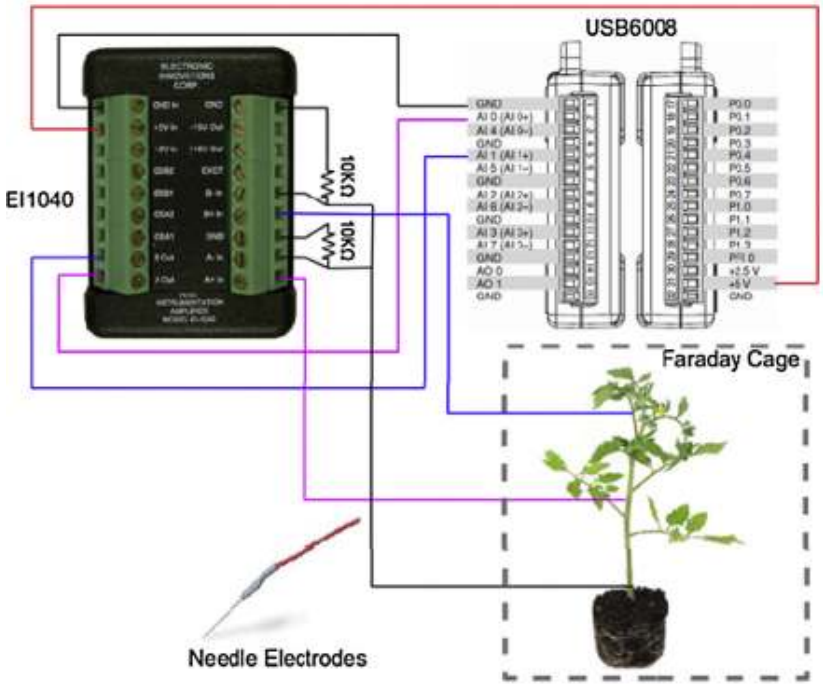


Figure 3.12: Connections between the devices employed to capture the electrical signals of the plants when subjected to light stimulus

Table 3.2: Conversion from Lux to PAR

LUX	**PAR** (μmol m^{-2} s^{-1})
1	0.0135

In order to get a robust model, the light pulse widths were varied during the experiments for 20 different plants, of two different species, and are given in Table 3.3. Each experiment involved a new plant and the data collected from it is referred to as Dataset (1, 2...and so on).

The aim of this exploration was to evaluate the behaviour of the plants to the presence (and absence) of light as a stimulus. The response included start-up transients, which was included in the overall behaviour of the plants. These transients were useful for evaluation of the models as a means to capture the plant's natural response.

Table 3.3: Variations in white-light pulse widths for each datasets (for 20 different plants)

Datasets	Plant species	Time (seconds)			
		First light pulse	Second light pulse	Third light pulse	Fourth light pulse
Main	*Laurus nobilis*	232.40	293.20	260.00	231.00
Dataset 1	*Zamioculcas zamiifolia*	147.80	151.80	167.00	148.00
Dataset 2	*Zamioculcas zamiifolia*	123.2	120	120	-
Dataset 3	*Zamioculcas zamiifolia*	119.99	119	121	-
Dataset 4	*Zamioculcas zamiifolia*	140.99	119.99	121	-
Dataset 5, 6,10-19	*Zamioculcas zamiifolia*	180	180	180	-
Dataset 7, 8	*Cucumis sativus*	180	180	180	-

Altogether, 19 experiments were performed for data collection by the project partners (as part of the EU project PLEASED) in Rome and Florence, Italy.

3.6 Signal pre-processing

As a starting point, the electrical signals from the plants were smoothed (moving average) using 10,000 samples (sampling rate was 1000 samples/sec), to reduce the stochasticity. This was done with an assumption that the on/off time of the light pulse event is embedded in the deterministic part (the trend) more than the stochastic part (the random part). The smoothing window of 10,000 samples was applied to all the datasets for uniformity. The raw versus smoothed signals are shown in Figure 3.13 in terms of number of samples. This smoothing function was applied to all the remaining 19 datasets and these are shown in Figure 3.14.

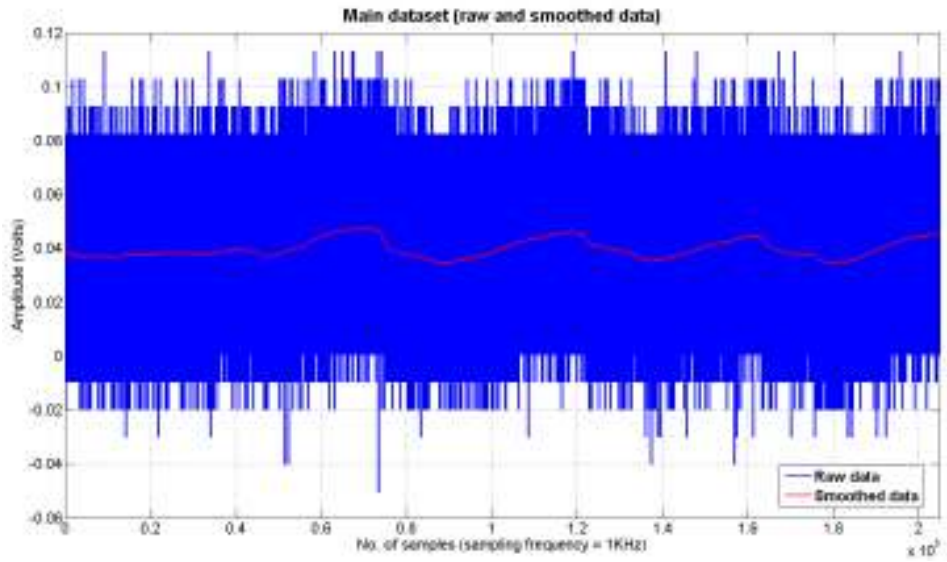
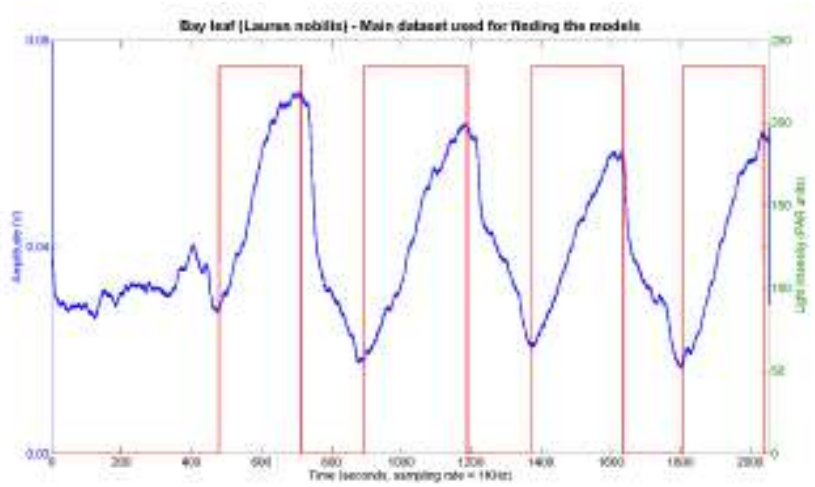


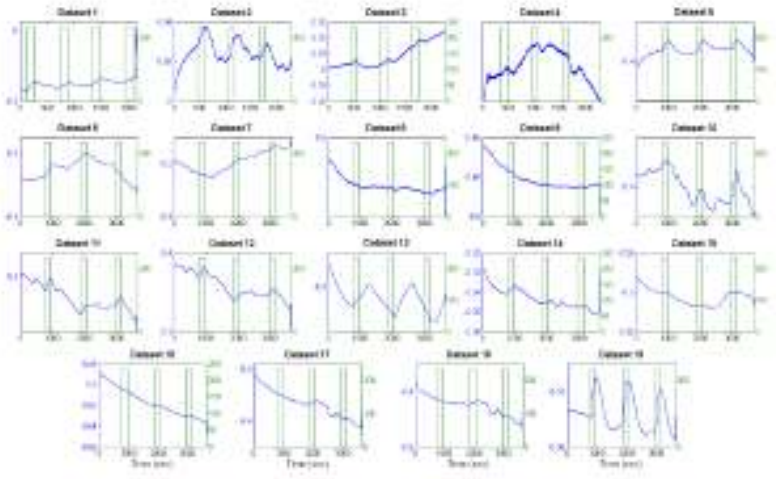
Figure 3.13: Raw versus smoothed electrical signal response of plant

Figure 3.14 (a) shows the electrical signal responses of the main dataset, when exposed to light pulses of varying widths. The y-axis of Figure 3.14 (a) shows both the amplitude of the electrical responses and the light intensities on parallel axes. Figure 3.14 (b) shows the 19 independent test datasets used to verify whether a similar response is seen for light pulses in order to aim for a common modelling framework.

A hypothesis was adopted that a common modelling framework would enable successful prediction of the stimulus for different plants. Whenever the light pulse was switched on or off, a change was observed in the gradient of the electrical response of the plants in all cases. These electrical responses were induced on the leaf tissue of the plants. Note that the extracellular measurements of the membrane potential on a plant leaf is a mixture of the individual responses of the guard cells, mesophyll cells, and the epidermal cell [116]. The morphology of the membrane potential of these cells can be different from what was recorded on the surface of the leaf tissue. The intention is to arrive at a black box model and thus the collective morphology of the membrane potential found on the tissue will suffice for the purpose of model building.



(a)



(b)

Figure 3.14: Plot of the variations in electrical signal response of the plant with respect to the incident light stimulus (a) main dataset (b) 19 independent test datasets

3.7 Results of forward and inverse modelling

Assuming that the electrical signals generated from plants came from a black box system, with the measured stimulus (light) and response (electrical signal) dataset, dynamical models were developed using the concept discussed in Section 3.4. For the present simulation studies, the *System Identification Toolbox* of MATLAB [134] were used to develop the input-output linear and nonlinear forward/inverse models. The idea was to develop a model whose predicted output best fitted the experimentally recorded or measured output, when the same

input was applied. Since this was the first exploration, some constraints were applied while trying to determine the best models. These constraints were varying certain model parameters such as input-output units and poles-zeroes and results were obtained within those constraints (Table 3.1).

The percentage fit shown in Figure 3.15 to Figure 3.21, to compare relative accuracies of different estimators, is given by the normalized root mean squared error (NRMSE) as equation (3.34).

$$fit = \left[1 - \frac{\|y - \hat{y}\|}{\|y - \bar{y}\|}\right] \times 100 \tag{3.34}$$

where $\hat{y}$ is the simulated or predicted model output, y is the measured output, and $\bar{y}$ is the mean of the output. When $\hat{y} > y$, fit was negative.

Figure 3.15 shows the plots having the best fits, using linear model variants ARX, ARMAX, BJ and OE, for both forward and inverse scenario of the main dataset. The y-axis shows the % fit with gradual increase in pole-zero order (as represented by na, nb, nc, nd, nf for various models previously described) from 1 to 10, as shown in the x-axis. Delay (nk) has been kept constant at 1 for all models. Of all four variants of linear LSE, the BJ model outperforms the others in both forward and inverse modelling, consistently giving higher accuracy.

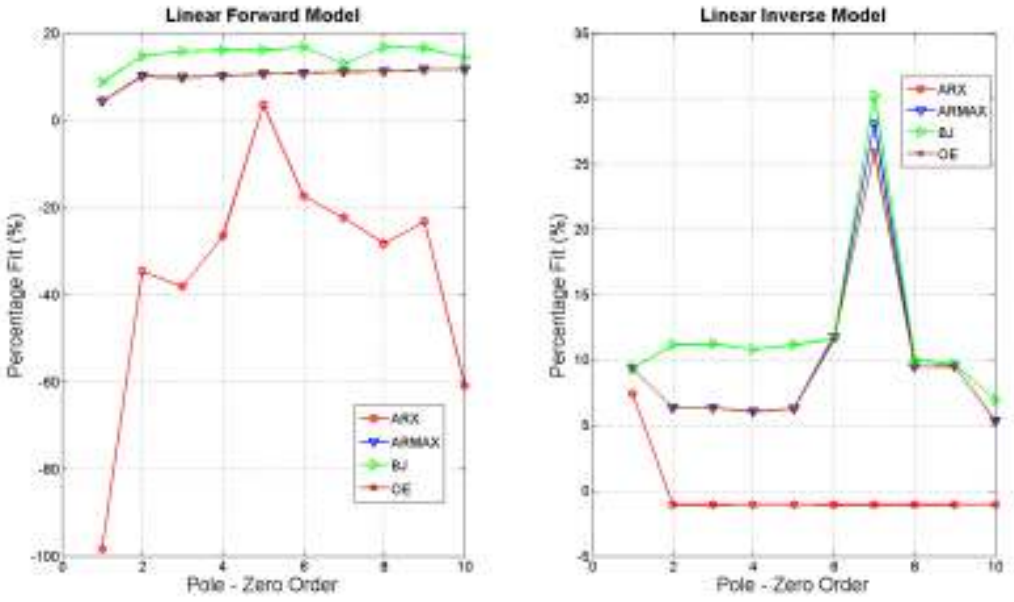


Figure 3.15: Linear models for forward and inverse modelling using the main dataset

It can also be seen that in forward modelling, the ARX model shows a high variation in percentage fit as the *pole-zero* order is varied, whereas ARMAX, BJ and OE models were more consistent while predicting the output electrical signal. In case of inverse modelling, the ARX model was more consistent than the other three models, but with a poorer fit compared to the others. ARMAX, BJ and OE showed a sudden jump in accuracy at a combination of pole-zero order as 7 (i.e. 7 poles, 7 zeros, and 1 sample delay).

All the linear model estimators performed poorly in both forward and inverse modelling settings. Similar simulations were undertaken using nonlinear modelling with NLARX variants wavelet network, sigmoid network and tree-partition, with a gradual increase in the number of regressors for input and output signal in both forward and inverse modelling. The modelling consistently gave negative percentage fits, the results have not been reported here. On the contrary, the NLHW estimators (with different static input-output nonlinearity) in most cases yielded better prediction accuracy as shown in Figure 3.16 to Figure 3.21.

Figure 3.16 shows the NLHW model with one dimensional polynomial as the nonlinearity type. This case uses two different pole-zero orders, 5 and 10 as starting points, and varied the number of input-output units for each order. While considering the forward model, a lot of variance was seen in percentage fit between input-output units from 1 to 4. Thereafter, from 5 to 8, the percentage fit seems to stabilize, and beyond 8 input-output units the fit suddenly drops.

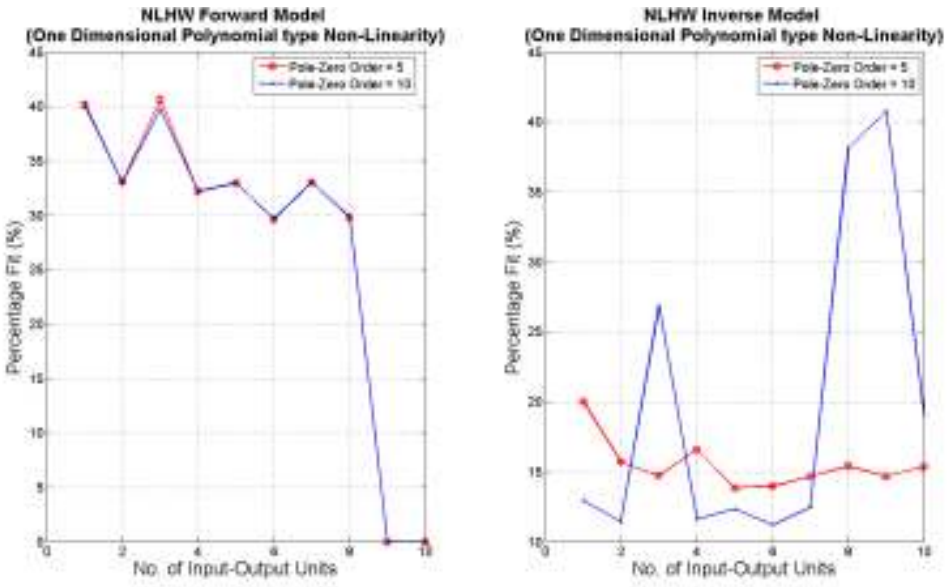


Figure 3.16: NLHW model with one dimensional polynomial as nonlinearity for forward and inverse modelling using the main dataset

The inverse modelling setting saw less variability in percentage fit when using pole-zero order of 5, but more variability when the pole zero order was set to 10, although the accuracy is improved for input-output units of 3 and 8-10. Again it was noticed that by using one dimensional polynomial type nonlinearity, the best fit was around 40%.

While using the dead-zone type nonlinearity, as shown in Figure 3.17, a stable prediction was observed during forward modelling for pole-zero orders higher than 2. However for inverse modelling with the same estimator configuration, it was found that the variation in the percentage fit was quite high. Again the fit was somewhere around 40% for both forward and inverse modelling.

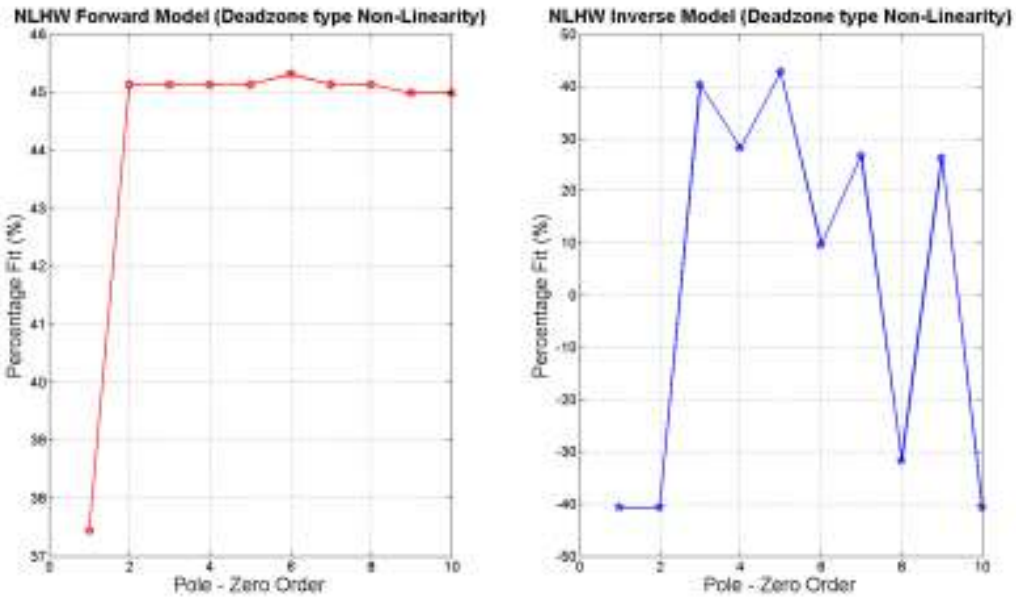


Figure 3.17: Nonlinear model with dead-zone as nonlinearity for forward and inverse modelling using the main dataset

In Figure 3.18, saturation type nonlinearity, in forward modelling, gave consistent prediction accuracy of around 52% for pole-zero orders above 1. In the inverse modelling case, some sort of consistency in accuracy was seen for pole-zero orders between 2 and 5 and randomly varying otherwise. The best fit during inverse modelling, using saturation type nonlinearity, was again around 40%.

For piecewise linear type static nonlinearity as the NLHW estimator, both the input-output units and also the pole-zero orders were varied for each case. Thus a surface plot was obtained of percentage fit as a function of these two, which is shown in Figure 3.19. The

input-output units were varied from 5 to 40, incrementing by 5 at a time. For each chosen pair of input-output units, the pole zero order was varied from 1 to 10. During forward modelling, the percentage fit was around 50% when using input-output units of 20 onwards and pole-zero orders between 2 and 8. Inverse modelling also had a percentage fit around 50% when the pole-zero orders were between 1 and 6 and input-output units were 35 and 40.

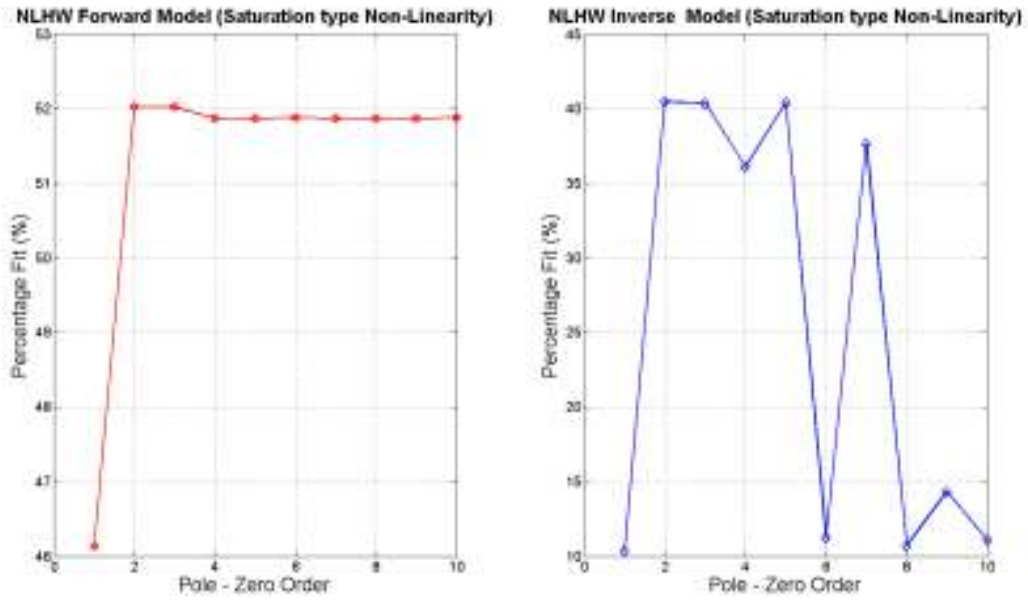


Figure 3.18: Nonlinear model with saturation as nonlinearity, for forward and inverse using the main dataset

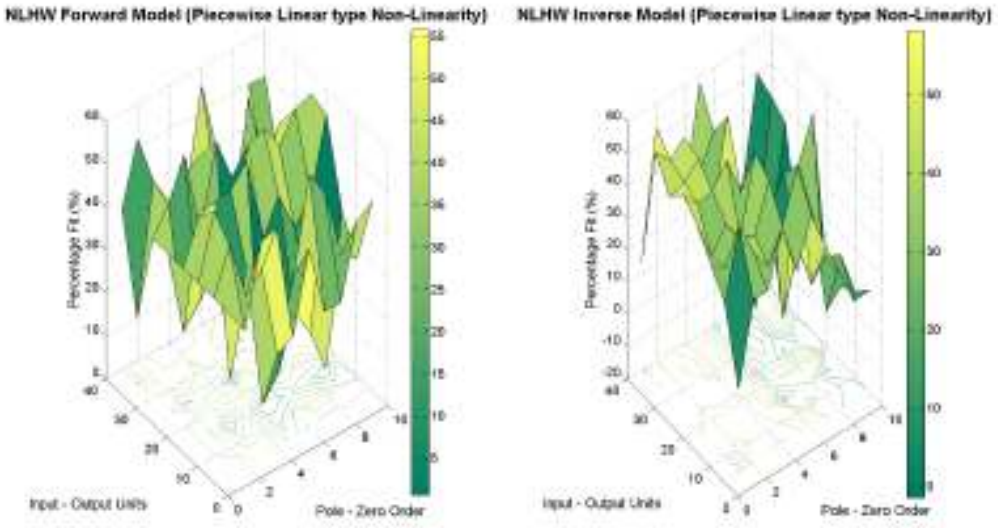


Figure 3.19: Nonlinear model with piecewise linear as nonlinearity for forward and inverse modelling using the main dataset

When using sigmoid type nonlinearity, Figure 3.20 shows that during forward modelling, percentage fit peak occurred at around 50% when input-output unit was 30 and decreased thereafter. For inverse modelling, a peak fit of around 80% occurred for number of units of 35 and pole-zero order of 10.

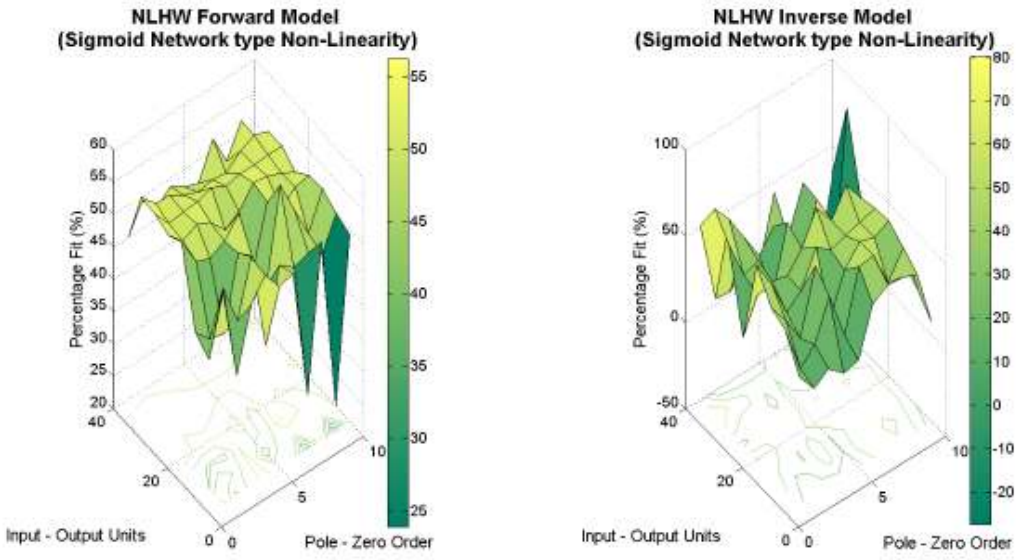


Figure 3.20: Nonlinear model with sigmoid as nonlinearity, for forward and inverse modelling using the main dataset

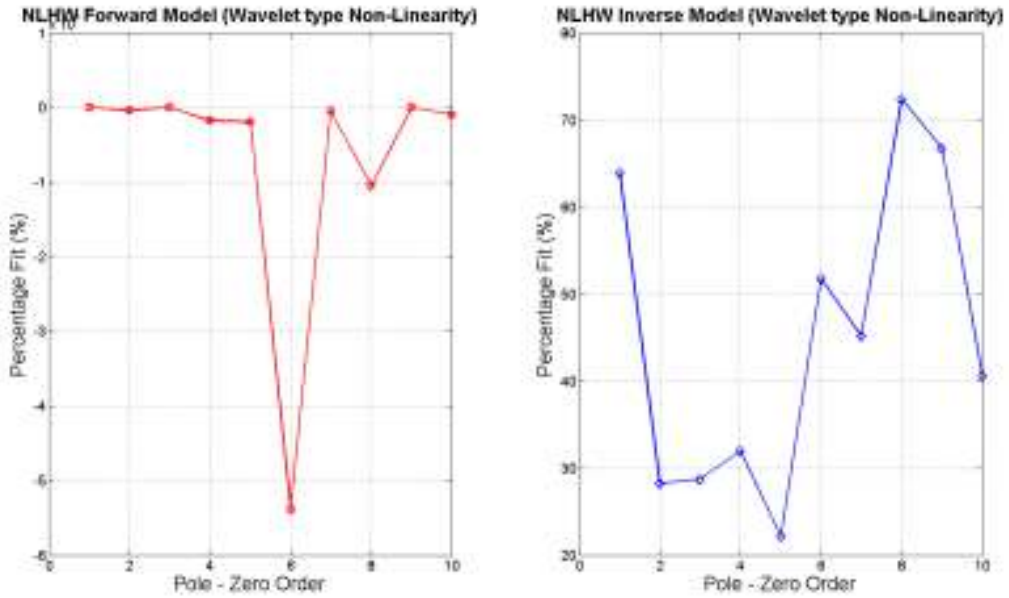


Figure 3.21: Nonlinear model with wavelet network as nonlinearity, for forward and inverse modelling using the main dataset

Lastly, Figure 3.21 shows the wavelet network type of static nonlinearity. During forward modelling a fairly consistent prediction accuracy (although in terms of negative percentage fit) was noticed, except pole-zero orders of 6 and 8. During inverse modelling a high variability in the prediction accuracy was observed with gradual increase in pole-zero order.

The best found forward models are shown in Figure 3.22, which were the linear BJ model with 6 pole-zero order and the NLHW model with piecewise linear type nonlinearity (10 pole-zero order and 25 input-output units). In the case of inverse modelling, the best configurations were achieved using NLHW with sigmoid nonlinearity with pole-zero order of 10 and 35 input-output units, amongst the nonlinear estimators, and BJ with pole-zero order of 7 amongst the linear estimators. This is shown in Figure 3.23.

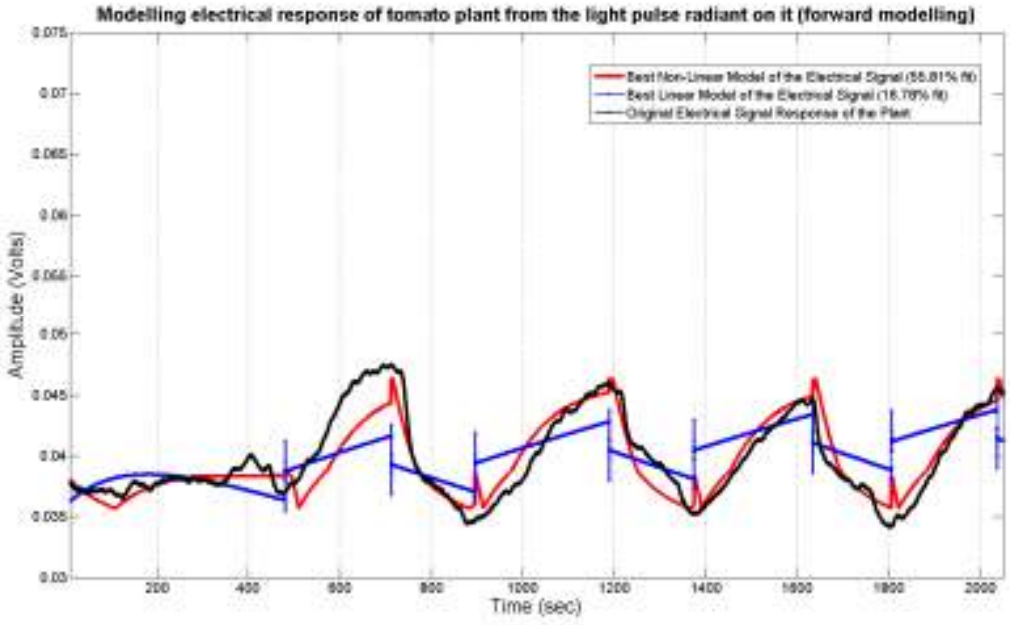


Figure 3.22: Best linear and nonlinear model estimates during forward modelling using the main dataset

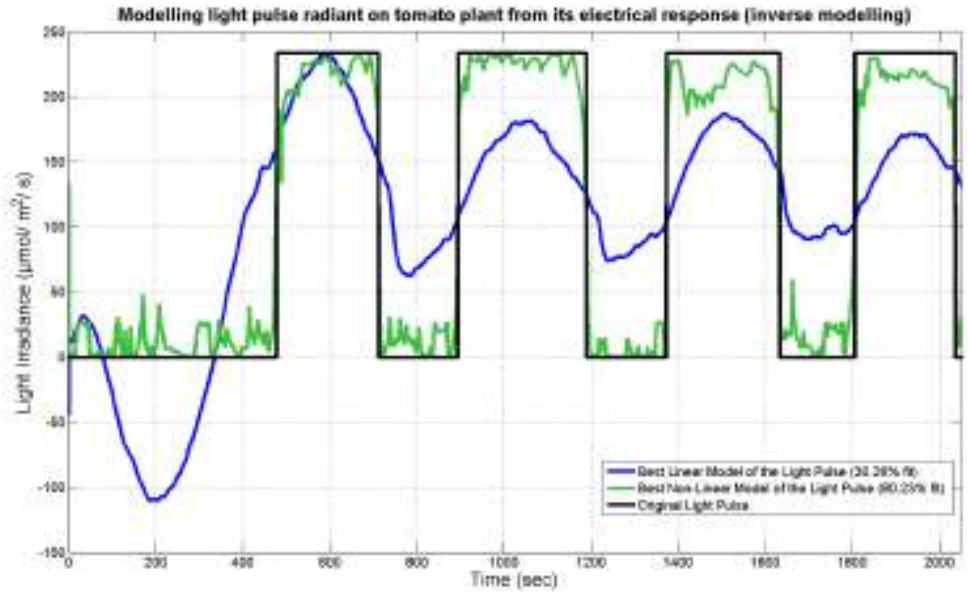


Figure 3.23: Best linear and nonlinear model estimates during inverse modelling using the main dataset

From the best forward and inverse models, shown in Figure 3.22 and Figure 3.23 respectively, the rising and falling of electrical signals and switching on and off of light pulses were captured as on-time (t_{on}) and off-time (t_{off}). The method of computing on-time and off-time can be seen (although only non-linear models are shown as an example) in Figure 3.24 and Figure 3.25.

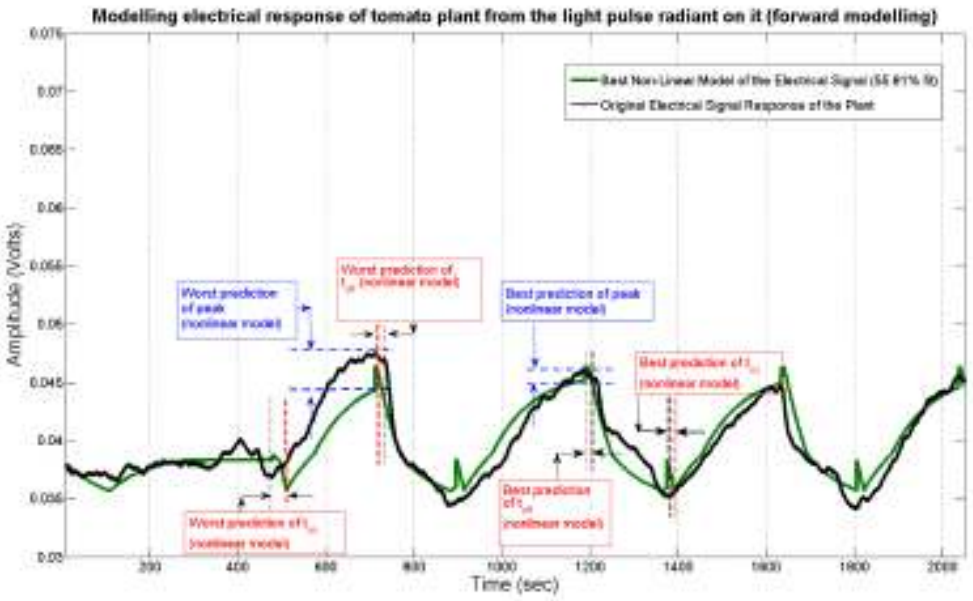


Figure 3.24: Measuring best and worst prediction of peaks and rise times/fall times of the plant electrical signal (forward modelling)

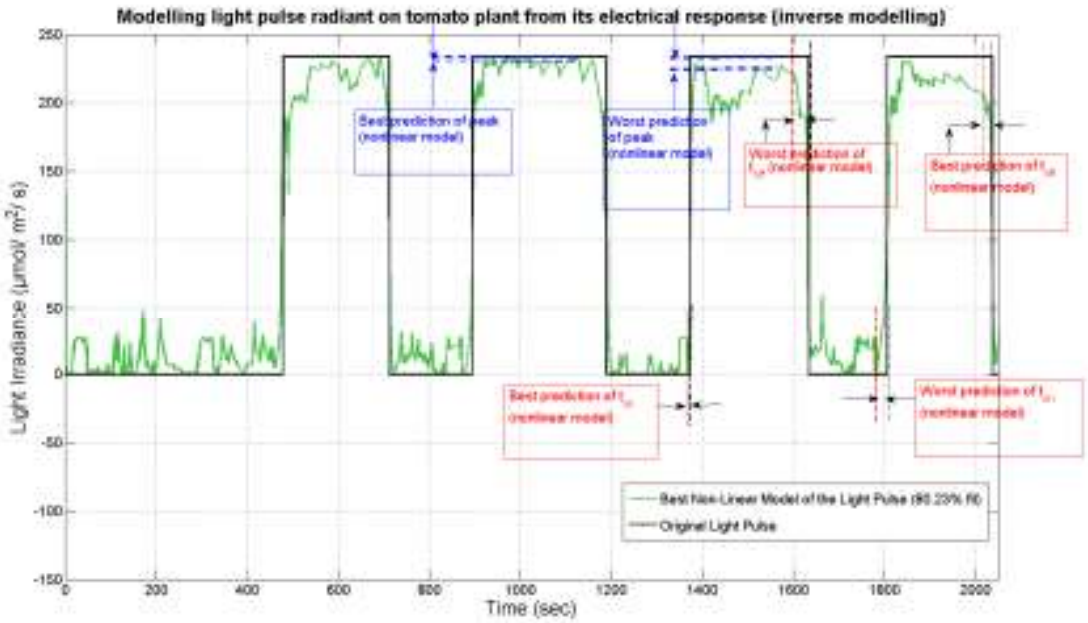


Figure 3.25: Measuring best and worst prediction of peaks and t_{on}, t_{off} of the predicted light-pulse (inverse modelling)

Next, the best and worst predictions were calculated from on-time and off-time for the light pulse (the difference between actual and predicted values) in the inverse modelling case. Similarly, rise time and fall time instants for the electrical signal (difference between actual and predicted values) in the forward modelling case were computed. These were denoted by Δt_{on}, Δt_{off}. The difference between the actual and predicted peaks of light intensity (inverse) and electrical signal (forward) denoted by Δpeak were also noted.

The methodology for defining the best and worst prediction accuracy for t_{on}, t_{off} and peak, amongst different light pulses is shown in Figure 3.24 and Figure 3.25 respectively from the forward and inverse modelling perspective, with the best and worst values of these three parameters reported in Table 3.4 and Table 3.5.

Table 3.4: Best *forward model* results with % fit, rise time instant (t_{on}), fall time instant (t_{off}) and peak of the electrical signal response for the main dataset

Estimator configuration	**% fit**	**Best Case**			**Worst Case**		
		Δ t_{on} (sec)	**Δ t_{off} (sec)**	**Δ peak (Volts)**	**Δ t_{on} (sec)**	**Δ t_{off} (sec)**	**Δ peak (Volts)**
Piecewise linear-10,10,1 (nonlinear)	55.81	+17	+14	+0.00079	–50.4	+62	+0.0431
BJ-6,6,6,6,1 (linear)	16.78	+39.3	+11	–0.00111	+107.2	+46	–0.00608

Table 3.5: Best *inverse model* results with % fit, switching on (t_{on}) and switching off (t_{off}) and peak of the light pulse for the main dataset

Estimator configuration	**% fit**	**Best Case**			**Worst Case**		
		Δ t_{on} (sec)	**Δ t_{off} (sec)**	**Δ peak (µmol/m^2/sec)**	**Δ t_{on} (sec)**	**Δ t_{off} (sec)**	**Δ peak (µmol/m^2/sec)**
Sigmoid-10,10,1 (nonlinear)	80.23	+7.1	+5.2	–2.0	+9.5	+6.0	–45
BJ-7,7,7,7,1 (linear)	30.26	+13	+26	–6.7	+18.8	+51.2	–63

A positive value of $\Delta\ t_{on}$ or $\Delta\ t_{off}$ denotes the simulated result occurs before the actual rise or fall time of the pulse respectively, whereas a negative value denotes the simulated value occurs after the actual event. Similarly, a positive value of Δ peak denotes a higher predicted peak amplitude than the actual value, while a negative value for Δ peak denotes the predicted value is lower than the actual value.

The estimator configurations BJ-6,6,6,6,1 for forward model and BJ-7,7,7,7,1 for inverse model in Table 3.4 and Table 3.5 represent the values for n_b, n_f, n_c, n_d and n_k, which denotes the order of the polynomials for the deterministic and stochastic part of the model structure and the delay unit respectively (see Section 3.4.2.3). Similarly for the NLHW class of models

with piecewise linear and sigmoid type static nonlinearity, the number of poles, zeroes and delay-unit is represented by 10,10,1 (i.e. 10 poles, 10 zeroes, 1 delay unit) respectively.

The main dataset (on *Laurus nobilis*) was used for the rigorous parameter estimation of different model structures, which showed that the NLHW model gave the best results for both forward and inverse model settings. The top three NLHW estimator settings (termed as model 1, 2 and 3), were therefore selected for both the forward and inverse problem, which yielded the best accuracies on the main dataset. These estimator configurations are given in Table 3.6.

Thereafter, the NLHW estimator configurations thus chosen were applied to 19 other independent datasets (17 *Zamioculcas zamiifolia* and 2 *Cucumis sativus*) to see if they consistently produce acceptable forward and inverse predictions. The results are reported in Table 3.7, from which it is evident that even in some of the other datasets, the NLHW models produced positive prediction accuracy. Any accuracy above 60%, for datasets 1-19, has been highlighted in bold (the main dataset is highlighted in blue).

The inverse models also produced comparable accuracy using the top three model configurations. Although the electrical signals in different plants were slightly different shapes, as can be seen in Figure 3.14 (a) and (b), the forward and inverse estimation accuracies seemed promising. The best settings of the estimators were found from the main dataset and hence some deviation or fall in prediction accuracy by the same estimator configuration for the other datasets was expected.

Note that the plant has been modelled as a *Single Input Single Output* (SISO) system (between light stimulus and obtained electrical response) and has ignored the impact of changing other factors such as temperature and humidity on plant electrophysiology, since their variation within the experimental time period was negligibly small.

As the photosynthetic light intensity was varied (i.e. light switched on and off), the polarization of the guard cells changed too, thereby affecting the K^+ ion concentration and this affected the ionic current recorded. In contrast, a natural (un-manipulated) variation of temperature and humidity can always be considered to be constant during the experiment conducted over a short period of time [40], and hence would not influence the ionic current within the plant cells.

Table 3.6: Top three estimator settings for the main dataset during forward and inverse modelling

Class of model	Model number	Nonlinearity in NLHW estimator	Input Units	Output Units	Poles	Zeroes	Delay
Inverse Models	1	Sigmoid	35	35	10	10	1
	2	Wavelet	-	-	8	8	1
	3	Sigmoid	35	35	1	1	1
Forward Models	1	Piecewise Linear	25	25	10	10	1
	2	Piecewise Linear	40	40	6	6	1
	3	Piecewise Linear	25	25	9	9	1

Table 3.7: Top three fits for each datasets during inverse modelling using the same model parameters

Dataset	Temp. (°C)	Humidity (%)	Inverse Models (% fit)			Forward Models (% fit)		
			Model-1	Model-2	Model-3	Model-1	Model-2	Model-3
Main	23.8	49	80.23	72.36	69.53	55.81	55.76	54.42
1	24.4	55	59.70	5.23	7.72	26.09	0.62	25.72
2	24.2	57	64.88	18.45	7.58	40.81	40.46	40.32
3	23.7	52	38.04	31.32	17.80	85.49	94.97	83.86
4	24.6	49	–2.91	26.34	36.49	14.01	62.98	19.50
5	22.8	47	45.29	48.06	63.90	59.24	33.43	44.14
6	22.4	54	2.07	12.73	62.17	27.00	70.44	14.28
7	26.3	47	56.38	28.15	55.81	91.88	42.43	91.88
8	23.3	49	28.04	54.76	53.46	21.50	50.28	78.57
9	23.3	49	22.01	73.90	8.75	86.86	91.85	85.74
10	24.4	53	36.39	53.41	54.12	42.09	65.40	20.40
11	24.6	52	2.23	9.22	56.50	67.81	66.27	65.14
12	24.6	52	5.64	53.41	3.18	67.76	26.05	66.03
13	23.7	52	69.51	17.78	41.79	50.44	39.10	18.81
14	24.8	56	72.99	26.46	34.30	23.56	77.90	79.38
15	23.1	50	25.34	60.90	47.93	90.41	91.81	93.31
16	23.1	35	30.42	75.06	67.38	97.70	98.34	46.49
17	23.1	50	31.30	71.77	53.55	91.18	11.00	90.48
18	23.1	50	72.95	31.45	54.02	74.95	18.77	25.08
19	24.9	58	50.43	30.44	67.81	72.20	57.42	57.18
Avg.	-	-	**39.54**	**40.06**	**43.18**	**59.33**	**54.76**	**55.03**

There are also some theoretical arguments that a change in temperature of at least 10° C is necessary to generate action potentials in plants [125]. Since within the duration of each

experiment, the temperature change was negligible, it has not been considered in the modelling.

In Table 3.7, the ambient temperature and humidity within the Faraday cage has been reported for each experiment (which were noted from the original database), but due to their static nature they have not been considered as additional inputs for the modelling. There was the possibility of an increase in leaf temperature if the light source was very close to the plant. However, since the light source used for the experiments were LEDs, which produce cool lighting, the temperature inside the Faraday cage only was measured as an indicator of the ambient as well as the plant surface temperature. A larger deliberate variation in temperature (heat or cold shock) could be undertaken as an additional input to investigate thermal effects on such models in future research.

The average of the fit values for all three models (for both Inverse and Forward modelling) are given in the last row of Table 3.7. These averages show that the top three models (and their poor fits on the test data) are not reflective enough of the plants as a system. Hence, there is a need to find the models from a pool of larger datasets (rather than just one dataset), and also to test them on another pool of separate test datasets. In order to gather more datasets, separate experiments were conducted as explained below.

3.8 New Experimental data exploration

As seen earlier, the results were not consistent and reliable. This is because the electrical response of the plants did not show any consistency compared to the main dataset. Hence, new experiments were set up with new plants. To achieve more robustness, the models need to be determined by exploring more data and then testing them on different data again. The exploration so far revealed that some models give a good fit for some datasets, while others such as NLARX does not. The same set of models will be explored as used on the main dataset, except that NLARX will be omitted.

In order to gather new data, exactly the same experimental setup was used. A microcontroller-controlled light bulb kit (Arduino shield) was obtained from Autohometion [144] and programmed for square light pulses of fixed widths (the Arduino code is listed in Appendix A) using the library provided by the manufacturer [145]. This shield is compatible

with Arduino and has a 2.4 Ghz radio transceiver on board that was used to control the smart LED bulbs.

The DTP-1309 light sensor originally used was replaced, as its software was outdated and in error, and the manufacturer had discontinued support for this sensor. Thus, a custom-built light intensity sensor (BH1750FVI) was used in conjunction with a *NodeMCU wifi-enabled* microcontroller to make light sensor readings every 10 secs, using an open source library available online [146]. The sensor readings were uploaded to a cloud IoT platform [147]. To ensure its accuracy, the readings from the custom-built sensor were compared with those from the commercial sensor on several occasions. From a sample of 30 comparison sensor readings, negligible average deviation of about 10 lux was found between the two sensors.

Figure 3.26 shows this new experimental setup, using the same DAQ (NI USB 6008, used by project partners in Florence/Rome). Using this setup, 30 new experimental datasets were collected from 15 experiments (two channels per experiment). Of these 30 new datasets, 18 were set aside for identifying the top models for forward and inverse modelling, while 12 were set aside for testing. Along with these 18 datasets, all 20 previous datasets were also used for training, making 38 training datasets altogether.

For the new experiments, five *Chrysanthemum* plants, approximately 10 weeks old, were obtained locally. Each plant was used for three experiments each, at different times. Thus, 18 datasets from three plants (3 plants x 3 experiments x 2 channels) were used for training, while 12 datasets from two plants (2 plants x 3 experiments x 2 channels) were used for testing. The response of the *Chrysanthemum* were found to be far more repeatable, and clear trends were found during switching on/off of the light pulses when compared with the old datasets comprising Cucumber and Zanzibar Gem plants.

These data for finding and testing the top models are shown in Figure 3.27 and Figure 3.28 respectively.

Figure 3.26: New experimental setup to collect data by controlling light pulses using an Arduino

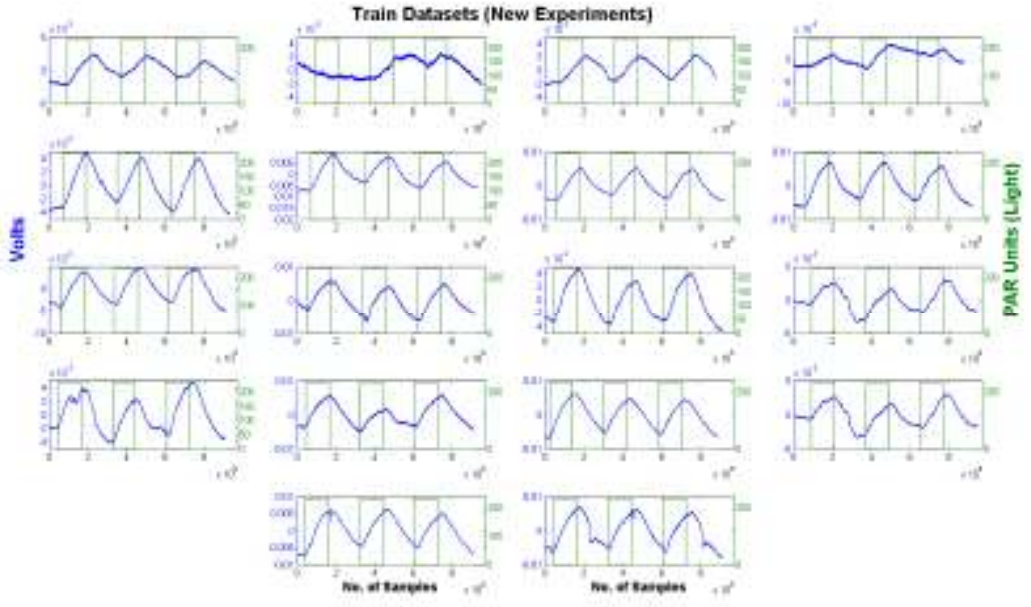


Figure 3.27: Electrical responses of *Chrysanthemum* – 18 new experimental data used for identifying the top models

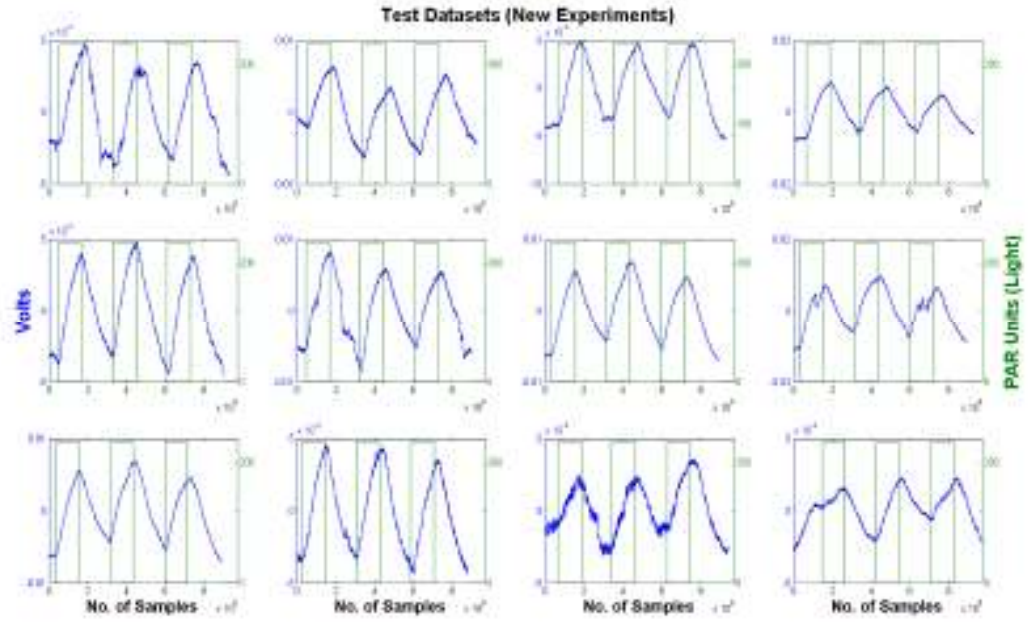


Figure 3.28: Electrical responses of *Chrysanthemum* – 12 new experimental data used for validating the top models

The data collection was done at 1 KHz sampling rate, exactly as before. The light pulses were of 120 seconds duration, followed by a 160 seconds of darkness. Three such pulses were repeated in each experiment. As can be seen from Figure 3.27 and Figure 3.28, the responses

from all the *Chrysanthemum* plants were found to be much more responsive to white light pulses (i.e. clear trends are visible) than most datasets shown in Figure 3.14.

The top three models found for each dataset during forward and inverse modelling scenario are given in Table 3.8 to Table 3.11.

The number of Poles, Zeros and Delays for any model is given as set of three digits such as 221. Delay was kept at 1 for all models. Similarly, the number of input/output units is described by N e.g. when N = 8, it means input/output units = 8. Other abbreviations are: Piecewise Linear (PL), Sigmoid Network (Sig), Wavelet Network (Wav), Box Jenkins (BJ) and One Dimensional Polynomial (Poly_1D) which were found to give the best results for different datasets.

Table 3.8: Top three models during forward modelling scenario for training with *new* datasets

Forward Models						
New Datasets	**Model 1**	**%**	**Model 2**	**%**	**Model 3**	**%**
Exp 1	PL 881/N=5	63.94	PL 991/N=30	63.93	PL 991/N=35	62.93
Exp 2	Wav 10101	56.04	PL 661/N=25	53.14	PL 661/N=20	44.02
Exp 3	Sig 771/N=35	82.61	Sig 661/N=20	82.34	Sig 881/N=40	82.23
Exp 4	Wav 111	64.83	Sig 111/N=35	61.60	Sig 111/N=30	60.98
Exp 5	PL 10101/N=15	80.04	PL 661/N=10	79.15	PL 331/N=20	76.49
Exp 6	PL 661/N=20	75.00	PL 661/N=35	74.92	PL 771/N=40	74.72
Exp 7	PL 991/N=25	85.77	PL 441/N=40	85.76	PL 881/N=35	85.74
Exp 8	Wav 10101	81.82	PL 771/N=20	81.12	PL 441/N=40	81.09
Exp 9	PL 221/N=35	82.22	PL 771/N=20	82.22	PL 991/N=10	82.16
Exp 10	PL 991/N=15	74.04	PL 221/N=35	71.24	PL 771/N=40	71.22
Exp 11	PL 221/N=40	79.25	PL 221/N=5	79.23	PL 221/N=35	79.23
Exp 12	PL 221/N=40	79.77	PL 661/N=25	79.75	PL 221/N=35	79.72
Exp 13	PL 221/N=30	61.05	PL 221/N=35	61.04	PL 221/N=40	61.03
Exp 14	PL 661/N=40	53.90	PL 661/N=30	53.89	PL 441/N=10	53.82
Exp 15	PL 221/N=30	83.57	PL 10101/N=10	83.53	PL 221/N=15	83.35
Exp 16	PL 10101/N=20	84.98	PL 10101/N=25	84.97	PL 331/N=30	84.96
Exp 17	Wav 991	72.21	PL 331/N=35	71.65	PL 10101/N=40	71.64
Exp 18	Wav 991	75.68	PL 331/N=35	75.25	PL 881/N=40	75.23

Table 3.9:Top three models during forward modelling scenario for training with *old* datasets

Forward Models						
Old Datasets	**Model 1**	**%**	**Model 2**	**%**	**Model 3**	**%**
Main dataset	PL 10101/N=25	55.81	PL 661/N=40	55.76	PL 991/N=25	54.42
Dataset 1	PL 881/N=40	45.61	PL 881/N=35	42.46	PL 661/N=30	37.99
Dataset 2	BJ 771	57.05	PL 10101/N=35	50.02	PL 441/N=40	46.78
Dataset 3	PL 10101/N=40	94.97	Sig 881/N=20	92.84	PL 221/N=35	92.59
Dataset 4	Wav 661	83.15	Wav 771	82.86	Wav 991	82.37
Dataset 5	Wav 661	87.52	PL 10101/N=25	59.24	Wav 771	57.28
Dataset 6	PL 661/N=40	70.44	PL 881/N=35	40.63	PL 661/N=30	36.62
Dataset 7	PL 991/N=25	91.88	BJ 771	55.47	PL 10101/N=35	48.69
Dataset 8	PL 771/N=20	99.20	PL 221/N=35	95.96	Sig 111/N=10	94.69
Dataset 9	PL 661/N=40	91.84	PL 10101/N=25	86.86	Wav 661	86.82
Dataset 10	PL 661/N=40	65.40	PL 881/N=35	42.60	PL 10101/N=25	42.09
Dataset 11	PL 10101/N=25	67.81	PL 661/N=40	66.22	BJ 771	59.41
Dataset 12	PL 441/N=40	95.71	Sig 881/N=20	93.55	PL 221/N=35	92.68
Dataset 13	Wav 661	94.97	PL 10101/N=25	50.44	PL 661/N=40	39.1
Dataset 14	PL 991/N=25	79.38	PL 661/N=40	77.90	PL 881/N=35	50.43
Dataset 15	PL 991/N=25	93.31	PL 661/N=40	91.81	BJ 771	67.5
Dataset 16	Sig 661/N=20	99.94	PL 221/N=30	99.42	Sig 881/N=40	99.41
Dataset 17	Wav 111	94.26	PL 10101/N=25	91.17	PL 661/N=10	37.69
Dataset 18	PL 10101/N=25	74.94	PL 881/N=40	62.91	Sig 771/N=35	58.58
Dataset 19	PL 10101/N=25	72.20	BJ 771	71.57	PL 10101/N=35	65.81

Table 3.10: Top three models during inverse modelling scenario for training with *new* datasets

Inverse Models						
New Datasets	**Model 1**	**%**	**Model 2**	**%**	**Model 3**	**%**
Exp 1	Wav 10101	93.95	Sig 441/N=20	85.91	Sig 661/N=35	85.57
Exp 2	Sig 551/N=5	49.14	Sig 661/N=40	34.56	Sig 991/N=20	32.46
Exp 3	Wav 661	96.13	Sig 10101/N=40	94.03	Sig 661/N=35	90.22
Exp 4	Sig 111/N=30	66.75	Sig 111/N=40	64.58	Sig 111/N=20	62.73
Exp 5	Wav 221	94.00	Sig 991/N=35	90.35	PL 221/N=15	78.25
Exp 6	PL 551/N=20	80.50	PL 551/N=10	78.18	PL 551/N=15	77.72
Exp 7	Sig 991/N=5	86.80	Sig 551/N=10	86.79	Sig 551/N=25	85.62
Exp 8	Sig 111/N=25	93.27	Sig 991/N=30	88.89	PL 221/N=10	86.95
Exp 9	Sig 441/N=40	87.68	Sig 991/N=30	83.94	Sig 661/N=35	81.96
Exp 10	Sig 111/N=40	86.02	Sig 111/N=25	85.78	Sig 551/N=5	85.66
Exp 11	Sig 10101/N=30	92.42	Sig 10101/N=20	91.32	Sig 441/N=10	91.21
Exp 12	Sig 10101/N=30	93.09	Sig 10101/N=20	91.74	Sig 441/N=10	91.87
Exp 13	PL 10101/N=40	89.96	Sig 10101/N=35	88.11	PL 661/N=25	85.75
Exp 14	PL 881/N=5	81.39	Sig 441/N=30	80.64	PL 331/N=20	79.23
Exp 15	Sig 111/N=20	85.33	Sig 661/N=35	83.35	PL 661/N=15	80.74
Exp 16	Sig 111/N=20	87.12	Sig 661/N=35	85.27	PL 661/N=15	82.48
Exp 17	Wav 441	91.23	Sig 221/N=35	84.41	PL771/N=15	78.99
Exp 18	Wav 441	94.92	Sig 221/N=35	87.84	PL771/N=15	82.65

Table 3.11: Top three models during inverse modelling scenario for training with *old* datasets

Inverse Models						
Old Datasets	**Model 1**	**%**	**Model 2**	**%**	**Model 3**	**%**
Main dataset	Sig 10101/N=35	80.23	Wav 881	72.36	Sig 111/N=35	69.53
Dataset 1	PL 661/N=40	76.70	PL 551/N=20	75.47	Sig 221/N=25	74.78
Dataset 2	Sig 10101/N=35	64.88	Sig 441/N=20	56.12	PL 661/N=40	55.71
Dataset 3	Sig 221/N=20	59.79	Sig 221/N=15	58.01	PL 991/N=30	57.81
Dataset 4	Sig 111/N=40	77.55	Poly 1D_5_Poles/ Deg = 10	56.48	Sig 991/N=35	50.24
Dataset 5	Sig 111/N=35	63.90	Poly 1D_5_Poles/ Deg = 10	61.15	Sig 441/N=40	56.91
Dataset 6	Sig 111/N=35	62.17	PL 661/N=40	57.79	Sig 441/N=40	57.35
Dataset 7	PL 661/N=40	78.02	PL 551/N=20	77.49	Sig 221/N=25	76.35
Dataset 8	Poly 1D_5_Poles/ Deg = 10	58.26	Sig 441/N=40	53.48	Sig 111/N=35	53.46
Dataset 9	Wav 881	73.90	Sig 221/N=20	60.78	Sig 221/N=15	59.54
Dataset 10	PL 661/N=40	59.62	Sig 441/N=20	58.95	Sig 551/N=25	57.31
Dataset 11	Sig 111/N=40	81.02	Sig 111/N=35	56.49	Wav 441	56.40
Dataset 12	Poly 1D_5_Poles/ Deg = 10	60.63	Sig 441/N=40	55.43	Sig 441/N=30	54.16
Dataset 13	PL 661/N=40	80.81	PL 551/N=20	80.06	Sig 221/N=25	78.95
Dataset 14	PL 661/N=40	74.49	PL 551/N=20	73.31	Sig 10101/N=35	72.99
Dataset 15	Wav 441	88.99	Sig 221/N=35	82.23	Sig 551/N=25	77.24
Dataset 16	PL 661/N=40	79.11	Sig 221/N=25	76.84	Sig 221/N=15	76.23
Dataset 17	Wav 881	71.77	Sig 441/N=20	58.87	PL 661/N=40	57.98
Dataset 18	Sig 10101/N=35	72.95	Sig 221/N=20	62.00	PL 991/N=30	60.46
Dataset 19	Sig 111/N=40	79.98	Sig 991/N=35	52.34	Poly 1D_5_Poles/ Deg = 10	58.27

The model exploration was carried out using a script to automate and optimise the whole process. The scripts were written to automatically output the results into Excel sheets and at the end of the simulation, an intimation will be sent out through an automated email so that another dataset could be used for model exploration. This script is provided in Appendix B.

The exploration revealed 44 different models for the Forward modelling scenario and 42 different models for the Inverse modelling scenario. These models were found by noting the top three models for each dataset. To validate and choose the overall top three models, all 86

models (44 forward and 42 inverse) were tested on the 12 datasets retained for this purpose. The top three generic models (based on the average fit for all the 12 test datasets) were then recommended to best capture the relationship between the light as stimulus and plant electrical signal response. The average results of all the models which were tested on the 12 separate datasets retained are given in Table 3.12 and Table 3.13.

From these tables, it is noted that Piecewise Linear (PL) for forward and Sigmoid Network (Sig) for inverse modelling scenarios, were the best structures found. The average fits and the models are given in Table 3.14 for reference. These average fits are found to be much better than those obtained during the first exploration.

Table 3.12: Evaluation of top 44 models during *forward* modelling scenario, on 12 test datasets

		Poles/Zeroes	N	Dataset-1	Dataset-2	Dataset-3	Dataset-4	Dataset-5	Dataset-6	Dataset-7	Dataset-8	Dataset-9	Dataset-10	Dataset-11	Dataset-12	Avg.	Rank
Model-1	BJ	7		0.1149346	0.036199	70.102	68.90192	70.99764	67.03483	76.73852	66.63821	59.90418	0.0130404	3.7085631	72.09066	46.35839	
Model-2	Wavelet Network	111	Auto	34.677577	28.52912	56.89955	55.67599	64.93825	59.61264	63.1711	60.25649	37.45375	43.827884	31.721691	64.447199	50.10094	
Model-3	Wavelet Network	661	Auto	63.975674	54.41291	78.35814	77.10695	70.51636	69.39832	77.01361	68.20868	67.5444	63.051176	57.489151	74.905073	68.49837	
Model-4	Wavelet Network	771	Auto	64.213647	54.54536	79.04259	77.40754	72.2192	66.33734	77.1989	68.6475	63.90738	64.513283	58.283279	71.532675	68.15406	
Model-5	Wavelet Network	991	Auto	62.6135	43.30331	77.22748	75.6257	74.40774	56.45595	77.20279	68.63672	67.53454	62.538324	46.853765	61.31958	64.47662	
Model-6	Wavelet Network	10101	Auto	64.189826	54.42962	79.01071	77.44483	74.64307	68.48736	76.94919	68.15462	67.25807	60.876296	57.933557	73.693817	68.58925	
Model-7	PW Linear	221	5	64.518683	60.91818	74.50966	72.59148	81.23641	74.58232	78.55623	67.10927	77.99925	80.440758	64.291634	80.209254	73.08026	
Model-8	PW Linear	221	15	65.465265	61.35894	76.87848	75.22954	78.81982	70.81137	73.6437	64.97967	79.21021	80.437083	64.866516	75.970947	72.30596	
Model-9	PW Linear	221	30	66.554876	61.35233	76.20894	74.42009	80.84486	70.60916	62.18123	68.57402	74.14108	80.609415	64.626615	75.986705	71.34244	
Model-10	PW Linear	221	35	65.501186	62.51182	65.02743	63.42409	65.37061	73.25765	76.84015	66.59422	69.2517	80.613741	66.274162	78.438215	69.42541	
Model-11	PW Linear	221	40	66.655318	61.39453	76.86818	75.4155	81.4371	71.61565	77.38261	68.58048	72.59079	71.530537	64.770792	76.451562	72.05775	
Model-12	PW Linear	331	20	66.468512	61.45352	74.79307	73.41426	81.43823	74.35005	65.0693	65.97612	73.27239	80.562007	64.810538	79.419165	71.75226	
Model-13	PW Linear	331	30	66.474173	61.36176	77.68439	75.59069	81.44731	74.17188	78.28924	68.56088	71.92415	55.949627	64.55939	79.18972	71.26693	
Model-14	**PW Linear**	**331**	**35**	**66.498185**	**61.08264**	**77.50198**	**76.11762**	**81.40994**	**74.31246**	**76.81875**	**68.12106**	**75.07129**	**80.611033**	**64.533338**	**79.037524**	**73.42632**	**1**
Model-15	PW Linear	441	10	65.434163	61.08594	68.10034	66.1904	81.28363	74.02363	78.03351	65.51662	69.15005	80.163736	64.557903	78.910204	71.03918	
Model-16	PW Linear	441	40	64.651333	61.53988	77.02542	75.40835	79.24104	74.51765	77.44785	56.05627	73.43979	77.38735	64.580073	80.112244	71.78394	
Model-17	PW Linear	661	10	65.333037	60.98543	67.75322	66.31243	81.15728	71.7384	78.33489	63.73076	73.21513	78.157011	64.602653	77.193492	70.71114	
Model-18	PW Linear	661	20	66.059604	61.1304	77.16825	75.43197	81.38586	69.9468	75.7414	63.19513	74.6471	80.378196	64.226489	74.736694	72.00399	
Model-19	PW Linear	661	25	65.834853	61.13109	75.33378	74.20689	81.44571	73.68157	77.45538	69.53326	66.90781	80.080957	64.306623	78.550796	72.37239	
Model-20	PW Linear	661	30	66.481685	61.29598	76.64784	75.46491	81.411	74.02152	66.23574	65.03209	73.61788	80.424198	64.890621	79.68114	72.10038	
Model-21	**PW Linear**	**661**	**35**	**66.563184**	**60.9506**	**77.91489**	**75.87486**	**81.40169**	**73.04657**	**73.59412**	**67.2474**	**79.35031**	**79.464866**	**64.037206**	**78.647508**	**73.17443**	**3**
Model-22	PW Linear	661	40	63.419496	61.54916	76.75408	74.72791	79.81077	72.89919	76.19534	54.8975	78.94191	79.151107	64.867092	77.943865	71.76312	
Model-23	**PW Linear**	**771**	**20**	**65.665932**	**61.04235**	**77.30037**	**75.34883**	**81.43638**	**74.871**	**77.62129**	**63.50878**	**78.48742**	**80.50004**	**64.146037**	**80.054224**	**73.33189**	**2**
Model-24	PW Linear	771	40	66.652238	61.54196	74.94987	72.91619	79.45557	72.62498	76.50698	60.15764	78.72435	80.425156	64.576646	78.062809	72.2162	
Model-25	PW Linear	881	5	65.538595	0.071205	74.91815	73.119	80.34256	57.83011	59.52152	62.82553	77.56236	75.536545	3.7809654	62.978562	57.83542	
Model-26	PW Linear	881	35	66.556292	0.043443	77.48776	75.5907	81.19246	73.34183	75.0383	65.88357	72.90857	80.414973	3.2678997	78.597773	62.52696	
Model-27	PW Linear	881	40	66.618606	0.042752	74.53111	72.71743	77.87474	73.77965	75.43525	62.2286	72.55722	80.440878	3.4193371	79.449935	61.59129	
Model-28	PW Linear	991	10	57.42774	60.86638	69.80207	68.48322	81.24556	74.09799	78.18818	65.16386	78.49586	73.797854	63.82937	79.316031	70.89284	
Model-29	PW Linear	991	15	65.470184	61.06902	71.36208	69.78969	78.72008	73.65649	57.94849	43.34331	78.61543	80.032016	64.047064	79.193997	68.60399	
Model-30	PW Linear	991	25	63.674729	61.01304	74.78927	72.88039	81.4004	74.85678	73.41996	67.3476	78.7699	80.323948	64.226496	80.312508	72.75125	
Model-31	PW Linear	991	30	66.403027	60.91521	77.16445	76.04252	81.01401	73.98915	74.58481	66.8144	79.06207	75.005125	64.125785	78.867352	72.83233	
Model-32	PW Linear	991	35	66.607369	61.09889	77.6684	75.99282	78.18772	74.63517	75.22197	67.39304	73.36043	72.374392	64.332982	80.146794	72.25166	
Model-33	PW Linear	10101	10	61.513772	58.75308	75.84906	74.29582	81.21585	74.25762	68.34354	60.25442	75.8655	79.899209	62.303112	78.991302	70.96186	
Model-34	PW Linear	10101	15	66.341265	61.11282	75.67254	74.49413	81.37765	72.81271	57.87307	47.56528	73.1647	79.943947	64.429038	78.256196	69.42028	
Model-35	PW Linear	10101	20	73.152549	61.33308	74.3462	72.78424	80.74121	70.8266	78.61816	63.11613	73.51904	80.343263	65.061164	76.039041	72.49006	
Model-36	PW Linear	10101	25	63.243522	61.32358	75.80122	73.84162	81.42536	72.77591	73.13266	69.28756	79.54234	80.50264	65.093794	77.833318	72.81863	
Model-37	PW Linear	10101	35	69.766321	61.14312	74.58141	72.84315	77.46734	74.51611	74.46272	66.23563	72.97236	80.448354	64.588813	79.990926	72.41802	
Model-38	PW Linear	10101	40	66.621908	60.76866	76.20752	74.34593	77.60126	73.48988	75.36611	53.57088	71.74905	80.568235	64.00267	78.916369	71.10071	
Model-39	Sigmoid	111	10	10.102732	53.60539	64.35182	62.29618	61.44056	63.95388	70.48556	64.61502	60.94737	58.96034	57.043376	69.11644	58.07656	
Model-40	Sigmoid	111	35	47.366726	53.39322	60.87006	59.42475	69.76839	62.70464	65.47888	61.14833	61.04875	58.567976	56.539944	68.246236	60.37983	
Model-41	Sigmoid	661	20	63.64256	55.84946	75.82596	73.95997	72.76283	69.25452	71.83376	67.38278	40.16251	77.901184	59.11363	74.790545	66.87331	
Model-42	Sigmoid	771	20	63.804398	55.40738	61.41465	59.42175	65.93499	69.17268	68.53399	66.72841	57.78493	62.145982	59.153094	74.708855	63.68426	
Model-43	Sigmoid	881	20	64.044686	0.043104	73.99727	72.44399	74.29749	66.83615	73.07761	65.41715	60.41048	60.229198	3.8043672	72.05721	57.22156	
Model-44	Sigmoid	881	40	64.004496	0.042871	65.14334	63.50178	78.16586	71.55496	73.62037	66.72452	63.98856	54.947117	3.0129832	76.438676	56.76213	

Table 3.13: Evaluation of top 42 models during *inverse* modelling scenario, on 12 test datasets

		Poles/Zeroes	N	Dataset-1	Dataset-2	Dataset-3	Dataset-4	Dataset-5	Dataset-6	Dataset-7	Dataset-8	Dataset-9	Dataset-10	Dataset-11	Dataset-12	Avg.	Rank
Model-1	Wavelet Network	221	Auto	52.270369	54.09648	15.0762	21.51615	60.82916	65.94095	64.35363	67.65893	28.095271	20.380223	52.187382	62.103741	47.04237	
Model-2	Wavelet Network	441	Auto	59.707716	47.27683	61.3064	66.29265	41.44702	85.04301	74.83352	67.07354	58.431179	66.935523	45.239761	72.25276	62.15333	
Model-3	Wavelet Network	661	Auto	52.546521	46.44429	57.44043	73.47435	65.02548	48.64617	82.57937	65.7115	76.4346.8	62.450056	44.006859	80.327149	62.9239	
Model-4	Wavelet Network	881	Auto	52.225788	47.34272	25.70687	49.69448	42.61207	66.02742	70.04298	69.77601	60.064167	30.781059	45.143379	67.802963	52.26833	
Model-5	Wavelet Network	10101	Auto	59.092885	47.96057	27.87192	28.72221	61.47108	41.49612	37.84812	34.48473	85.840068	32.921715	45.792379	35.170277	44.88934	
Model-6	PW Linear	221	:0	64.143191	84.77285	58.16806	16.10072	68.8562	83.21072	67.84226	59.24116	21.512954	63.615118	82.534478	65.213547	61.2676	
Model-7	PW Linear	221	:5	73.365015	53.49708	67.01847	13.721	72.87217	39.66756	70.01739	43.62373	26.257482	72.436827	50.919496	67.615791	54.251	
Model-8	PW Linear	331	20	73.647038	69.0263	51.91133	-5.42064	0.561663	70.24967	55.31366	35.47291	16.760601	56.910094	66.648159	53.589025	45.38915	
Model-9	PW Linear	661	:5	64.143191	84.77285	58.16806	16.10072	68.8562	83.21072	67.84226	59.24116	21.512954	63.755216	82.917778	65.740633	61.35515	
Model-10	PW Linear	661	25	73.365015	53.49708	67.01847	13.721	72.87217	39.66756	70.01739	43.62373	26.257482	71.823771	51.023837	68.05317	54.24506	
Model-11	PW Linear	661	40	68.608929	76.96302	58.85981	60.91812	75.01732	37.01237	82.60782	50.84216	76.646683	63.648128	74.429494	80.163944	67.14315	
Model-12	PW Linear	771	:5	69.00137	83.79717	61.37546	63.22546	68.90812	51.01953	71.16407	55.76477	81.272098	67.063709	81.603466	68.978041	68.59777	
Model-13	PW Linear	881	5	57.510428	69.8417	68.85298	46.35695	77.89645	15.30714	68.44382	55.45871	5.8845032	74.198052	67.504308	65.746403	56.08345	
Model-14	PW Linear	991	30	59.461356	77.48023	76.27635	23.23505	73.65728	44.60527	90.35108	55.62185	60.476168	81.766527	75.60456	88.454579	67.24919	
Model-15	PW Linear	10101	40	73.007262	80.45817	44.22314	19.29869	28.78853	72.85885	85.31595	62.25325	65.203789	49.088526	78.447182	82.836919	61.81502	
Model-16	Sigmoid	111	20	71.094235	59.11624	61.98102	56.20857	63.92046	64.18661	1.794668	55.74402	90.10608	67.01487	57.081382	-0.2732262	53.99791	
Model-17	Sigmoid	111	25	72.45574	61.39885	25.97343	36.06327	62.32847	33.59452	14.21631	-2.2915	12.48546	30.694089	58.988681	12.125225	34.83605	
Model-18	Sigmoid	111	30	46.489607	80.37664	71.76339	61.82519	68.47761	73.83274	45.87036	47.96083	4.4009389	77.042849	77.67109	43.31002	58.25177	
Model-19	Sigmoid	111	35	46.465669	68.51283	40.07437	63.65879	71.62185	71.76943	50.58264	42.58259	85.141717	45.716488	66.116404	48.61974	58.40521	
Model-20	Sigmoid	111	40	58.619483	71.79975	71.11439	51.31206	82.91524	27.96347	45.31018	23.02592	22.814648	76.279709	69.8293	42.922016	53.65885	
Model-21	Sigmoid	221	:5	71.868318	69.37429	76.68799	49.91729	78.4106	77.85132	52.17135	34.24093	79.627975	81.757256	66.877485	49.687631	65.70604	
Model-22	Sigmoid	221	20	0.1327234	70.63067	75.99388	57.53065	46.17463	68.70481	19.59836	52.40756	13.540172	81.226945	68.866032	17.589142	47.69963	
Model-23	Sigmoid	221	25	84.888191	72.68854	9.575326	57.71093	65.9151	63.37359	86.27134	67.72202	34.34873	14.384607	69.984101	83.878614	59.22842	
Model-24	Sigmoid	221	35	77.726751	3.132193	69.64356	36.13472	68.67071	70.96362	67.64922	66.40881	-1.7263336	75.127344	1.2245874	65.444366	50.0333	
Model-25	Sigmoid	441	:0	82.843415	71.68281	47.26168	-0.6512	68.33717	35.20628	81.6803	1.95048	19.949783	52.370076	69.159934	79.39241	50.76526	
Model-26	Sigmoid	441	20	76.608216	76.87065	13.23857	0.580593	1.932081	39.75299	24.15369	38.78305	0.5099083	18.10641	74.40457	22.232485	32.26443	
Model-27	**Sigmoid**	**441**	**30**	**73.067983**	**78.79257**	**75.00144**	**62.71883**	**57.46959**	**80.55445**	**89.17969**	**69.05776**	**55.389678**	**80.106502**	**76.522223**	**87.002558**	**73.73861**	**1**
Model-28	Sigmoid	441	40	71.130727	67.33985	76.44137	61.40768	64.16414	89.0189	46.12981	66.0139	63.617631	81.433883	65.202512	44.32765	66.35234	
Model-29	Sigmoid	551	5	84.902632	64.86231	73.74381	28.02942	61.51247	9.870125	78.94628	35.6867	67.086595	78.763055	62.736021	76.365553	60.20875	
Model-30	Sigmoid	551	:0	76.63739	48.21014	63.85588	52.78287	69.92653	72.16071	64.96898	53.89469	63.67867	68.587417	46.457146	63.150485	62.02591	
Model-31	Sigmoid	551	:5	85.166745	68.07949	63.40044	33.6515	76.57389	68.67346	29.00736	54.32685	55.594479	68.359314	65.835544	26.563639	57.93606	
Model-32	Sigmoid	551	25	81.821888	63.77549	64.31489	40.57303	81.80407	85.18059	38.14773	55.08197	55.151262	69.496787	61.677308	36.371415	61.11637	
Model-33	Sigmoid	661	35	62.919405	50.90716	64.65529	65.21047	79.77491	50.69054	56.00054	70.18263	83.108069	69.516132	48.804373	53.345581	62.92626	
Model-34	Sigmoid	661	40	59.963293	66.66439	62.70503	40.84065	89.50977	28.788	60.60499	36.50057	65.336051	67.80084	64.561972	58.435973	58.47596	
Model-35	**Sigmoid**	**991**	**5**	**76.69153**	**65.40126**	**54.03012**	**74.48167**	**83.1037**	**64.94354**	**67.02109**	**70.75109**	**85.055793**	**59.202737**	**63.055312**	**64.313213**	**69.00426**	**2**
Model-36	**Sigmoid**	**991**	**20**	**70.360885**	**66.00945**	**51.35712**	**63.32036**	**93.91127**	**80.11101**	**71.87545**	**50.56712**	**81.519527**	**56.603967**	**63.98402**	**69.295468**	**68.24297**	**3**
Model-37	Sigmoid	991	30	64.620527	65.92329	63.38985	60.35604	78.37827	62.50759	24.67926	64.4701	73.556522	68.661247	63.507963	22.815297	59.4055	
Model-38	Sigmoid	991	35	74.572802	74.56945	53.62297	60.40517	78.5192	63.18789	29.22891	66.05168	63.602656	58.675346	72.228241	27.147575	60.15099	
Model-39	Sigmoid	10101	20	85.160667	79.45991	52.76458	67.9628	59.67901	67.83562	25.21471	50.19715	47.453743	57.699538	77.496561	23.324597	57.85407	
Model-40	Sigmoid	10101	30	77.711891	76.28774	54.09329	80.49312	69.3678	10.97139	51.66576	79.6805	88.641706	58.962418	74.074766	49.644818	64.2996	
Model-41	Sigmoid	10101	35	7.4086222	47.6157	26.30267	-5.21961	63.03425	67.42585	6.7642.02	2.380882	17.962321	31.583126	45.510747	4.1839872	26.24606	
Model-42	Poly 1D	5	:0	84.954681	71.58559	64.98308	80.36929	41.21654	3.156869	43.98953	53.41513	-11.61.1649	70.031716	69.267677	41.938468	51.10808	

Table 3.14: Top three models found after evaluating 12 held out test datasets

Forward modelling scenario		Inverse modelling scenario	
Models	**Avg. fit %**	**Models**	**Avg. fit %**
PL 331/N=35	73.42	Sigmoid 441/N=30	73.73
PL 771/N=20	73.33	Sigmoid 991/N=5	69.00
PL 661/N=35	73.17	Sigmoid 991/N=20	68.24

Thus, it can be said that the top three models reported in Table 3.14 for both forward and inverse modelling scenarios provide the best and reliable results for plant electrical signal response to light. Hence a mathematical relationship has been established between stimuli and plant electrical response in the form of modelling.

3.9 Summary

By using the electrical response data of 20 plants to incident light stimulus (three different species, *Laurus nobilis*, *Zamioculcas zamiifolia*, and *Cucumis sativus*), the rising and falling edges of the light were predicted within a forward and inverse modelling dynamical system framework. The best prediction for detecting the instants of turning on/off and peak intensity of light was obtained by a set of NLHW models over linear and NLARX models. NLARX produced negative percentage fits consistently and hence have been ignored. The top three NLHW model settings, which showed reasonable prediction ability for the main dataset, were tried on 19 other datasets. The results showed the top three model settings produced 39-43% fit during inverse modelling and 55-59% fit for forward modelling scenarios. Only a few datasets were found to have fit of around 60% and above, and the models performed poorly on the remaining datasets. This was as expected as the models were found from one main dataset and tried on 19 others.

The electrical response of these 19 other plants were different from the response of the Bay leaf plant (main dataset) which was used to train the models. Hence new experiments with *Chrysanthemum* plants were conducted with light pulses as stimuli, and 30 new datasets were added. This new data showed consistent and repeatable trends. Thus combining 20 old and 30

new datasets, provided 50 datasets for exploration of the models and independently validating them. Of the 50, 12 datasets were set aside for validating the top three models resulting from the best fits on the 38 datasets used.

It was found that the average fit percentage increased drastically to around 73% (models found from 38 datasets and tested on 12 datasets). The significance of using more data from multiple species of plant overrules any species bias and so it can be said that the final models are generic. As with any data analysis exploration, the more data are available, the better the results and the analysis. Although the results presented in this chapter are based on 50 datasets, there is the possibility of adding more species for further experimental data collection and exploration of the models. The addition of independent test data makes the models much more generic. Also as a first study, the goal of establishing a mathematical relationship has been achieved.

This method of system modelling, based on the system identification approach, can be further explored by using plant electrical response data to determine a variety of stimuli such as introduction of gas or chemicals to the soil, thus paving the way towards conceptualizing plant-based environmental biosensors. Although the system identification technique was employed as a first step to determine whether there was enough information contained within the plant electrical signals about the applied light stimulus within the model constraints (e.g. ignoring temperature and humidity, limited variation of parameters such as input-output units and poles-zeroes), in future the model constraints could be changed such as expanding the number of poles and zeros in separate orders (keeping in mind that poles $\geq$ zeros for realizing proper transfer functions). In fact, all the constraints mentioned in Table 3.1 could be changed. This work will provide a solid foundation for future work on plant electrical signals using *system identification* techniques to explore the on-off times and amplitudes of other stimulus types.

The black-box modelling may not be suitable if the goal is to interpret the stimulus *type* (from multiple categories) and that too from a short segment of the electrical signal response of the plants. The necessity for analysing a short segment of the electrical signal response is for conceptualising an environmental sensor (based on the plant electrical signal response) which will have a very short decision time to predict which stimuli affected the plant. Not knowing when the stimuli affect the plants, any segment from the response should be able to give an

indication of the stimuli. However, as a first exploration, the goal of showing the possibility of establishing a forward as well as an inverse relationship between the light as a basic stimulus and the plant electrical signal response in terms of fit and on/off times has been achieved. Now the focus shifts to more complex stimuli and analyse whether those stimuli can successfully be mapped to the plant electrical signals.

4 Exploring Strategies for Classification of External Stimuli Using Statistical Features of the Plant Electrical Response

4.1 Introduction

To achieve the goal of using plants as biosensors, capable of detecting which external stimulus is affecting them, the requirement is to develop a methodology by which it is possible to detect multiple environmental stimuli from the electrical response of the plants. One approach seemed to be a supervised learning methodology where certain features, extracted from the raw plant electrical signals, may be used to build classification models.

The work so far worked has been on light, essential for the optimal growth of plants. This stimulus is important especially in the case of precision agriculture and greenhouse irrigation. From the ability of a plant to sense its natural environment, the exploration should address whether a unique response is being provided by the plant that can be extracted by standard measurement techniques. Such a response should be processed and give meaningful results, providing clues about the stimulus to which the plant was exposed. This requirement is the focus of this chapter, where plants were exposed to more complex stimulus, under laboratory conditions, to extract their electrical signal responses. These responses were then analysed to see if the stimulus can be identified from them.

System Identification techniques enabled the time of application of the stimulus as well as its amplitude to be found. In this chapter, methods are employed which help identify the type of stimulus. This provides two options for extracting information from plant electrical signal response: finding the time and amplitude of the stimulus (established using basic light pulse as stimulus), and identifying the type of stimulus.

This chapter extends the experiments to four different stimuli: Sodium Chloride (NaCl in two different quantities – 5 ml and 10 ml), Sulphuric Acid (H_2SO_4), and Ozone (O_3), which were applied to different plants in a laboratory setting to generate the plant electrical signal response. It also explores whether the short duration of the electrical signal could help detect the type of external stimulus. To achieve this, 11 statistical features were explored, which were computed from the raw, non-stationary, signal time series, to predict the stimulus applied by using classification algorithms in six classification settings. To achieve this, four

different *discriminant analysis* based algorithms and one *minimum distance* based algorithm were explored, which successfully established that there is enough information in the raw electrical signal to detect the type of stimulus, i.e. classify it. The aim was also to identify which features and which classification algorithm provided the best results for the available data. As an initial exploration using the classification methodology, *univariate* and *bivariate* feature explorations were carried out.

Although there have been some recent attempts on signal processing, feature extraction and statistical analysis using plant electrical responses [110], [112]–[116], [141], there has been no attempt to associate features extracted with different external stimuli. As far as is known, no work has been carried out to detect the type of external stimulus, and thus this exploration is the first of its kind.

The methodology of classification of stimulus type (also referred to as *class*) from the plant electrical signal was based on three fundamental parameters:

- Plant electrical signal pre-processing
- Feature extraction
- Classification

First, a trained model was obtained using a set of data (i.e. a set of feature values) and then the model was tested on another set of data (i.e. a separate set of feature values) to validate it, as shown in Figure 4.1.

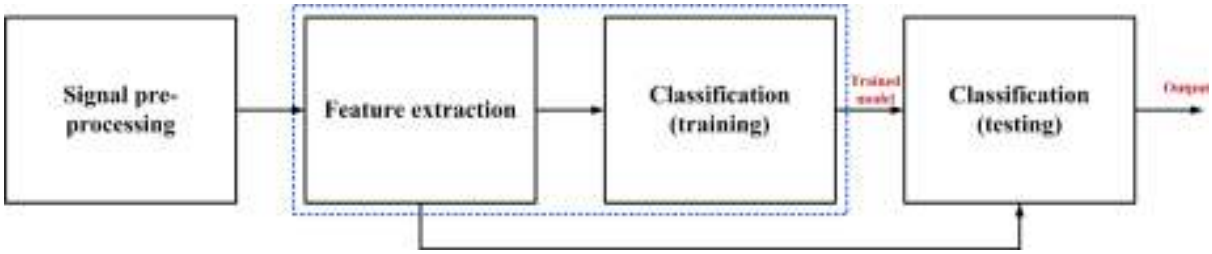


Figure 4.1: Methodology for classification of stimulus from plant electrical signal response

The rest of the chapter is organized as follows. Section 4.2 provides details of the experimental data collection. Section 4.3 discusses the signal processing steps carried out on the data collected. Section 4.4 provides information about the features extracted. The method for ranking of the features (in binary classification settings) is discussed in Section 4.5. The methodology behind the classifiers is discussed in Section 4.6, while Section 4.7 discusses the results.

4.2 Experimental data collection

Four sets of experiments were conducted with H_2SO_4, NaCl (5 ml) and NaCl (10 ml) each as the stimulus, and eight sets of experiments were conducted with O_3 as the stimulus. Table 4.1 gives the details of the experiments conducted and resulting blocks of data collected. The data blocks were obtained after dividing the entire signal into blocks of 1000 samples.

Table 4.1: Different stimulus, plants species and number of data-blocks obtained

Stimulus	**Plant species used**	**Concentration and application**	**Number of data-blocks (each block containing 1000 samples)**
Ozone (O_3)	Tomato/Cucumber	16 ppm for a minute, every 2 hours	1881
Sulphuric acid (H_2SO_4)	Tomato	5 ml of H_2SO_4 0.05 M in the soil once	496
Sodium Chloride (NaCl) – 5 ml	Tomato	5 ml of NaCl 3 M in the soil once	812
Sodium Chloride (NaCl) – 10 ml	Tomato	10 ml of NaCl 3 M in the soil once	612

For experiments with NaCl and H_2SO_4, four different tomato plants (similar age, growing conditions and heights) were used, each plant being exposed to the stimulus only once. Thus for 12 experiments (i.e. four experiments for each of the three stimuli), 12 tomato plants were used. For O_3 as the stimulus, six cucumber plants and two tomato plants were used for 8 experiments, each plant being subjected to only one experiment (but involving multiple applications of the stimulus).

Three stainless steel needle electrodes were used for each plant, one at the base for reference, one in the middle and the other at the top of the stem as shown in Figure 4.2. The electrodes from Bionen s.a.s. were 0.35 mm in diameter and 15 mm long, similar to those used in EMG, and were inserted around 5-7 mm into the plant stem so that the sensitive active part of the electrodes (2 mm) was in contact with the plant cells [148]. The electrodes were connected to the amplifier-Data Acquisition (DAQ) system in the same way as previously (see Section 0) [148]. The plants were then enclosed in a plastic transparent box with proper openings to allow the presence of cables and inlet/outlet tubes, and exposed to artificial light conditions (LED lights responding to plant's photosynthetic needs, mimicking a day/night cycle of 12 hours each). Each experiment was conducted in a dark room to avoid any external light

interference. The whole setup was then placed inside a Faraday cage to limit the effect of electromagnetic interference, as shown in Figure 4.2.

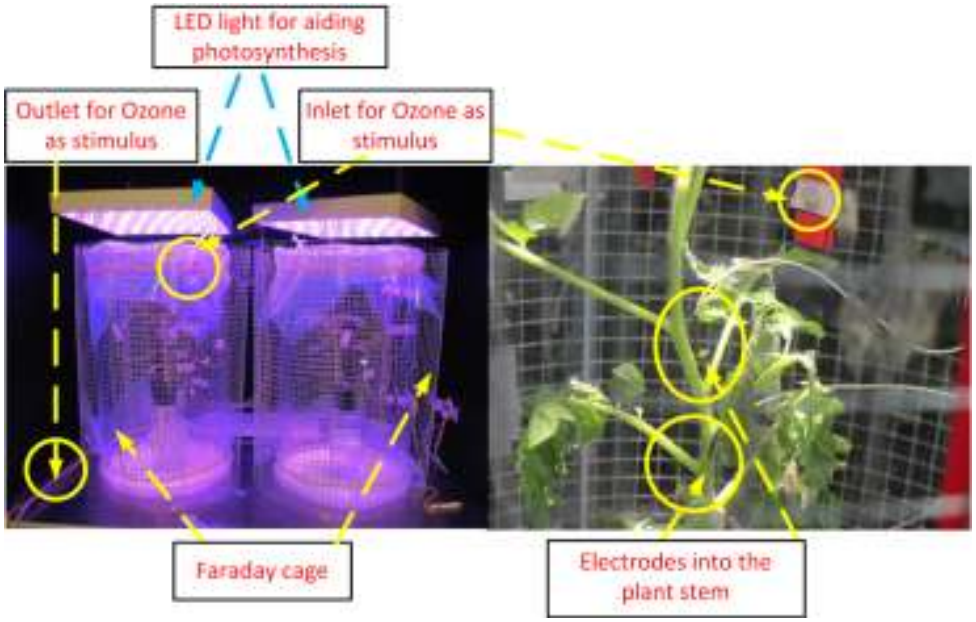


Figure 4.2: Experimental setup showing a tomato plant inside a plastic transparent box, kept inside a Faraday cage. The placement of the electrodes on the stem is also shown.

After the insertion of the electrodes into the plant, a waiting period of about 45 minutes was allowed for the plant to recover before starting the stimulation. Electrical signals acquired by the electrodes were provided as input to a 2-channel high impedance (10^{15} Ω) electrometer (DUO 773, WPI, USA) while data recording was carried out through the 4-Channel DAQ (LabTrax, WPI) and its dedicated software, LabScribe (WPI) [149].

The sampling frequency was set to 10 samples a second for all the recordings (as it was deemed sufficient to capture the electrophysiological response of the plants appropriately). For the treatments with liquids, i.e. H_2SO_4 (5 ml, 0.05 M) or NaCl (5 or 10 ml of 3 M solution), a syringe was placed outside the Faraday cage and connected to a silicone tube inserted into the plant soil. This was used to inject the solution, as shown in Figure 4.3(a). O_3, produced by a commercial ozone generator (mod. STERIL, OZONIS, Italy) [150], was injected into the box through a silicone tube (1 minute spray every 2 hours, 16 ppm), while a second outlet tube withdrew the O_3 from the box into the chemical hood, as shown in Figure 4.3(b). The concentration of O_3 inside the box was monitored using a suitable sensor.

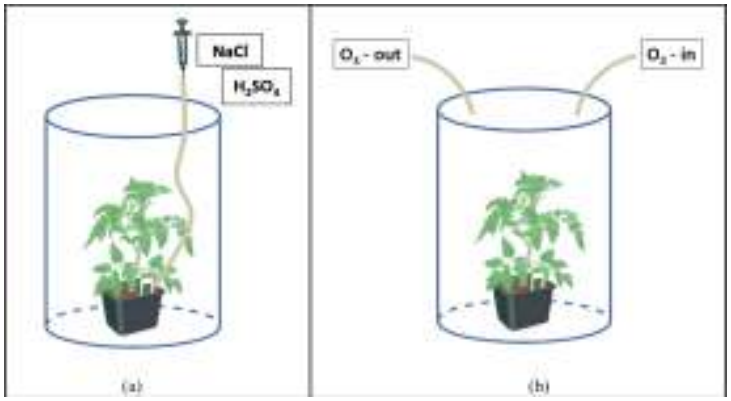


Figure 4.3: Tube system for introducing pollutants into the box

4.3 Pre-processing the entire signal

The raw plant electrical signal response, obtained from various experiments, were processed in order to

- Increase the number of data to train the classifiers with, for obtaining a robust classification model [151]
- Transform the increased data from non-stationary to stationary signals, so that each data point can be considered as independent and identically distributed (IID)

These pre-processing steps are described briefly below.

4.3.1 Increasing the number of data through segmentation - Resampling method

Each dataset was obtained after one (for H_2SO_4, NaCl 5 ml and NaCl 10 ml) or multiple (for O_3) applications of that particular stimulus. This is illustrated in Figure 4.4 where the application of the stimulus is marked by a vertical dotted line with the post-stimulus part of the time series on the right side and the background or pre-stimulus part indicated on the left side of this line. Multiple applications of the O_3 stimulus is shown by multiple markers.

In general, there were sudden spikes in the signal after the application of H_2SO_4 and O_3 as stimuli. However, for the NaCl 5 ml and 10 ml stimuli, the changes in the electrical signal response were relatively slow. These characteristics were observed for most of the datasets used in this work. Thereafter, for each experiment, the data was divided into a post-stimulus part and a pre-stimulus part for each experiment. Each of the pre- and post-stimulus parts were appropriately labelled to identify the experiment and channel number (each experiment

had two channels, see Section 4.2). Where multiple O_3 stimuli were applied, the data was divided so that the signal duration between consecutive applications of the stimulus is considered a separate post-stimulus response (a possibility existed that subsequent response due to O_3 may have included signatures of the previous application of the O_3. This was ignored in this work). The result was several post- and pre-stimuli datasets for all four stimuli. Next, each of these datasets was segmented into blocks of fixed window length of 1000 samples (100 seconds), which is shown in Figure 4.4.

The reason for exploring the possibility of classifying stimuli from a small segment of plant's electrical signal response, was an environmental sensor should be able to capture a small segment of data from any part of an incoming stream of signals and be able to classify the stimuli successfully.

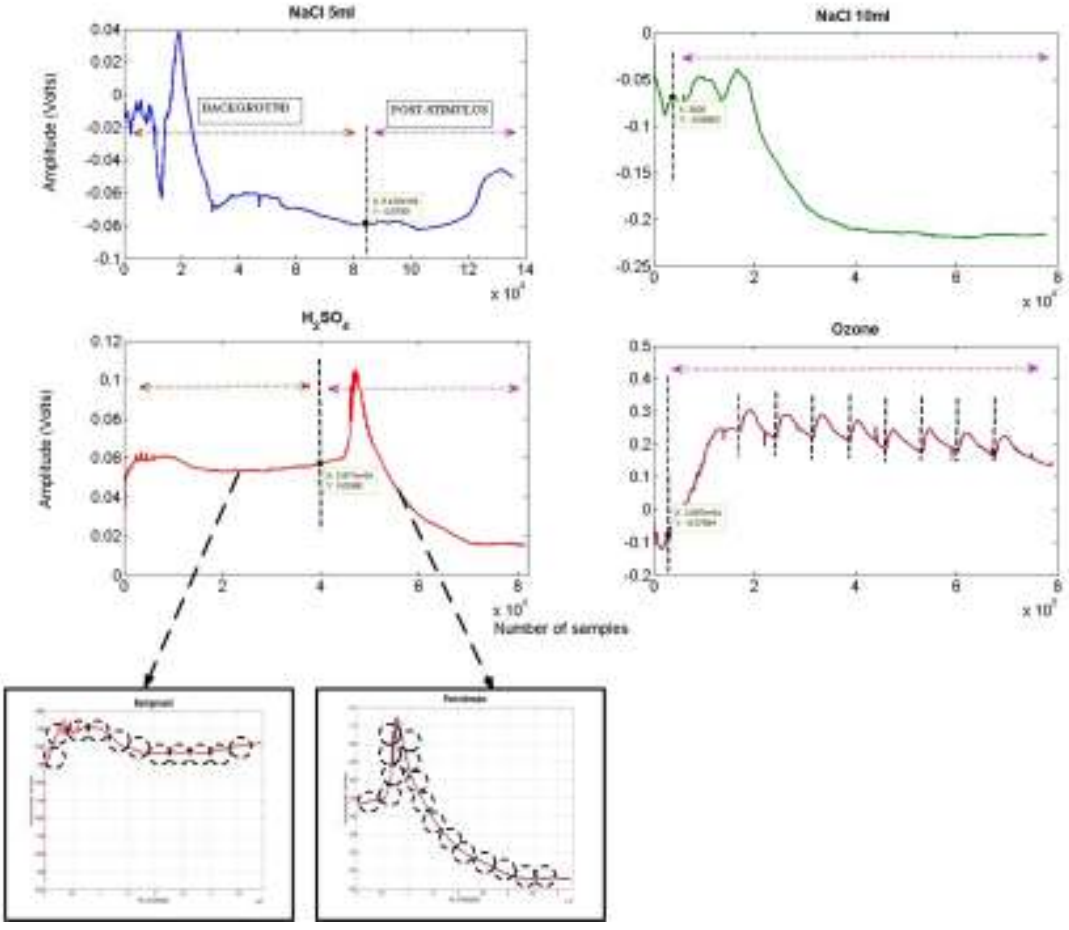


Figure 4.4: (Top) The vertical dotted lines mark the application time of the four stimuli. (Bottom) Separating the plant electrical signal into background and post-stimulus parts and then dividing them into smaller blocks of 1000 samples (dashed circles).

The segmentation methodology, which is a part of resampling methods proposed in standard machine learning literature [135] has been used for some time [152]–[159] to analyse various time-series. This method, especially used to analyse ECG and EEG signals, are an important topic for processing temporal information contained within a long, and often high-dimensioned, series. Several segmentation procedures have been developed, which compare the characteristics of segments of time with the whole time series on which the segmentation is carried out. The resulting segmented series are then used for analysis based on statistical methods, clustering, decision tree, etc. Depending on the purpose, the time series can be analysed at the level of individual segments, the number of segments, or as the entire series. By creating internally homogeneous segments of the time series, it is also possible to identify significant changes in the behaviour of the series, the change-points, and other temporal information [160].

Usually, the process of segmenting a time series is as follows:

- Some known data behaviour is observed that defines precisely the time duration of each segment.
- Thereafter, a detailed analysis is carried out on these segments for classification, assessment and/or prediction of behaviour of the observations.

The same segmentation technique is applied in this analysis due to shortage of data available for carrying out classification. However, as classification using plant electrical signals has never been attempted before, it was not known which part of the plant electrical signal response contained the embedded information about the applied stimuli. Hence the entire response was segmented into blocks of fixed window length. Through this, we adopted the *hypothesis* that there is enough information about the applied external stimulus within *any* block of 1000 samples (or 100 seconds) from the entire time series data.

Although segmentation method is widely used as a part of resampling method, the drawback of this resampling approach is that it considers the data in each segment as being independent and identically distributed (IID) and considers correlation only within the segment [135]. More specifically, dividing long time series signals (with fewer independent realisations) into smaller segments ignores the correlation lost between successive segments. This is often a valid assumption for practical time series classification problems if the goal is to detect the cause of the time series by looking at a small portion of the signal [135]. In the next section, the steps taken to remove the correlation between the segments is described.

4.3.2 Transforming non-stationary data to IID sample

The plant electrical signal is a typical non-stationary biological signal [47] which needs to be transformed into stationary signal so that any segments from this signal can be considered as IID samples for training the classifier [135]. This is carried out so that there are more number of data to carry classification. Two approaches could be taken for pre-processing the plant electrical signals to transform them from non-stationary to stationary signals.

These steps are given below [135]:

- When using raw signals – use segmentation method to divide the entire signal into segments of fixed window lengths. Thereafter use these segments from both pre- and post-stimulus parts of the signals, for feature extraction. The mean of each feature is then calculated from the pre-stimuli parts and this mean was subtracted from the corresponding feature of the post-stimulus parts. This ensured that the resultant pre-stimulus mean subtracted features of the post-stimulus parts, reflected only the change in the statistical parameters due the introduction of the stimulus. Hence any bias due to the difference in pre-stimulus starting points was diminished (more details about the background subtraction method is provided in 4.7.2).

- Explore optimum filtering criteria for the plant electrical signals (see Chapter 5) and use this to filter the entire raw electrical signal. Thereafter, use the segmentation method on the filtered signals to get individual blocks which could be used extracting features for training the classifiers.

By using these two methods, each of the epochs/segments/blocks were assumed to be IID samples belonging to the same class as the entire time series. In this chapter, the first method was employed as a first exploration to validate the hypothesis that a small segment of raw plant electrical signal contains sufficient information about the type of the stimulus to provide a good classification accuracy.

A successful classification of the stimuli type from the features of such a small signal block would enable a fast decision time if and when implemented as a sensor in a natural environment, because a smaller buffer-size is needed compared with the whole length of the signal acquisition, making it easier for possible online implementation. Since this is the first exploration of its kind, the choice of 1000 samples was made, to enable sufficiently good

classification accuracy to be obtained. However, there is scope for further exploration of an optimum window length to classify the stimuli which was explored in Chapter 5. The classifier was trained using only the blocks of samples belonging to the post-stimulus part of the plant signal.

The stimuli-induced plant signals contain both deterministic and random dynamics, i.e. local and global variations in amplitudes and values of different statistical measures (for different data segment lengths) [47], [54], [82], [108], [161].

4.4 Extracting statistical features from segmented time series

11 features were explored, which are predominantly used in the analysis of other biological signals [162]. Different descriptive statistical features were calculated, such as *mean* (μ), *variance* (σ^2), *skewness* (γ), *kurtosis* (β) as given in (4.1) and *Interquartile range* ($IQR = Q_3 - Q_1$, i.e. the difference between the 1st and 3rd Quartile).

$$\mu = E\left[x_i\right], \sigma^2 = E\left[x_i - \mu\right]^2, \gamma = E\left[\left(x_i - \mu\right)/\sigma\right]^3, \beta = E\left[\left(x_i - \mu\right)/\sigma\right]^4 \tag{4.1}$$

In the calculation of these four basic moments, x_i is the segmented raw electrical signals (each segment containing 1000 samples) and E[.] is the mathematical expectation operator. Apart from these five, the remaining six features taken were – *Hjorth mobility*, *Hjorth complexity*, *detrended fluctuation analysis* (*DFA*), *Hurst exponent*, *wavelet packet entropy*, and *average spectral power*, which are briefly described below.

4.4.1 Hjorth's parameters

The Hjorth mobility and complexity, described in [163]–[165], quantify a signal from its mean slope and curvature by using the variances of the deflection of the curve and the variances of their first and second derivatives.

As shown in Figure 4.5, let the signal (depicted by measurement of amplitudes at discrete time intervals) be a with a_n being the amplitude value at time t_n. The measure of the complexity of the signal is based on the second moments in the time domain of the signal, and the signal's first and second derivatives.

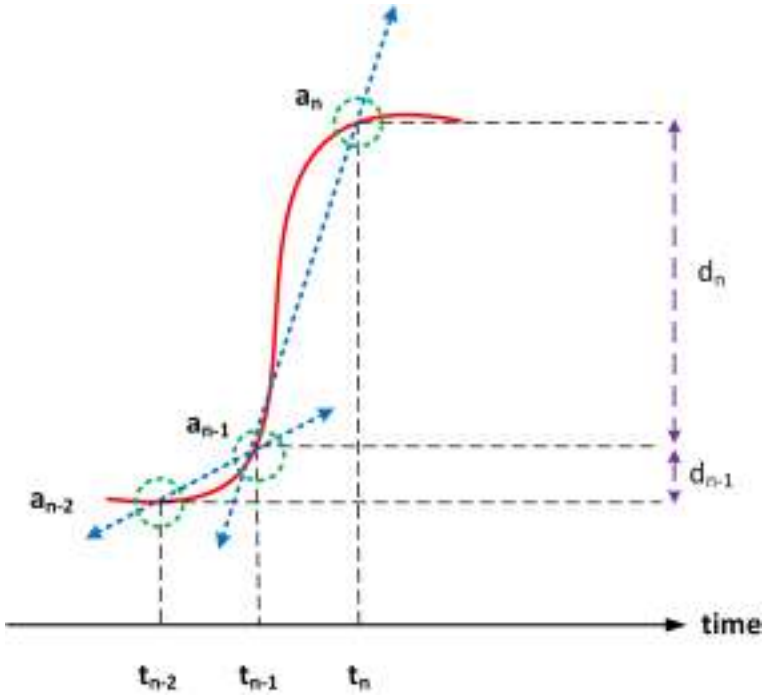


Figure 4.5: Identifying variations in slope and curvature of a signal [163]

The finite differences of the signal or time derivatives can be seen in (4.2).

$$d_n = d'_n = a_n - a_{n-1}, \text{ where } n = 1,2,\cdots,(N), \text{ and}$$

$$d''_n = d_n - d_{n-1}, \text{ where } n = 1,2,\cdots,(N) \tag{4.2}$$

The *variances* are then computed as following [163].

$$\left\{ \begin{array}{c} \sigma_a^2 = \frac{1}{N}\sum_{n=1}^{N} a_n^2 \\ \sigma_d^2 = \frac{1}{(N)}\sum_{n=1}^{N} (d_n)^2 \\ \sigma_{dd}^2 = \frac{1}{(N)}\sum_{n=1}^{N} (d_n - d_{n-1})^2 \end{array} \right\} \tag{4.3}$$

These variances are used to calculate the Hjorth mobility (m_{H}) and the Hjorth complexity (c_{H}) [163]–[165] as shown in (4.4).

$$\left\{\begin{array}{c} m_H = \dfrac{\sigma_d}{\sigma_a} \\ c_H = \sqrt{\dfrac{\sigma_{dd}^2}{\sigma_d^2} - \dfrac{\sigma_d^2}{\sigma_a^2}} \end{array}\right\} \tag{4.4}$$

4.4.2 Detrended fluctuation analysis (DFA)

DFA has been introduced for identifying long range correlations in non-stationary time series data. By using a scaling exponent (α), one can describe the significant autocorrelation properties of signals with a provision of capturing the non-stationary behaviour as well [166], [167]. The different values of α represent certain auto-correlation properties of the signal [166], [167]. For a value of less than 0.5, the signal is described as anti-correlated. A value of exactly 0.5 indicates an uncorrelated (white noise) signal, whereas a value greater than 0.5 indicates positive autocorrelation in the signal. When α = 1, the signal is defined to be $1/f$ (pink) noise and a value of 1.5 indicates the signal to be random walk or Brownian noise [166], [167].

4.4.3 Hurst exponent

The Hurst exponent (*H*) is a dimensionless estimator similar to DFA, and is used as a measure of the long term *memory* of a time series data x_i [168], [169]. The value of the Hurst exponent lies between 0 and 1, with a value between 0 and 0.5 indicating *anti-persistent* behaviour. This denotes that a decrease in the value of an element will be followed by an increase and vice versa. This characteristic is also known as *mean reversion*, which is explained as the tendency of future values to return to the longer term mean value. The mean reversion phenomenon gets stronger for a series with *exponent* value closer to zero [168], [169]. When the value is close to 0.5, a random walk (e.g. a Brownian time series) is indicated. In such a time series, there is no correlation between any element and predictability of future elements is difficult [168], [169]. Lastly, when the value of the exponent is between 0.5 and 1, the time series exhibits persistent behaviour. This means the series has a trend or there is a significant autocorrelation in the signal. The closer the exponent value gets towards unity, the stronger the trend is for the time series [168], [169].

4.4.4 Wavelet entropy

A time series may be represented in frequency and/or time-frequency domains by decomposing the signal in terms of basis functions, such as harmonic functions (as in Fourier analysis), or wavelet basis functions (with consideration for non-stationary behaviour). Given such a decomposition, it is possible to consider the distribution of the expansion coefficients in this basis pace. Quantification of the degree of variability of the signal could be done using the entropy measure, where high values indicate less ordered distributions. The wavelet packet transform based entropy (Wentropy or WE) measures the degree of disorder (or order) in a signal [170]–[172]. A very ordered underlying process of a dynamical system may be visualized as a periodic single frequency signal (with a narrow band spectrum). The wavelet transformation of such a signal will now be resolved in one unique level with value near 1, all other relative wavelet energies being minimal (almost zero) [170]–[172].

On the other hand, a disordered system represented by a random signal will portray significant wavelet energies on all frequency bands. The wavelet (Shannon) entropy gives an estimate of the measure of information of the probability distributions. This is calculated by converting the squared absolute values of the wavelet coefficients s_i of the i^{th} wavelet decomposition level, as shown in (4.5).

$$WE = -\sum_i s_i^2 \log\left(s_i^2\right) \tag{4.5}$$

4.4.5 Average spectral power

The average spectral power ($\overline{P}$) is the measure of the variance of signal power, distributed across various frequencies [173]. It is given by the integral of the power spectral density (PSD) curve $\left|X\left(e^{j\omega}\right)\right|^2$ of the signal $x(t)$ within a chosen frequency band of interest (bounded by the low and high frequency ω_l, ω_h respectively) as shown in (4.6)

$$\overline{P} = \int_{\omega_l}^{\omega_h} \left|X\left(e^{j\omega}\right)\right|^2 d\omega \tag{4.6}$$

The features *mean*, *variance*, *skewness*, *kurtosis*, *IQR*, *Hjorth's parameters*, *DFA* and *Hurst exponent*, were computed using the MATS toolbox [162], whereas *wavelet-entropy* and *average spectral power* were computed using scripts written in MATLAB.

4.4.6 Normalization of features

The 11 features which were extracted from the electrical signal response of the plants (blocks of 1000 samples) were normalized using the formula in (4.7) so that the feature values lie within a minimum {0} and a maximum {1} in order to avoid any unnecessary emphasis of some of the features on the classifier weights due to their larger magnitude than the others.

$$\tilde{x} = \left(x - x_{\min}\right) / \left(x_{\max} - x_{\min}\right) \tag{4.7}$$

Here, $\tilde{x}$is the normalized feature value, $x$is the original feature value, $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the feature vector respectively.

4.5 Ranking of features – Fisher ratio

Since there were 11 features altogether, it was essential to find out which features gave the best separation between two different stimuli. For this reason, a criterion called *Fisher discriminant ratio* (*FDR*), which is the ratio between the means and the standard deviation of the features of the two classes *A* and *B*, were used. This is given by the following equation

$$FDR = \frac{\left(\mu_A - \mu_B\right)^2}{\left(\sigma_A^{\,2} + \sigma_B^{\,2}\right)} \tag{4.8}$$

Based on FDR, higher ranking will be assigned to those features which have higher difference in the mean values and small standard deviation, implying compact distantly located clusters. The higher the value of the Fisher ratio for a feature, the better that feature is in quantifying the separation between the two classes involved.

The mean and variances in this case should not be confused with those of the raw signal in (4.1). Because multiple stimuli are used in this work, the FDR based feature ranking is applied to each of the stimulus pairs.

4.6 Theoretical background of discriminant analysis based classification techniques

The next step was to use the extracted features to actually distinguish between the stimuli, using some form of decision-making algorithm. Such an algorithm should be able to find a recognizable pattern amongst the features for a particular stimulus and therefore be able to differentiate between different patterns, i.e. for different stimulus. This problem of identifying

the similar and dissimilar patterns is a part of pattern recognition/machine learning or classification theory [174]. In a typical m-class classifier problem, the need is to identify the class of an unknown sample that could be achieved by a well-trained classifier.

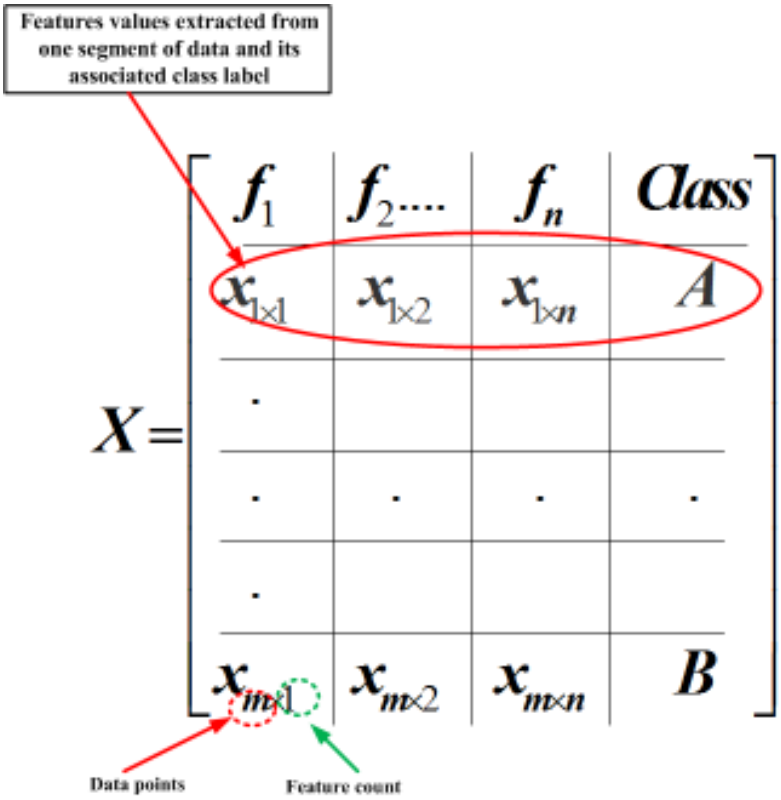


Figure 4.6: Example of a training dataset containing feature values for two classes: A and B

A classifier is basically an algorithm, which is developed using certain mathematical rules, and used on training data sets containing features extracted from different classes (here, the features were extracted from segmented signals) as shown in Figure 4.6. Although the class labels in Figure 4.6 are shown as *A* and *B* for there, they are usually numeric. The feature values are also referred to as *predictors* or *independent variables*. The training data sets provide an insight into how future data from a particular class might look. Usually, if the training data for a particular class is widely varied (i.e. there are enough data points from each class), the better trained and generalized will be the classifier algorithm for observations for that particular class [151], [175]. In principle, a classification approach consists of two steps as follows.

1) *Training*, where the class labels (different classes, identified by different class labels, here representing different stimuli) of a certain portion of the known (*training*)

datasets is used to create a model which maps the class labels to the feature vectors belonging to that class.

2) *Testing*, where the trained model thus obtained is used to predict the class labels for a group of unknown feature vectors (*test dataset*). The test dataset is not used for creating the model and is set aside separately for testing the already created model. However, the class labels in the test dataset are known to the user so that the model prediction can be validated against the known labels for ascertaining the model accuracy. The testing could either be done using different *cross-validation* schemes or on an independent dataset or a combination of both. Once it is found that the testing phase is successful using the trained model, the algorithm (and the trained model) can be used to identify which class an unknown data belongs to.

4.6.1 Choice of classifiers

A variety of algorithms can be used to determine how accurately classification identifies the four different stimuli. Four *discriminant analysis* based and *minimum distance* based classification methodologies were chosen for their simplicity. In cases where the class distinction is easily achievable, discriminant analysis classifiers such as *Linear Discriminant Analysis (LDA)* could be effective. Where such distinctions are not straightforward, nonlinear classifiers such as kernel-based techniques like *support vector machine* (SVM), which are relatively complex, can be applied.

A *discriminant* is a function which maps an output class label to an input variable. A discriminant function which is linear in the input variables is termed as linear discriminant function [151], [175], [176]. The aim of discriminant analysis is to reduce the dimensionality of the input feature space and project an *n*-dimensional input data (for *n* features) to a single dimensional space/line. However, not all projections will optimally separate the different classes. *Fisher's linear discriminant analysis* is used to find the optimal direction of the projected space/line which separates all the classes optimally. For a two class problem, this is shown in Figure 4.7 [151], [175], [176].

The figure clearly shows that the separation of the two classes is well defined in projection **Wa** (the red line) but not on projection **Wb** (the blue line). It can be inferred that the direction in which the variance of the classes is minimal, is the optimal direction for projection (as it

avoids overlapping of the projected clusters). The method for computing the optimal projection is discussed in Section 4.6.3.

The choice of a classifier (discriminant or complex kernelized SVM) may sometimes be determined by looking at the distribution plot of the features of the two groups. If the distribution plots show two well-separated means, a simple linear or other discriminant analysis based classifiers should be able to classify the data sufficiently.

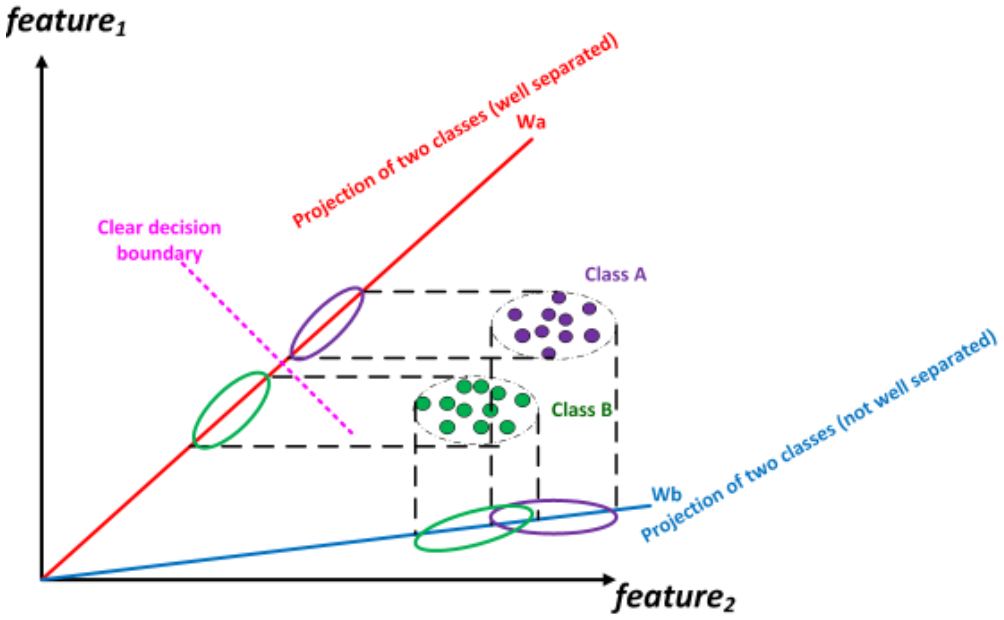


Figure 4.7: Projection of two classes (Class A and Class B) in a 2-dimensional feature space

Figure 4.8 gives an example of this, where the distribution of two different observations is shown (i.e. two different classes, created using random numbers) with means close to each other ($\mu_1 = 2.3$ and μ_2=2.5) with standard deviations of $\sigma_{1,2} = 0.1$. Observe that the distributions are not clearly separated. By contrast, in Figure 4.9 (a) shows two distribution plots, with means at μ_1=2.3 and μ_2=5 with standard deviations of $\sigma_{1,2} = 0.1$, with a clear separation between the distributions. Figure 4.9 (b) shows two such clearly separable distributions, represented as a 2D scatter plot, where each axis represents a feature.

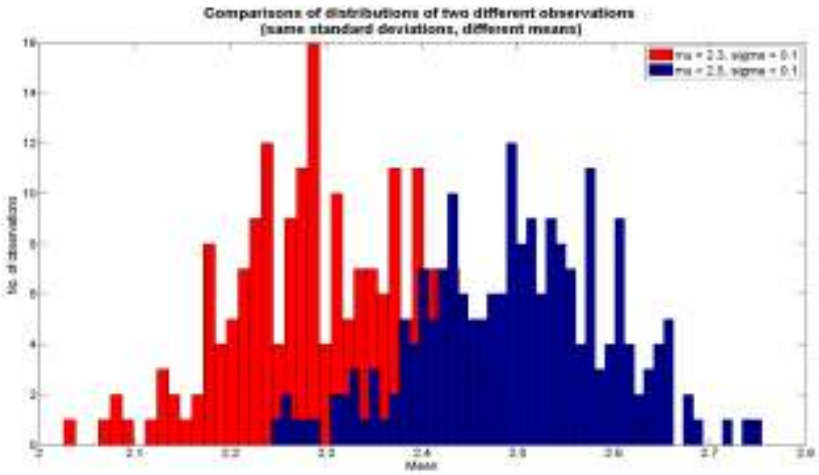
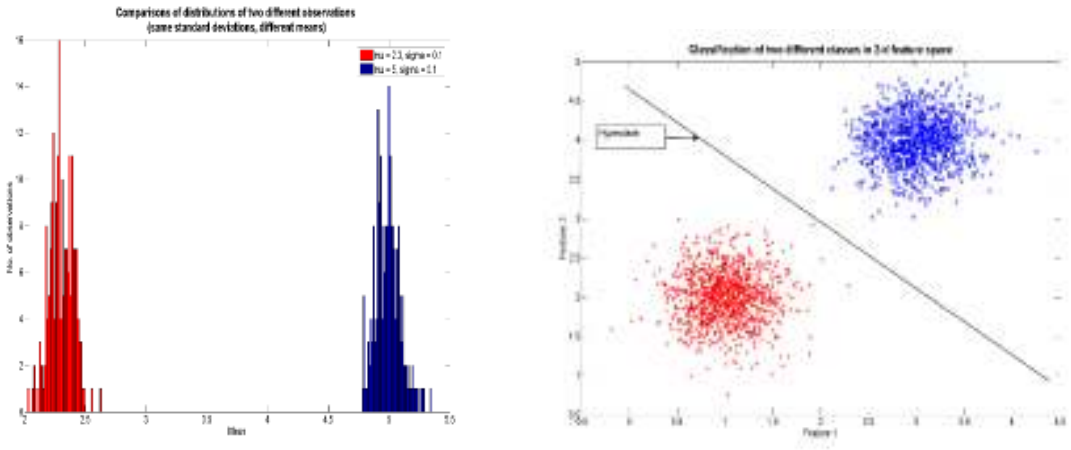


Figure 4.8: Histogram plots of two different distributions with means close to each other



(a) (b)

Figure 4.9: (a) Histogram plots of two different distributions with means well separated from each other; (b) Scatter plot showing two classes in 2 dimensional feature spaces

Figure 4.9 (b) shows that if the data is well separated, it is easier to draw a straight boundary line between the two. In such cases, a discriminant analysis based classifier is usually a sensible option. In cases where the distributions are not easily separable, nonlinear classifiers, which can draw nonlinear boundaries between the groups, is the correct option to use.

Unnecessarily involving a complex nonlinear classification technique often gives high classification accuracy on the training dataset, but is prone to over-fitting. The present work focused on four different discriminant analysis classifiers and one distance based classifier (*Mahalanobis* distance), using MATLAB for training on the datasets available.

Different discriminant analysis classifiers were used because of their simplicity, to see the characteristic changes traced in the features due to the stimuli. Two approaches could be taken in classification: 1) choice of meaningful statistical features followed by simple classifier, 2) simple features followed by a complex classifier. The former option is preferable in the current work since it may help in understanding the changes in statistical behaviour of

the signal which might be indicative of some consistent modification of the underlying biological process. Also, since this exploration was the first of its kind, a decision was made to use simple classifiers initially.

4.6.2 Cross Validation

Cross validation schemes are often employed to avoid the introduction of any possible bias from the training dataset and to enhance the generalization capability of the classifier [175]. This work used the *leave one out cross validation* (LOOCV) scheme where, if there are *N* data-points, then *N–1* samples are used for training the classifier and the remaining sample is used to test the trained model. Thereafter, the single test sample will be included in the next training set, and again a new sample from the previous training set will be set aside as the new test data. This loop will repeat *N* times, till all the samples have been tested and the average classification accuracy for all the *N* instances has been calculated. The final result is the average of all the *N* runs ensuring that data belonging to all classes (i.e. all class labels) are well shuffled and the best average estimate of the final classifier accuracy is obtained [175]. The advantage of LOOCV over other N-fold cross-validation schemes is that the introduction of undesired class bias is minimal.

4.6.3 Binary Classification – two classes A & B

When two classes need to be identified, the class means could be either close to each other (Figure 4.8) or well separated from each other (Figure 4.9). In either case the spread of the data (variance) plays a crucial role in separating the two classes.

Figure 4.9 shows an ideal case of binary class separation, where the straight line (b) is the decision boundary that separates the two classes. This is also seen in Figure 4.7, where the decision boundary is seen to be orthogonal to the optimal plane of projection (of single dimension, with maximal separability for the two classes), denoted by **Wa**. The **Wa**, corresponding to the direction of optimal separation of the two classes, is determined by maximizing the ratio of the difference of the means of the two classes and the sum of the class variances. Once the optimal projection has been obtained (through training the classifier), some kind of threshold on this projection can be set which will help assign a *new observation* to one of the two classes by the inner product of the weight vector **Wa** (representing the optimal projection) and the new observation. By the term *new observation*,

is denoted the n-dimensional feature vector f_i extracted from the *new observation* (i.e. from a new block of 1000 samples of the plant electrical signal) belonging to class i (say, O_3).

$$y = (Wa)^T f_i \tag{4.9}$$

From the linear combination of the features (components) from the new observation, comes the scalar dot product shown in equation (4.9). T can be set as some threshold value (usually the midpoint of two class labels), so that if $y \geq T$, then $i = A$, else $i = B$, where A, B are the two classes we are trying to classify [151], [175], [176].

Thus, the dimensionality of the observation is being reduced from n-dimensions to a single dimension by merely projecting it onto a line, through linear combinations.

The line of projection with maximal separation between the two classes can be found by looking at the mean of each class. The mean vector of *each class* can be written in generic form as

$$\mu_i = \frac{1}{N_i} \sum_{x_i \in i} f_i \tag{4.10}$$

where f_i is an n-dimensional feature matrix belonging to class i (i.e. Class A or Class B in the binary class scenario), with total N_i components.

Also, the mean of the class on the projected line is given by (4.11) [151], [175], [176]:

$$\tilde{\mu}_i = Wa^T \left(\frac{1}{N_i} \sum_{f_i \in i} f_i \right) = Wa^T \mu_i \tag{4.11}$$

Here $\tilde{\mu}_i$ is the projection of μ_i. Thereafter, the difference between the means of the two projected classes can be used, shown in (4.12), as the objective function to be maximised.

$$\left| \tilde{\mu}_{i=A} - \tilde{\mu}_{i=B} \right| = \left| W_a \left(\mu_{i=A} - \mu_{i=B} \right) \right| \tag{4.12}$$

However, referring back to Figure 4.9 (a), notice that the optimal separability also depends on the standard deviation of each class. Thus the solution given by *Fisher* is to maximize the ratio of the distance between the two class means and the within-class variation. This variation, on the projected line y, is referred to as *scatter* and is given for each class as follows [151], [175], [176]:

$$\tilde{s}_i^{\,2} = \sum_{y \in i} \left(y - \tilde{\mu}_i \right)^2 \tag{4.13}$$

Thus, the objective function to be maximized, in order to get a good separation between the two classes, is $J\left(W_a\right)$ and is given by (4.14) [151], [175], [176]

$$J(Wa) = \frac{\left|\tilde{\mu}_A - \tilde{\mu}_B\right|^2}{\tilde{s}_A^{\,2} + \tilde{s}_B^{\,2}} \tag{4.14}$$

Here, $\tilde{s}_A^{\,2} + \tilde{s}_B^{\,2}$ is defined as the total *within class scatter* for the two classes $i = A, B$

Equation (4.15) can be modified into [151], [175], [176]

$$\tilde{s}_i^2 = \sum \left(Wa^T f_i - Wa^T \mu_i \right)^2 \tag{4.15}$$

or

$$\tilde{s}_i^2 = \sum Wa^T \left(f_i - \mu_i \right)\left(f_i - \mu_i \right)^T Wa \tag{4.16}$$

or

$$\tilde{s}_i^{\,2} = Wa^T S_i Wa \tag{4.17}$$

Here, S_i is defined as the covariance matrix. Thus, for the two classes, we get

$$\tilde{s}_A^2 + \tilde{s}_B^2 = Wa^T S_W Wa \tag{4.18}$$

where S_W is the *within-class* scatter matrix of the projected samples of both the classes [151], [175], [176].

Similarly, the difference of the means in the projected line is expressed in terms of the differences of the means in the original feature space as follows [151], [175], [176]:

$$\left(\tilde{\mu}_A - \tilde{\mu}_B\right)^2 = \left(Wa^T \mu_A - Wa^T \mu_B\right)^2 = Wa^T \left(\mu_A - \mu_B\right)\left(\mu_A - \mu_B\right)^T Wa \tag{4.19}$$

or

$$\left(\tilde{\mu}_A - \tilde{\mu}_B\right)^2 = Wa^T S_B Wa \tag{4.20}$$

The term S_B is called the *between-class* scatter matrix of the original feature vector [151], [175], [176].

Hence, the *Fisher* criteria can be expressed by within-class and between-class scatter matrix as follows [151], [175], [176]:

$$J(Wa) = \frac{Wa^T S_B Wa}{Wa^T S_W Wa} \tag{4.21}$$

Since the aim is to maximize the function $J(Wa)$ given in (4.22), we can solve the generalized eigenvalue problem [151], [175], [176]:

$$S_W^{-1} S_B Wa = \lambda Wa \tag{4.22}$$

where λ is the eigenvalue (a constant). The eigenvectors corresponding to the largest eigenvalues constitute the projection where the separability of the two classes is maximised.

Thus *Fisher's linear discriminant*, which is the optimal direction of projection on to a one dimensional hyperplane for maximal separation of two classes, has been found [151], [175], [176].

4.6.3.1 Linear discriminant analysis classifier

The *linear discriminant analysis* (LDA) algorithm, takes an input feature vector *F* and assigns it to one of the classes (binary in this work). The output of an LDA is given by (4.23) [151], [175], [176].

$$y = \sum_{j=1}^{M} w_j^T f_j + b \tag{4.23}$$

Here, w_j are the components of the weight vector Wa (optimal projection) and f_j are the components of the *M*-dimensional feature vector *F*. *b* is the bias/offset which determines the location of the class-separating hyperplane. The weight vector Wa determines the orientation of the hyperplane. Thus, a real number *y* is given as a linear combination of all the input features and their corresponding weights. This real number *y* reflects the observation in reduced (single) dimensions. The LDA classifier performs well when the data is linearly separable. Where the data is not linearly separable, more complex decision boundaries may be required to separate the data [151], [175], [176]. For LDA, the covariance matrix is

assumed to be identical for all classes and hence only one covariance matrix is estimated (pooled) for all classes [151], [175], [176].

4.6.3.2 Quadratic discriminant analysis classifier

The linear discriminant function can be extended to include the products of pairs of components of the features (variables) and (4.23) can be extended to (4.24) [151], [175], [176].

$$y = \sum_{j=1}^{M} w_j x_j + \sum_{j=1}^{M} \sum_{k=1}^{M} w_{jk} x_j x_k + b \tag{4.24}$$

Therefore, the additional $\frac{M\,(M\,+1)}{2}$ terms in the quadratic discriminant function in (4.24) allow more complex decision boundaries to be drawn between the classes [151], [175], [177]. For QDA, it is assumed that the covariance matrix is separate for each class and hence estimated separately [151], [175], [176].

4.6.3.3 Diaglinear and Diagquadratic classifiers

Diaglinear and Diagquadratic classifiers, also known as Naïve Bayes method, is based on the assumption that the features (variables) are independent, thus ignoring any information sharing between the features. That is why the non-diagonal elements in the covariance matrix for the linear and the quadratic form (given by (4.23) and (4.24)), are presumed zero and only the diagonal elements are taken into account [151], [175], [177].

4.6.3.4 Mahalanobis distance classifier

Instead of the usual Euclidean distance, when variances are different in different directions the *Mahalanobis distance* is a common measure between two points (usually a point and the mean of the multivariate data). The *Mahalanobis distance classifier* is used to classify a point based on the least distance between the point and the class mean [151], [175], [177].

The Mahalanobis distance between a feature vector F and the mean vector m_i, of a class $i \in A, B$ is given in (4.25) and explained briefly below [151], [175], [176]:

$$r_i = \sqrt{\left(F - m_i\right)^T \cdot \Sigma^{-1} \cdot \left(F - m_i\right)} \tag{4.25}$$

The estimate of the covariance matrix is denoted by Σ. If $r_A > r_B$, then the feature vector belongs to class B, else it belongs to class A. In a 2D space (e.g. when two features are

chosen), the region of constant Mahalanobis distance forms an ellipse. For spaces higher than 2D, ellipsoid or hyperellipsoids are formed. The Mahalanobis distance of 1 unit corresponds to 1 standard deviation along all primary axes of variance (i.e. 2 primary axes when 2 features are chosen, etc.). If $\sum$ happens to be the identity matrix, then the Mahalanobis distance becomes the common Euclidean distance [151], [175], [177].

4.7 Results

This section presents the results obtained for a binary classification. Since four stimuli were used, there were six possible binary classification settings and the classification results were obtained for all six.

4.7.1 Conditions and measures

If a particular stimulus/class was present, it is termed as *Positive*. When the classification algorithm correctly found the presence of a *Positive* in a set of observations (i.e. the test dataset), it was called *True Positive* (TP).

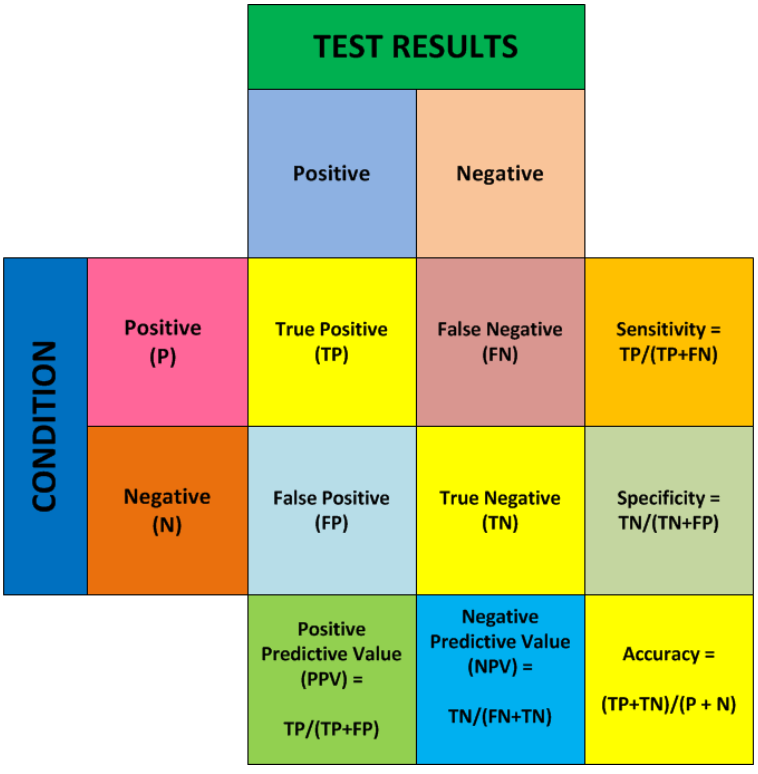


Figure 4.10: Confusion matrix showing different measures of classification

When the algorithm detected a *Positive* out of the set, but in reality it was not a *Positive*, this situation is called *False Positive.* Similarly, when the algorithm detected the presence of a

Negative correctly (i.e. the presence of the other class in a binary class setting), this was called *True Negative*. However, the detection of a *Negative*, where in reality it was a *Positive*, is called a *False Negative*. Figure 4.10 summarizes the conditions and the various measures which describes them. This figure is called the Confusion matrix and is widely used in classification tasks [151], [177].

The five measures – *Sensitivity*, *Specificity*, *PPV*, *NPV* and *Accuracy* were computed for all the binary classification scenarios here. Accuracy was taken into account to determine the classification success of any two different stimuli detected through the features extracted from the plant electrical signal response [151], [177].

4.7.2 Pre-processing for pre-stimulus parts of the data

Figure 4.4 showed the four plant electrical signal responses to four different stimuli started at different amplitude levels. The background signal (even before the application of the stimuli) was different in all four cases. This could have biased the final classification result due to the background information already separated within the multiple features considered. As a precautionary measure, this background information had somehow to be removed from the signals. Because of these different backgrounds, a clear separation could be seen between the stimuli for some features such as Hjorth mobility, Hjorth complexity and skewness, in Figure 4.11, where histogram plots for each of the features for each stimulus are plotted (without the subtraction of any background information).

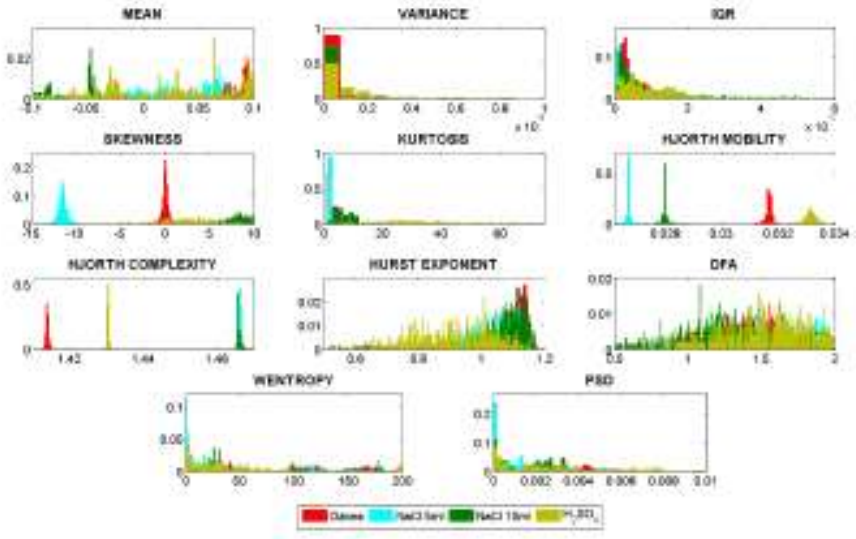


Figure 4.11: Normalized histogram plots for 11 individual features showing stimuli separability (no background subtraction).

This encouraged an inspection of only *incremental values* of the features under different stimuli. The incremental values were obtained by subtracting the mean of every feature extracted from the pre-stimulus part of the signal, from the corresponding feature extracted from the post-stimulus part. The histogram plots of the incremental values of the individual features, after the background was subtracted, are shown in Figure 4.12, which shows a less separability in the stimuli, as expected. Therefore, the approach here was to carry out this background subtraction prior to feature normalization.

These incremental values of the features were used to see how good they were in providing a successful classification (using five different classifiers) between any two stimuli (six binary combinations of four stimuli). As an example, although the histogram plots in Figure 4.12 shows clear separation of the distributions for NaCl and O_3 using skewness as a feature (due to their peaky nature), the frequency of occurrence of the histograms show that the distributions have wider spread (for other classes) which has been reflected by the moderate rate of classification reported in the next subsections using skewness as a feature. Figure 4.12 shows that though skewness shows a good discrimination between NaCl 5 ml and other stimuli, Table 4.3 shows that the average classification accuracy using skewness as an individual feature is very low, because skewness as an individual feature did not give good classification results between the remaining stimuli combinations.

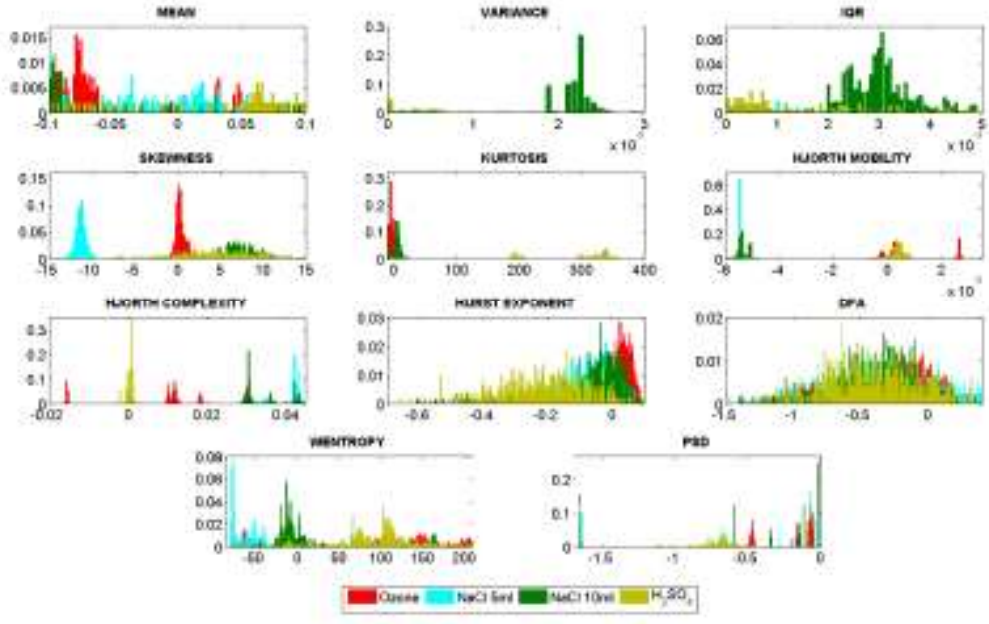


Figure 4.12: Univariate histograms of each of the 11 features
for four different stimuli (with background subtraction)

4.7.3 Correlation between features to avoid redundancy

A correlation test was carried out to find out the inter-dependence between all the features. The result of this test, given in Table 4.2, was obtained by checking the *Pearson correlation coefficient* values between all feature pairs (across all classes). The *Pearson correlation* is given by (4.26), where *A*, *B* are any two features with *N* elements.

$$\rho_{A,B} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\overline{A_i - \mu_A}}{\sigma_A}\right)\left(\frac{B_i - \mu_B}{\sigma_B}\right) \tag{4.26}$$

where μ_A, μ_B represent the means of the features *A*, *B*, and σ_A, σ_B represents their standard deviation.

Table 4.2: Correlation coefficient between 11 statistical features extracted from plant electrical signals (after subtracting the mean of the pre-stimulus features from the post-stimulus ones)

Features	μ	σ^2	IQR	γ	β	m_H	c_H	H	α	WE	$\overline{P}$
$f_1 = \mu$	1.00	0.09	–0.03	–0.06	0.07	0.04	0.03	–0.11	–0.22	***0.70***	0.26
$f_2 = \sigma^2$	*	1.00	***0.83***	0.01	0.10	–0.05	–0.23	–0.10	0.21	0.02	0.07
$f_3 = IQR$	*	*	1.00	–0.04	0.02	–0.08	–0.31	0.01	***0.53***	–0.12	0.05
$f_4 = \gamma$	*	*	*	1.00	0.29	0.00	–0.06	–0.09	–0.07	–0.08	0.00
$f_5 = \beta$	*	*	*	*	1.00	–0.01	–0.23	–0.14	–0.03	0.14	0.06
$f_6 = m_H$	*	*	*	*	*	1.00	0.34	–0.07	–0.10	0.06	0.02
$f_7 = c_H$	*	*	*	*	*	*	1.00	–0.12	–0.28	0.06	0.09
$f_8 = H$	*	*	*	*	*	*	*	1.00	***0.64***	–0.15	–0.16
$f_9 = \alpha$	*	*	*	*	*	*	*	*	1.00	–0.29	–0.06
$f_{10} = WE$	*	*	*	*	*	*	*	*	*	1.00	–0.09
$f_{11} = \overline{P}$	*	*	*	*	*	*	*	*	*	*	1.00

A correlation value near to +1/–1) indicates a strong positive/negative correlation between a feature pair, whereas a value closer to zero indicates the feature pairs are independent and thus more informative about the underlying process. A good classification strategy should ideally involve uncorrelated features, in order to avoid redundancy in training the classifier. This work initially took all features into account and then ignored the ones with high correlations.

4.7.4 Classification using univariate features

As a first attempt, the classification results were obtained in two ways – using univariate and bivariate features, to make the analysis intuitive and simpler to infer. That is, instead of taking all the features together to get a multivariate classification (which may give good classification accuracy but becomes less intuitive due to the increase in complexity and dimensions of the feature matrix), the results with 11 individual features and 55 possible feature pairs, were explored as a starting point.

Table 4.3 presents the classification accuracies, obtained using each individual feature, averaged across all the six stimuli combinations and all the five different classifier variants. The features $f_1 \cdots f_{11}$ were ranked by classification accuracy obtained using each one of them individually (for all binary stimuli combinations). The final ranked features were labelled $F_1 \cdots F_{11}$, with F_1 being associated with the highest classification accuracy achieved and F_{11} being associated with the lowest accuracy. Although the *Fisher ratio* was computed to rank the features for every binary stimuli combination, it was not sufficient to determine which feature provided the best accuracy for *all* binary stimuli combinations.

Table 4.3 shows that the feature *mean* on its own produced the best classification result for all combinations. However, since the features were extracted from the raw non-stationary signals, *mean* was not a very reliable feature to base any conclusions on because it could be influenced by various artefacts and noise during measurement or from various environmental factors (e.g. a sudden breeze could shake the electrodes connected to the plant body, etc.). The next five best features (best average classification accuracies), when taken individually, were *Wavelet packet entropy* (F_2), *Hjorth complexity* (F_3), *Interquartile range* (F_4), *Variance* (F_5) and *Average spectral power* (F_6). From here onwards, only these features were considered as candidates for the top five features. Table 4.3 also reports the best accuracy achieved by the best classifier using each of the single features to discriminate the four

stimuli within a '*one vs. the rest*' strategy (i.e. NaCl 5 ml vs. the rest, NaCl 10 ml vs. the rest, and so on). This highlighted the possibility of isolating one particular class from the other classes using a single feature, with a certain degree of confidence. The averaged results of classification for six binary stimuli combinations using individual features has now been produced. The next approach was to find the best classified binary stimuli combination using only the top five individual features (F_2 to F_6) and using the five classifier variants. As a result, five classification accuracies (for five individual features) were obtained for every classifier for each of the six binary stimuli combinations. This resulted in 25 classification accuracies obtained for each of the six binary stimuli combinations, as shown in Figure 4.13.

Table 4.3: Average accuracy and best accuracy for classification using individual features

Ranked Features		**Average classification accuracy for all *binary* stimulus combinations (%)***	**Best classification accuracy for all *one vs. the rest* stimulus combinations (%)****
F_1	Mean (μ)	70.87	73.00, *Mahalanobis*
F_2	Wentropy (*WE*)	69.79	62.26, *Mahalanobis*
F_3	Hjorth Complexity (c_H)	66.61	60.82, *Mahalanobis*
F_4	Inter Quartile Range (*IQR*)	65.07	63.62, *LDA/Diag-LDA*
F_5	Variance (σ^2)	63.57	65.58, *LDA/Diag-LDA*
F_6	Average spectral power ($\overline{P}$)	60.51	61.58, *Mahalanobis*
F_7	DFA (α)	60.14	61.28, *Mahalanobis*
F_8	Kurtosis (β)	58.06	62.64, *Mahalanobis*
F_9	Hjorth Mobility (m_H)	57.45	61.44, *Mahalanobis*
F_{10}	Skewness (γ)	54.55	62.09, *Mahalanobis*
F_{11}	Hurst Exponent (*H*)	52.38	61.40, *Mahalanobis*

* averaged across all six *binary* stimuli combinations and all five classifier variants. This column infers how good a feature is in terms of classifying different binary stimulus combinations.
** averaged across four *one vs. the rest* stimuli combinations. This column infers how good a classifier is in terms of *one vs. the rest* classification using the corresponding features.

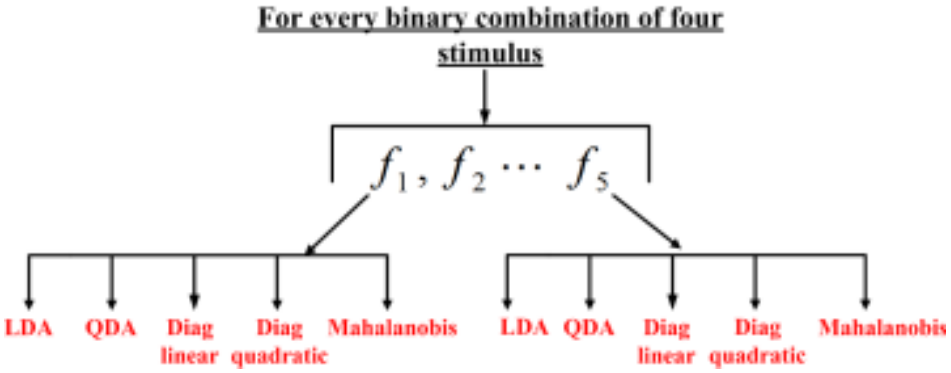


Figure 4.13: Classification using five classifiers for the top five features

All these 25 results were averaged for each stimuli combination and given in Table 4.4. This shows the best discrimination possible is for H_2SO_4 and O_3 (highlighted in bold) with an average classification accuracy of around 73%. Additionally, discrimination between NaCl (both concentrations) and O_3/H_2SO_4 also shows promising results, with accuracy over 65% and 63% respectively.

The average classification results presented in Table 4.4 gave a preliminary overview of how a particular binary stimuli combination was classified on average, across all five classifiers with the top five chosen features (when used individually). This encouraged a search, from the best results achieved for each stimuli combination, to see whether there was any consistent feature giving good classification results.

Table 4.4: Accuracy using top five individual (univariate) features (F_2 through F_6) and averaged across five classifiers (average separability between different stimulus combinations)

Stimulus	**NaCl 5 ml**	**NaCl 10 ml**	**H_2SO_4**	**O_3**
NaCl 5 ml	-	57.20%	64.02%	65.94%
NaCl 10 ml	*	-	63.17%	67.29%
H_2SO_4	*	*	-	**73.03%**
O_3	*	*	*	-

Table 4.5: Best accuracy taking individual features for each stimulus combinations (best separability between different stimulus combinations)

Stimulus	**NaCl 5 ml**	**NaCl 10 ml**	**H_2SO_4**	**O_3**
NaCl 5 ml	-	74.36% (F_3, LDA classifier)	75.09% (F_3, Mahalonobis classifier)	78.95% (F_3, LDA classifier)
NaCl 10 ml	*	-	72.13% (F_8, LDA classifier)	82.27% (F_9, QDA classifier)
H_2SO_4	*	*	-	94.95% (F_2, QDA classifier)
O_3	*	*	*	-

From Table 4.5 it was evident that F_3 (*Hjorth complexity*) gave the best result for three out of six different binary stimuli combinations, with an accuracy over 74%. Overall, the best accuracy was achieved for classification between H_2SO_4 and O_3, with an accuracy >94% using F_2 (Wentropy) and QDA classifier. Although in Table 4.4, the discrimination between

NaCl and O_3/H_2SO_4 are shown in terms of the average accuracy, which might seem to be relatively low (63% to 67%), the best cases for such a discrimination can be found in Table 4.5 (accuracies of 72% to 82%) between the same set of stimuli. Thus, out of all binary combinations, O_3 *vs*. H_2SO_4 achieved the best classification accuracy with the best results obtained using *Wavelet entropy* and *QDA* classifier. It is also noted that the best classification accuracies for NaCl 10 ml *vs*. H_2SO_4 was achieved using F_8 and the best classification accuracy for NaCl *vs*. O_3 was achieved using F_9. However, both F_8 and F_9 do not produce overall good classification accuracies across all binary stimuli combinations, hence their ranking is low.

4.7.5 Classification using feature pairs (bivariate)

This experiment sought to find out whether there was any improvement on the classification accuracy when a feature pair was used, rather than individual features.

The effect of all possible feature pairs was looked at using 11 individual features (totalling 55 independent feature pairs) on the classification results between six different stimuli combinations, averaged across all classifiers. These classification accuracies are shown in Figure 4.14 together with the difference in accuracy (error) when the background is not subtracted as discussed earlier.

The features mentioned as $\{1, 2, \cdots, 11\}$ in Figure 4.14 were the unranked features designated by $\{f_1, f_2, \cdots, f_{11}\}$ respectively in Table 4.2. Although the classification accuracies using all 55 bivariate combinations were observed, the combinations of the top five features ($F_2 \cdots F_6$ ignoring *mean*) were addressed, just as for the individual feature experiments. The results obtained using each of these bivariate features, using all five classifier variants, were averaged and are given in Table 4.6. These results were found to be better than the averaged results obtained using just univariate features given in Table 4.4 (for all the binary stimuli combinations, except NaCl 5 ml *vs*. H_2SO_4).

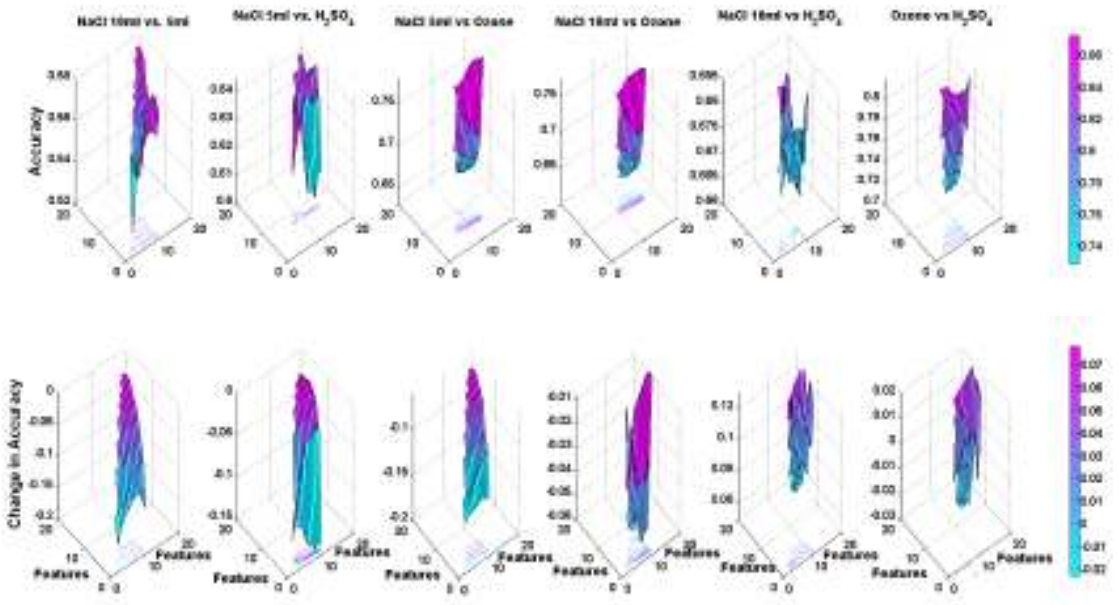


Figure 4.14: (top) Classification accuracy for different feature combinations with background information removed; (bottom) deterioration in accuracy for the features including background information.

Table 4.6 also shows that the top two best accuracies were obtained for stimuli combinations of NaCl 10 ml *vs*. O_3 and H_2SO_4 *vs*. O_3 (highlighted in bold). From Table 4.4 and Table 4.6 can be inferred that H_2SO_4 *vs*. O_3 combination is better classified than all other binary stimuli combinations. Further, the best feature pairs, among binary combinations of all 11 features, are given in Table 4.7.

Table 4.6: Average accuracy obtained using top five feature combinations (bivariate) and five classifiers (average separability between different stimulus combinations)

Stimulus	**NaCl 5 ml**	**NaCl 10 ml**	**H_2SO_4**	**O_3**
NaCl 5 ml	-	59.52%	58.21%	72.69%
NaCl 10 ml	*	-	64.66%	**76.60%**
H_2SO_4	*	*	-	**74.60%**
O_3	*	*	*	-

Table 4.7: Best accuracy for each stimulus combination using bivariate features (best separability between different stimulus combinations)

Stimuli	NaCl 5 ml	NaCl 10 ml	H_2SO_4	O_3
NaCl 5 ml	-	63.18% (F_4-F_6 with Diaglinear)	65.87% (F_4-F_6 with LDA)	82.69% (F_4-F_5 with Diaglinear)
NaCl 10 ml	*	-	73.18% (F_4-F_5 with Diagquadratic)	92.06% (F_4-F_5 with Mahalanobis)
H_2SO_4	*	*	-	87.48% (F_4-F_5 with QDA)
O_3	*	*	*	-

We observed that a combination of F_4 (IQR) and F_5 (variance) resulted in the best classification accuracies for four out of six different stimuli combinations. For the remaining two stimuli combinations, a feature pair of F_4 and F_6 (average spectral power) provided the best classification accuracies. Again, comparing Table 4.5 with Table 4.7, notice that three out of six binary stimulus combinations, NaCl 5 ml *vs*. NaCl 10 ml, NaCl 5 ml *vs*. H_2SO_4, and H_2SO_4 *vs*. O_3 were classified better with individual features than feature pairs.

4.7.6 Finding the most reliable combinations of feature/feature pair and classifier variant

Individual features F_2, F_3, F_8 and F_9, and feature pairs F_4 & F_5 and F_4 & F_6, produced the best classification results for one or more (out of the six) stimuli combinations. Each of these individual features and feature pairs were explored for *all* stimuli combinations. Table 4.8 gives the results of classification when F_2, F_3, F_8 and F_9 were used as an individual feature using all classifier variants for all *binary* stimuli combinations. Table 4.8 shows that F_2 or F_3 provided consistently better average classification accuracies than F_8 or F_9. Also notice that although F_2 provided a better classification for the stimuli combinations NaCl 10 ml *vs*. O_3 and O_3 *vs*. H_2SO_4, F_3 provided more consistent and better results for the remaining combinations. When considering a single feature for discriminating the four stimuli, the best average result (73%) could be obtained using the F_2 (Wentropy) with Mahalanobis classifier, although it is highly correlated with the signal mean (F_1) as shown in Table 4.2. Since *mean* as a feature was ignored due to its susceptibility to artefacts, ideally *Wentropy* should also be ignored. If *Wentropy* is ignored, then *Hjorth complexity* may be proposed as the next best

individual feature along with *LDA*/*Diaglinear* classifier, for achieving an average classification accuracy of around 71% across the six stimuli combinations.

Table 4.8: Accuracy (in %) of different classifiers for six stimuli combinations using the best individual features

Individual feature	**Classifier variant**	**NaCl 5 ml vs 10 ml**	**NaCl 5 ml vs H_2SO_4**	**NaCl 5 ml vs O_3**	**NaCl 10 ml vs O_3**	**NaCl 10 ml vs H_2SO_4**	**O_3 vs H_2SO_4**	**Average Accuracy (%)**
F_2 (Wentropy)	LDA	55.3	66.4	73.4	77.6	59.5	82.8	69.2
	QDA	52.2	67.6	62.0	74.0	56.4	95.0	67.9
	Diaglinear	55.3	66.4	73.4	77.6	59.5	82.8	69.2
	Diagquadratic	52.2	67.6	62.0	74.0	67.3	95.0	69.7
	Mahalanobis	55.5	73.1	73.7	78.2	63.8	94.4	**73.1**
F_3 (Hjorth Complexity)	LDA	74.4	73.9	78.9	66	68.3	66.9	**71.4**
	QDA	74.1	47.1	61.6	71.5	67.5	41.8	60.6
	Diaglinear	74.4	73.9	78.9	66	68.3	66.9	**71.4**
	Diagquadratic	74.1	47.1	61.6	71.5	67.5	41.8	60.6
	Mahalanobis	74.4	75.1	62.4	51.1	69.4	81.5	69.0
F_8 (Kurtosis)	LDA	57.1	66.5	57.6	60.5	72.1	66.1	63.3
	QDA	47.8	38.4	69.3	47.1	71.8	22.5	49.5
	Diaglinear	57.1	66.5	57.6	60.5	72.1	66.1	63.3
	Diagquadratic	47.8	38.4	69.3	47.1	71.8	22.5	49.5
	Mahalanobis	57.7	58.2	38.5	81.3	71.3	81.0	64.7
F_9 (Hjorth Mobility)	LDA	60.4	73.9	68.5	66.0	56.0	35.8	60.1
	QDA	49.7	47.7	76.4	82.3	48.7	81.6	64.4
	Diaglinear	60.4	48.8	68.5	66	56	35.8	55.9
	Diagquadratic	49.7	47.7	76.4	82.3	48.7	81.6	64.4
	Mahalanobis	66.2	54.6	30.1	24.7	58.4	20.5	42.4

Table 4.9: Accuracy of different classifiers for six stimuli combinations (in %) using the best feature pairs

Best feature set	**Classifiers**	**NaCl 5 ml vs 10 ml**	**NaCl 5 ml vs H_2SO_4**	**NaCl 5 ml vs O_3**	**NaCl 10 ml vs O_3**	**NaCl 10 ml vs H_2SO_4**	**O_3 vs H_2SO_4**	**Average Accuracy (%)**
F_4 & F_5 (IQR-Variance)	**LDA**	56.61	62.61	82.53	81.57	69.24	85.09	72.94
	QDA	57.78	61.62	82.24	86.06	64.69	87.49	***73.31***
	Diaglinear	63.17	53.8	82.69	83.47	62.06	80.33	70.92
	Diagquadratic	60.34	58.43	82.00	86.08	73.19	81.98	***73.67***
	Mahalanobis	62.17	50.23	80.24	92.06	53.64	78.91	69.54
								Avg. = 72.07%

F_4 & F_6 (IQR-Average spectral power)	**LDA**	56.88	65.87	71.61	70.06	67.94	81.27	**68.94**
	QDA	57.82	57.01	73.89	81.00	65.19	78.48	68.90
	Diaglinear	63.18	54.59	68.67	78.16	64.92	71.39	66.82
	Diagquadratic	58.11	60.57	79.19	75.70	67.96	79.66	68.31
	Mahalanobis	62.61	52.99	62.72	79.62	58.08	60.84	63.20
								Avg. = 67.23%

When using bivariate feature pairs, it is evident from Table 4.9 that the top two classification accuracies (>73%) were obtained using the F_4 & F_5 combination and Diagquadratic and QDA as classifiers. However when using the F_4 & F_6 feature pair for all stimuli combinations, the classification accuracies for all stimuli combinations were inferior to those obtained using F_4 & F_5. Referring back to Table 4.2, shows that *IQR* (F_4) and *variance* (F_5) were highly correlated with each other. However, as good results were being achieved using these two features, it was proposed to calculate *IQR* and *variance* from a block of 1000 samples of raw non-stationary signals, along with the *QDA* or *Diagquadratic* classifier, which would provide an average of 73% classification accuracy (highlighted by italics in Table 4.9) in identifying the external stimulus that caused the particular signature in the signal. However, this statement is limited because of the datasets on which these explorations were carried out. Figure 4.15 shows the plots separating the four stimuli, looking at 12 pairs of almost uncorrelated features, which provided best average classification accuracies (obtained across all stimuli combinations and using all five different classifiers). This is a 2D normalized histogram (volume being unity). Except the first subplot with $f_2 - f_3$ (F_4 & F_5), all the remaining combinations are almost uncorrelated and still give good classification performance.

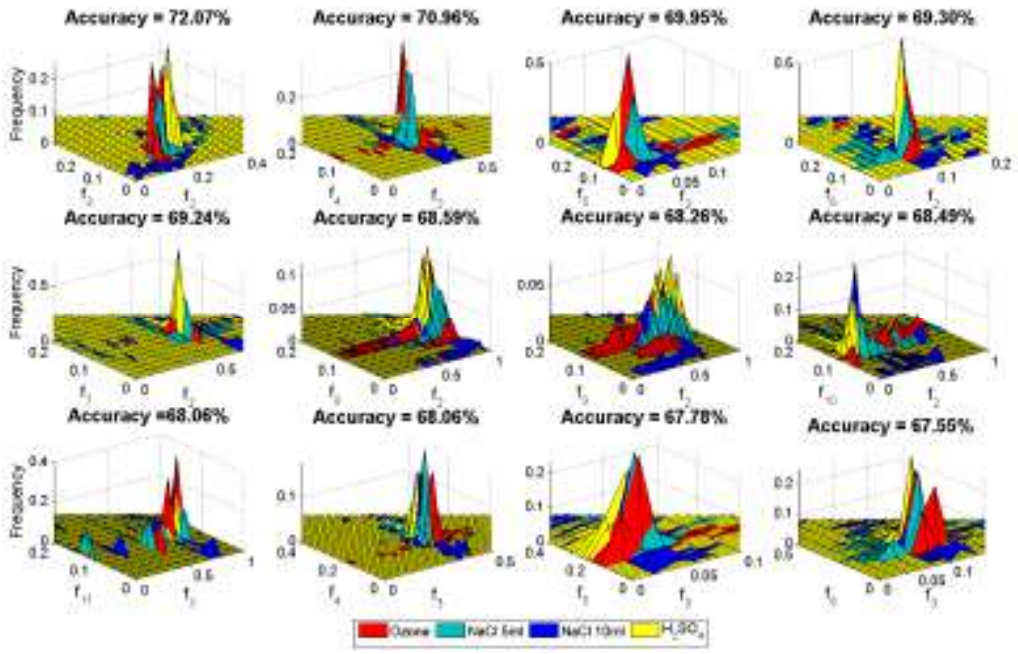


Figure 4.15: Bivariate histograms of top feature pairs with highest classification accuracy for all four stimuli

Average classification accuracy obtained is also shown (averaged over all stimuli combinations and classifiers) using particular feature pairs (denoted by $f_1, f_2, \cdots, f_{11}$ in the x and y axis, as described in Table 4.2). The second best average classification accuracy was achieved using *variance* (F_5 / f_2) and *skewness* (F_{10} / f_4) as features, which are almost uncorrelated (correlation ~0.01 in Table 4.2). The accuracies obtained using *variance* and *skewness* along with the five different classifiers is shown in Table 4.10.

Thus as a reliable measure of analysis, it can be proposed that the *variance* and *skewness* as uncorrelated features, calculated from a block of 1000 samples of raw plant electrical signals, along with the *LDA*/*Diagquadratic* classifier will be able to give an average (over all six stimuli combinations) accuracy of around 72% during binary classification of the stimuli. Again, this statement is limited to the datasets on which these explorations were carried out.

In the bivariate classification scheme, the mean (f_1) has not been considered as one of the features. Also, the best bivariate accuracies were achieved involving the *variance* (f_2) along with all the other features ($f_4 \ldots f_{11}$) in Figure 4.15, when ignoring the (f_2 - f_3) combination due to their high inter-dependence. In summary, a more reliable classification scheme is

expected (67%-70%) involving bivariate features, as shown in Figure 4.15, with respect to the univariate features given in Table 4.3 (<67%, ignoring *mean* and *Wentropy*).

Table 4.10: Accuracy (in %) of different classifiers for six stimuli combinations using Variance and Skewness

Best feature set	**Classifiers**	**NaCl 5ml vs 10ml**	**NaCl 5ml vs H_2SO_4**	**NaCl 5ml vs O_3**	**NaCl 10ml vs O_3**	**NaCl 10ml vs H_2SO_4**	**O_3 vs H_2SO_4**	**Average Accuracy (%)**
F_5 & F_{10}	**LDA**	56.07	63.51	80.96	80.03	68.51	84.26	**72.22**
	QDA	57.00	59.79	81.25	85.61	64.10	83.22	71.82
	Diaglinear	63.07	52.61	81.18	82.42	62.41	78.78	70.07
	Diagquadratic	59.69	58.52	81.45	82.09	71.54	81.73	**72.50**
	Mahalanobis	61.63	49.12	78.19	90.36	53.14	76.58	68.17

Avg. = 70.96%

4.8 Summary

This chapter met the objective of showing that the raw plant electrical signals can also be used to distinguish between more complex stimuli using simple binary classification techniques.

The exploration using raw signals, was to assist in realizing a plant electrical signal based bio-sensor to classify environmental stimuli. Binary classification as a methodology was used to pursue the goal of mapping an external stimulus of small duration with the plant's electrical signal response.

The classification scheme was based on 11 statistical features extracted from post-stimulus parts of the segmented raw signals. To account for any bias of the pre-stimulus parts, which were different for different experiments, a pre-stimulus subtraction was carried out so that the incremental feature values from the post-stimulus parts could be addressed. This was followed by normalization and ranking of the features. These features were then used for rigorous univariate and bivariate feature-based classification using classifiers based on four different discriminant analyses and on one minimum distance. The correlation of features was also investigated and features discarded that were highly correlated with each other.

External stimuli H_2SO_4, O_3, and NaCl in two different amounts (5 ml and 10 ml) were classified using the adopted approach with 11 statistical features, capturing both the stationary and non-stationary behaviour of the signal. The classification yielded a best

average accuracy of around **72%** (across all stimuli and five classifier variants) when using *variance* and *IQR* as feature pairs, and the best overall accuracy of around **73%** (across all stimuli) when using *variance* and *IQR* as feature pairs along with *QDA/Diagquadratic* classifier. The fact that, by looking at the statistical features of plant electrical response, one can successfully detect which stimulus caused the signal, is quite promising.

In the exploration presented in this chapter, the data from two channels per plant (per experiment) were used to record the electrical response, and statistical features were then calculated from both the channels and pooled. The location on the plant body for the data extraction was ignored, as the work was focussed on the possibility of classification of applied external stimuli from the signal extracted. Similarly, the effect of different species of plants have also been ignored, except for the introduction of cucumber for the O_3 experiments. The purpose is a classification scheme based on generic plant signal behaviour and not on a specific species. However, such isolation forms a good study and could be taken up as future scope of work. Possible confounding effects could occur from the position of the electrodes or plant species, but such effects will be minimal due to the large number of data samples used and the use of cross-validation to test the performance of the classifiers.

Moreover, the present classification scheme is based on the raw non-stationary plant signal. In bio-signal processing literature [178], use of a high-pass filter is recommended to make a bio-signal stationary instead of extracting features from the raw non-stationary signal. This was avoided here since, with ad-hoc filtering, some useful information in the data might get lost as the cut-off frequency for plant signal processing is not yet well established. That is why incremental features were considered (i.e. subtracting the mean of the features in the pre-stimulus part from the features of the post-stimulus parts), for training the classifier, by removing any possible bias of the channel or plants.

The segmentation of the signal in a block of 1000 samples also disregarded the temporal information of the stimuli, since the question to be answered was whether classification was indeed possible by looking at any segment of the post-stimulus part of the signal. Also, in a practical application, the algorithm would not know when the response to a particular stimulus started. So the classification needs to be based on the in-coming stream of live data.

5 A Decision Tree Based Classification Strategy to Detect External Chemical Stimuli from Raw and Filtered Plant Electrical Response

5.1 Introduction

Chapter 4 explored classification using features extracted from raw signals and obtained results for six binary stimulus combinations, involving four stimuli. This chapter explores *multi-class classification* where an incoming stream of feature vectors can be classified as belonging to a particular stimulus from a group of stimuli.

To carry out the multiclass classification, 15 features were used of which 11 are those reported in the previous chapter [179]. Four additional features have been explored. Discriminant analysis and minimum distance based classifiers are used to establish a decision tree based classification system using *One-Versus-One* (OVO) [151], [175], and *One-Versus-Rest* (OVR) [151], [175] structures. The classifiers were validated in two ways:

1) Leave One Out Cross Validation (LOOCV) on ~73% of the available data (*retrospective study*), and
2) Independent testing of the remaining ~27% of the available data (*prospective study*).

Datasets from experiments using three different stimuli were used: NaCl (combined datasets from NaCl 5 ml and NaCl 10 ml), H_2SO_4, and O_3, and can be found in [180].

The work presented in this chapter contains the following differences from the last chapter:

- An exhaustive set of experimental data with 28,070 data blocks (each block containing 1024 samples) were used for training the classifiers, about7.4 times greater than the previous data used.
- *Cabbage* was also included for extracting experimental data, in addition to *Tomato* and *Cucumber* plants which were used for the previous experiments. This was done to achieve more robust and generalized classification results.
- Features from NaCl 5 ml and NaCl 10 ml were combined and labelled as NaCl in the present work, and considered as a single class. This was to negate the fact that the soil could contain some NaCl to begin with, and see whether any classification could be achieved for NaCl as a class, irrespective of its quantity.

- Multivariate features have been used here to explore whether classification accuracies improve, despite using the same variants of the classifiers as previously.
- A systematic decision tree has been implemented here, whereas the previous classification were average results for six binary stimuli combinations.
- Along with retrospective study employing LOOCV, a separate retained dataset was used for prospective study here.

The structure of the chapter is: Section 5.2 presents the experiments conducted to record the electrical signals from plants. Section 5.3 presents the methodology for pre-processing, feature extraction, and classification, which were carried out on the raw signals. Section 5.4 presents the results and discussion of classification of stimuli, followed by an introduction to the analysis of filtered signals in Section 5.5. The results of filtered signal analysis are presented and discussed in Section 5.6, while Section 5.7 compares the results of raw and filtered signals.

5.2 Recording electrical signal from plants

Raw electrical signals from different experiments were acquired from different plants involving O_3, NaCl and H_2SO_4 as external stimuli, under lab conditions. Each experiment was conducted on a new plant, thereby eliminating the risk of any residual effect of previous experiments infiltrating the current electrophysiological response. The experimental setup was exactly as described in previous chapter (Section 4.2).

Table 5.1 gives details of the experiments conducted to extract the datasets. For experiments with H_2SO_4, 5 ml of the solution (0.025 M or 0.05 M) were applied once or twice. For experiments with O_3, the gas was injected into the box for 1 min every hour and the maximum concentration ranged from 13 to 16 ppm. NaCl treatment consisted (in the addition to whatever may be present in natural soil) of 5 or 10 ml of 3 M NaCl solution. These have been combined as a single class, i.e. NaCl.

The experiments were conducted on multiple species of plants, *Solanum lycopersicum* (Tomato), *Cucumis sativus* (Cucumber), and *Brassica oleracea* (Cabbage), so that a generalized classifier could be built which is not species-specific but picks up the common signature (representing the stimulus) in the electrical response. Whether some plants are more sensitive to certain stimuli, is left for future exploration.

Therefore, this chapter aimed at a more generalized and robust detection of chemical stimuli, rather than focusing on quantification of the variability due to the species and concentration of the stimuli.

Table 5.1: Details of the experiments with different chemical stimuli

Stimulus	Plant species used	Concentration	Number of applications of stimulus
Ozone (O_3)	Tomato & Cucumber & Cabbage	16 ppm / 13.07 ppm	Multiple
Sulfuric acid (H_2SO_4)	Tomato & Cabbage	5 ml of 0.05 / 0.025 M solution	Once / twice
Sodium Chloride (NaCl)	Tomato	5 / 10ml of 3 M solution	Once

5.3 Methodology

The methodology adopted for the multiclass classification using features extracted from the raw plant electrical signal was based on the following three processes:

- Signal pre-processing
- Feature extraction
- Classification

5.3.1 Preprocessing and data Segmentation

Since plant electrical signals are slow by nature [47], it was conjectured that there would be sufficient information about the external stimulus with ~1.5 min of data (at 10 samples/second). Thus the raw signals were divided into segments of 1024 samples using well established resampling techniques [135] as explained in Section 4.3. Although blocks of 1000 samples was used previously in Chapter 4, here the block size was increased to 1024 samples so that in a future application, the data could be saved and processed in a platform embedded in the plant.

From the time stamps it was known when the stimulus was applied for each experiment. The portion of the signal prior to the application of the stimulus was termed the pre-stimulus part of the time series, while the portion of the signal from the time of application of the stimulus was termed the post-stimulus part. This produced several blocks of 1024 samples from both

pre-stimulus and post-stimulus parts. This procedure was repeated for both channels used for every experiment conducted.

In total, some 38,000 such blocks from the post-stimulus parts were obtained from different experimental datasets. Of these, about 10,000 blocks (~27% of the total) were set aside for *prospective study*, i.e. independent testing of the classifier settings. These were obtained from completely separate experiments which were never used for training the classifiers.

The number of blocks used for each stimuli, from the post-stimulus parts, are shown in Table 5.2. This shows that the main imbalance in signal length was caused by O_3 as a class, which contributed more data blocks due to the longer duration of the experiments than NaCl and H_2SO_4 [179].

Table 5.2: Blocks (of 1024 samples) for each stimulus in different validation schemes of the classifiers

	No. of blocks belonging to post-stimulus parts of each stimuli		
Validation schemes	**NaCl**	**H_2SO_4**	**O_3**
Retrospective study (LOOCV)	352	1340	26378
Prospective study (independent testing)	276	148	9692

The best classifier and feature combinations were selected first by employing the LOOCV [181] on about 73% of the total data.

The reason behind using only 73% of the data to find the best feature-classifier combination were to

a) avoid finding an over-fitted model of the total data, and

b) test the best classifier-feature combination obtained on a section of independent data, i.e. to estimate the performance of this combination in classifying unseen data.

The aim was to find the most successful decision tree architecture out of the two (OVR/OVO), consisting of the classifier and feature combination, which would give the best multiclass classification result for both retrospective and prospective study.

5.3.2 Statistical feature extraction from segmented time series

In total, 15 features were extracted from each block of 1024 samples. Of these, 11 features have been dealt with previously, and include the statistical features *mean* (μ), *variance* (σ^2), *skewness* (γ), *kurtosis* (β), *Interquartile range* (*IQR*). Also included were features capturing

nonlinear and non-stationary behaviour [162]: *Hjorth mobility*, *Hjorth complexity*, *detrended fluctuation analysis* (DFA), *Hurst exponent*, *wavelet packet entropy*, and frequency domain feature *average spectral power* [179].

The additional four features are described below.

5.3.2.1 Hyperskewness and hyperflatness

Two out of these four additional features are higher standardized moments capturing the non-Gaussian nature of the signal given in (5.1)

$$S_{N=5,6} = \left(\left(x - \mu \right) / \sigma \right)^N \tag{5.1}$$

When $N = 5$ this is described as *hyperskewness* and when $N = 6$ as *hyperflatness*, and are further measures of the shape of the distribution of the data [182].

5.3.2.2 Fano factor

The third additional feature is the *Fano factor* or the index of dispersion described as

$$F = \left(\sigma^2 / \mu \right) \tag{5.2}$$

which also characterizes the shape of the underlying probability density function [183]. Different distributions will have different values *fano factors*. For example, the Poisson distribution takes the value 1, the negative binomial distribution takes a value higher than 1, signifying over-dispersion, and the binomial distribution takes a value between 0 and 1, signifying under-dispersion.

5.3.2.3 Correlation dimension

Nonlinear dynamical systems can be characterized by *invariant measures*, which can be estimated from a given time series. These invariant measures are indicators of the complexity of the system dynamics and one such is the dimension of the systems attractor – the set of points on the *phase plot* obtained from the system [162].

The fourth additional feature considered here is the *correlation dimension* (*D*) which is a measure of the attractor's fractal dimension [184]. The correlation dimension is estimated from the density of pairwise distances of points, usually with a *time delay embedding* for delay *t* and embedding dimension *m*. The correlation sum is first computed for a range of distances *r* (Euclidean norm is used to calculate the distance) and then inspected for self-

similarity in a log-log plot [184]. The slope of this graph gives the correlation dimension [162].

5.3.3 Classification methodology

The mean of each feature vector from the pre-stimulus parts was subtracted from corresponding feature vector from the post-stimulus parts, as explained in Section 4.7.2. This way, a feature matrix is created representing incremental values of features due to application of the stimulus.

The 15 pre-stimulus subtracted features, extracted from the blocks of the post-stimulus parts, were normalized as previously (in Section 4.4.6).

Five classifier variants were used: LDA, QDA, Diaglinear, Diagquadratic, and the Mahalanobis distance. The Diaglinear and Diagquadratic classifiers, also known as Naïve Bayes classifier, use a simple linear and quadratic kernel along with only diagonal estimates of the covariance matrix (thereby neglecting any cross-terms). The Mahalanobis distance classifier modifies the distance measure instead of using the standard Euclidean version [151], [175].

Two decision tree architectures were explored, for help in the multiclass classification problem involving the three stimuli O_3, NaCl, and H_2SO_4. These architectures are shown in Figure 5.1 and Figure 5.2. In both configurations, the features and classifier combinations were explored which would produce the best results for both cross-validation and independent testing. Thus a multiclass problem has been reduced to one or more binary classification problems [185], [186] using two distinct decision tree structures, OVR [187] and OVO [188].

5.3.3.1 One vs. Rest (OVR) Scheme

In the OVR scheme [189], [190], the three classes (stimuli) were classified in a two-node set up. In the first node, the *best separable* class (of the three) was determined (depicted as Class A in Figure 5.1) along with the best features-classifier combination. In the second node, the remaining two classes were classified.

To determine the best separable class, three binary classification settings were set up for the first node. In each binary setting, one of the stimuli was considered as *one* class and the other two were combined as the *rest* (thereby reducing the problem to a binary classification

scenario). For each of these settings, the features were ranked using the Fisher Discriminant Ratio (FDR) [151], [175].

A systematic feature selection method, the *Sequential Forward Search* (SFS) algorithm [151], [175] was carried out on the ranked features, with five different classifiers, for each of the three binary classification settings (i.e. *Class A vs. rest*, *Class B vs. rest* and *Class C vs. rest*). SFS uses incremental features for classification, starting with the highest ranked feature (FDR) and moving on to the least ranked feature (i.e. feature 1, then features 1&2, then features 1&2&3, and so on).

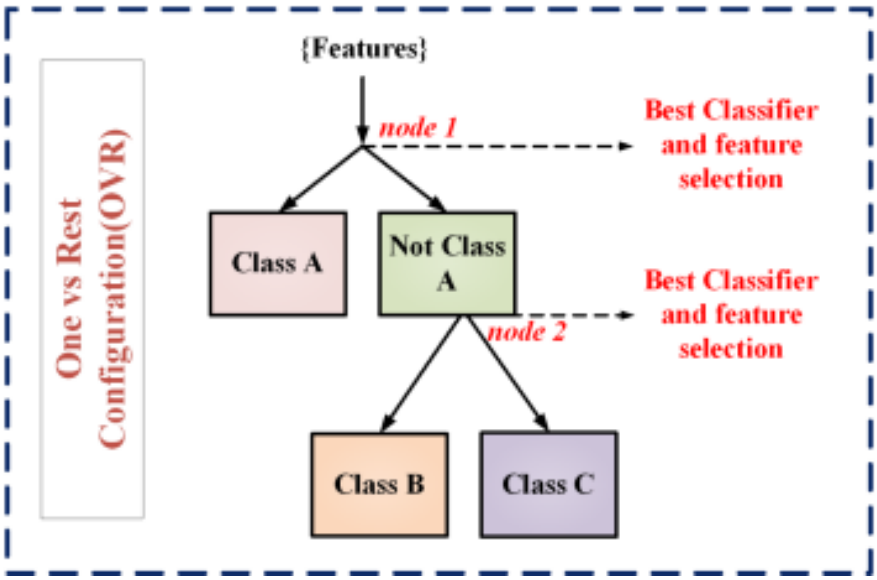


Figure 5.1: Decision Tree incorporating One Versus Rest (OVR) configuration for multiclass classification

Thus the best feature-classifier combination is determined for the best separable class. The best feature-classifier combination for the remaining two classes was also determined using a similar approach (i.e. ranking the features using FDR, then employing SFS along with five different classifiers).

5.3.3.2 One vs. One (OVO) Scheme

For the OVO configuration [189], [190], three binary classification settings – O_3 *vs.* H_2SO_4, H_2SO_4 *v.s NaCl* and *NaCl vs.* O_3, were simultaneously carried out (as shown in Figure 5.2). If two classifiers affirmed the presence of a particular class then only that particular class was predicted [191]. In the case of contradictory decisions by two classifiers, the assignment was termed as *unknown* by the algorithm. For each binary classification setting, the best features-classifier combination was determined. The feature ranking for each of the three settings were

again performed using FDR. SFS was employed again on these ranked features for all five classifier variants to determine the best feature-classifier combination for each of the three binary classifiers.

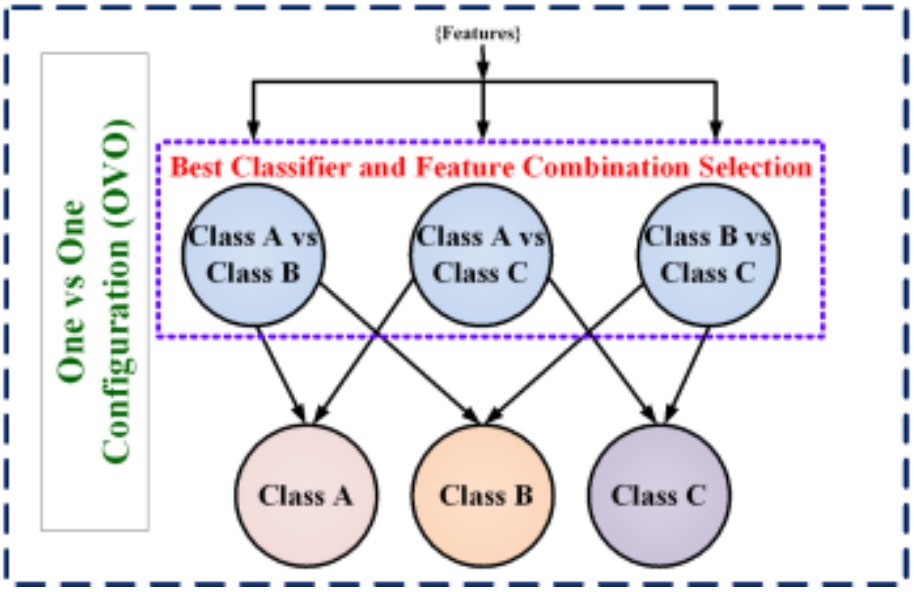


Figure 5.2: Decision Tree incorporating One Versus One (OVO) configuration for multiclass classification

5.3.3.3 Retrospective Study (Using LOOCV)

The results identifying each class in a binary classification were obtained in terms of sensitivity and specificity given by the confusion matrix shown previously (Figure 4.10).

For cases where $P \approx N$, accuracy of the classification is obtained using the traditional notion of accuracy in (5.3).

$$Accuracy = \frac{(TP + TN)}{(P + N)} \tag{5.3}$$

In the case of unbalanced data [192] involving two classes (i.e. when $P \gg N$ or vice versa), the *balanced accuracy* [193]–[196] is often used to determine the accuracy of the classification as shown in (5.4). The derivation of (5.4) is given in [194], [195] for a single run of the classifier and for a fixed threshold.

$$Accuracy(balanced) = \frac{(sensitivity + specificity)}{2} \tag{5.4}$$

Since the dataset here is unbalanced due to different duration of exposure of the plants to different stimuli, the balanced accuracy was used, called simply *accuracy* hereafter.

Identifying a difficult class within an unbalanced dataset as attempted here has been the focus of several recent works [197].

The best combination of features and classifier were explored to get the optimum classification results within the available datasets. This was done in the following way:

- The features for every binary classification setting (OVR and OVO) were ranked using FDR.
- Using these rankings, an SFS algorithm was employed with all five classifiers to find the best feature-classifier combination.

The ranked features for each binary classification settings are provided in Appendix C.

5.4 Raw signals – Results and discussion

Figure 5.3 shows the univariate histogram plots of the features, extracted from the pre-stimulus information subtracted raw data, which helped visualize the class separation and overlaps for each individual feature. It is evident that, in most cases, the individual features did not allow a straightforward separation of the classes.

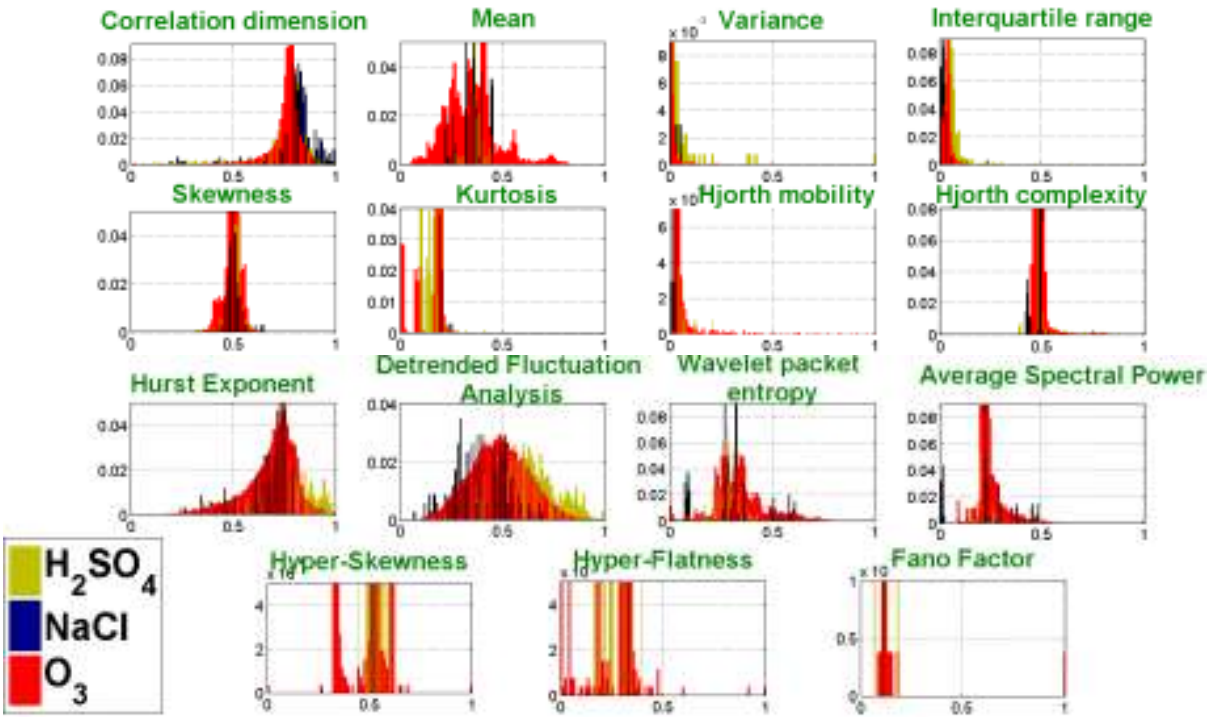


Figure 5.3: Histogram plots for 15 features (computed from raw data), showing overlap of classes

Since the separation of the classes using individual features were not that clear from the histogram plots, an attempt was made at plotting the features on a 3-D LDA Basis space.

5.4.1 Visualization of class separability on the LDA Basis

The LDA basis vectors (*fisherfaces*) [151], [175] were found by solving the generalized eigenvalue problem given by (5.5)

$$S_W^{-1} S_B W = \lambda W \tag{5.5}$$

The terms S_B and S_W are called between-class and within-class scatter matrix respectively, for the original n-dimensional feature vector extracted from the plant signals.

The eigenvectors corresponding to the largest eigenvalue gave the optimal projection when the variance between the class features is maximized. The three dimensions of the LDA Basis were obtained from eigenvectors corresponding to the three largest eigenvalues.

The separability of the three stimuli on 3-D LDA basis using raw signals was plotted in Figure 5.4. The data shown are those which were used for training the classifiers during retrospective study (~73% of total data). The basis vectors, designated as LDA Basis – 1, 2, 3, were obtained by linear combinations of the feature values and weights of the features. Those weights are the elements of the eigenvectors belonging to the three highest eigenvalues, when solving the generalized eigenvalue problem given by (5.5).

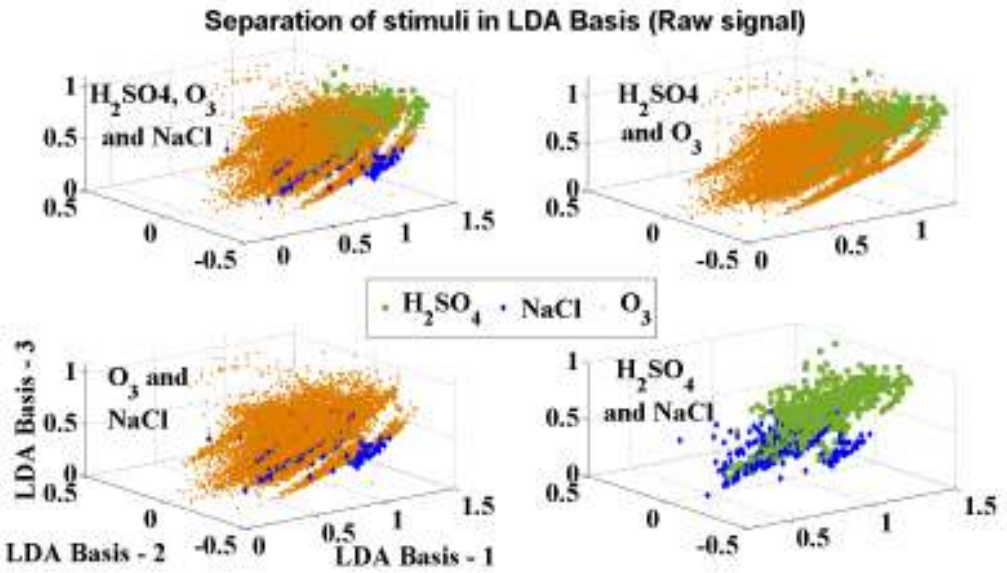


Figure 5.4: LDA Basis showing separation of three stimuli using raw signals

In addition to showing the separability of all three stimuli in one plot, the separability of all possible binary combinations of the three stimuli are also shown. Figure 5.4 shows that H_2SO_4 and NaCl looks to be separated to a good extent. This separation appears to be better than the separation of the other two binary stimuli combinations.

5.4.2 Retrospective study using the raw plant signals

The classification accuracy obtained using the SFS algorithm on ranked features and using all five variants of classifiers are shown in Figure 5.5 (OVR) and in Figure 5.6 (OVO). From these diagrams, the classifier and feature combination giving the best accuracy for each binary classification scenario can be identified. These results are obtained by carrying out *LOOCV* on the ~73% training dataset for the retrospective study. The rank of each feature for each binary classification scenario is given in Appendix C.

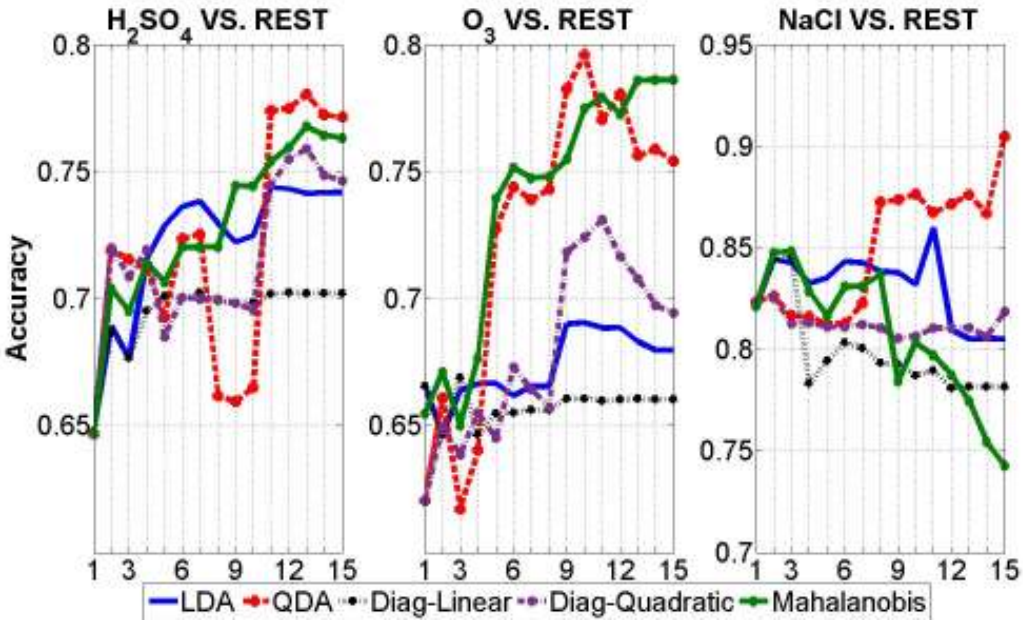


Figure 5.5: Accuracy vs. increment in features (SFS) for OVR setting (using raw signal) – first node setting

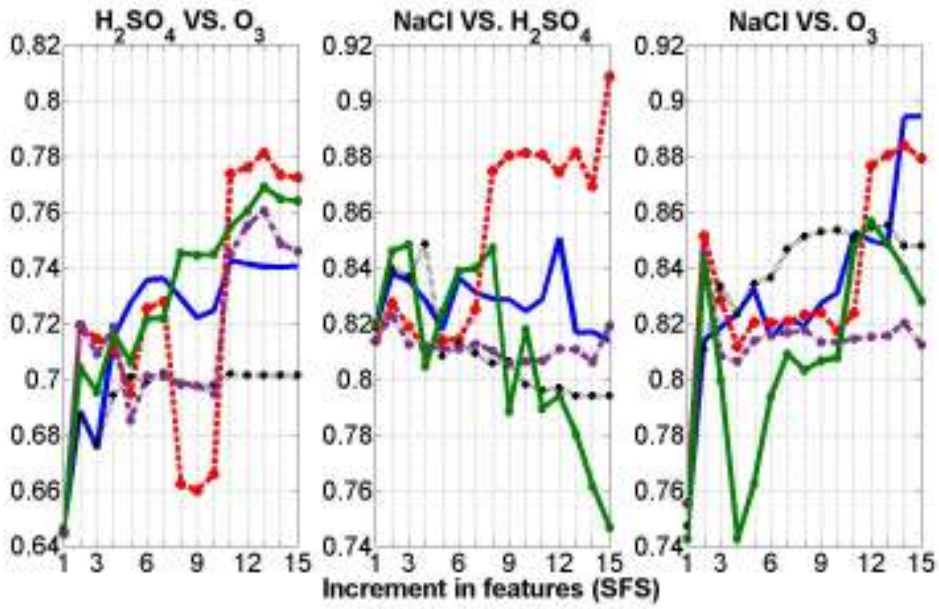


Figure 5.6: Accuracy vs. increment in features (SFS) for OVO setting (using raw signal)

The term "top-*N* features" has been used while discussing the results. By this is meant that by using all ranked features from the highest, i.e. 1st ranked to N^{th} rank. The best results obtained during the retrospective study (~73% dataset, LOOCV) in OVR/OVO setting is given in Table 5.3 from which is determined that NaCl was the best separable class, leaving H_2SO_4, O_3 as the next stimuli to be classified, i.e. Class $\{A = NaCl, B = H_2SO_4, C = O_3\}$ respectively in the OVR setting.

Table 5.3: Best Classification Accuracy for the Raw Signal (Features + Classifier Combinations)

	OVO		**OVR**
Stimuli	**NaCl**	**Ozone**	**Rest**
H_2SO_4	89.43% (top 14 features), *LDA*	78.11% (top 13 features), *QDA*	78.04% (top 13 features), *QDA*
NaCl	-	90.85% (top 15 features), *LDA*	90.47% (top 15 features), *QDA*
Ozone	*	-	79.58% (top 10 features), *QDA*

Table 5.3 demonstrates that, using features extracted from raw signal, ~90% classification accuracy was achieved for *NaCl vs. rest* in OVR setting and ~78% accuracy was achieved for H_2SO_4 *vs.* O_3. For both results, QDA performed the best, along with top 15 and top 13 features for *NaCl vs. rest* and H_2SO_4 *vs.* O_3 respectively.

It can also be seen that *NaCl vs.* H_2SO_4 and *NaCl vs.* O_3 provided classification accuracies of ~89% (using top 14 features) and ~90% (using top 15 features) respectively. Both these results were obtained by using LDA. This possibly suggests that electrical response of the plant due to *NaCl* is statistically dissimilar to electrical response due to O_3 or H_2SO_4.

5.4.3 Constructing the decision tree for prospective study using results from retrospective study

The retrospective study determined that the decision tree in OVR configuration requires the top 15 features (*NaCl vs. Rest*) with a QDA classifier at the first node to test an incoming feature vector as belonging to NaCl. If that vector is not from NaCl, then at the second node, the feature vector is tested for belonging to O_3 or H_2SO_4, using the top 13 features and a QDA classifier. The retrospective study determined that the decision tree in OVO configuration would test an incoming feature vector for three binary classification settings simultaneously. The classifier setting of *NaCl vs.* O_3 requires the top 15 features with an LDA classifier. The classifier setting of *NaCl vs.* H_2SO_4 requires the top 14 features with an LDA classifier. The

classifier setting of *O_3 vs. H_2SO_4* requires the top 13 features with a QDA classifier. These feature-classifier settings were used to design the decision tree for the prospective study.

One experimental dataset from each stimulus was set aside for the independent test dataset for the prospective study. Figure 5.7 shows the test feature matrix, where each set had all the blocks from the stimulus.

Figure 5.7: Test feature matrix for prospective study

These features were extracted from a block of 1024 samples from the post-stimulus section of the signals (with pre-stimulus information subtracted) belonging to the test dataset, as given in Table 5.2.

For each of the two decision tree configurations, OVR and OVO, a decision criterion, based on the value of the discriminant function (in a binary class setting), interpreted which stimuli the features pointed towards. The answers were then measured against the known class labels of the test dataset and thus the accuracy of the configurations determined. The results for each of the three stimuli were collected according to equation (5.6).

$$Accuracy_{prospective} = \frac{(c_a + c_b + c_c)}{(n_a + n_b + n_c)} \tag{5.6}$$

Here $\{c_a, c_b, c_c\}$ denotes the correct number of blocks belonging to each of the three classes detected by the decision tree. The total number of blocks belonging to the three stimuli are denoted by $\{n_a, n_b, n_c\}$ taken from Table 5.2.

However, when the OVR/OVO configurations were tried on unseen data, using the same feature-classifier combinations found in the retrospective study, the results of classifying the three stimuli were poor. These results are provided in Table 5.4 and show that H_2SO_4 and NaCl were poorly detected, while O_3 was mis-classified.

Table 5.4: Prospective study results using best settings obtained from retrospective study

Stimulus	OVR	OVO
	Correctly detected blocks / total blocks tested	
H_2SO_4	2 / 148	2 / 148
NaCl	0 / 276	0 / 276
O_3	10114 / 9692	10114 / 9692

Therefore, the feature-classifier settings which produced good results for both retrospective as well as prospective study were revisited.

5.4.4 Redesigning the Decision Tree for Prospective Study

The results of *prospective* and *retrospective* studies obtained using five different classifiers were now compared using SFS. It was decided to again check the effect of incrementing features (using SFS algorithm) in order to see which features and classifier produces good results for both studies. First, the classifier in every node was kept the same, for both OVR and OVO. The features at every node were also incremented simultaneously. With these settings, the results produced for both retrospective and prospective study are shown in Figure 5.8 and Figure 5.9. Thus, rather than having different classifiers at different nodes of the two decision structures, the same classifiers were kept at every node. The aim was to obtain good results for both retrospective and prospective study.

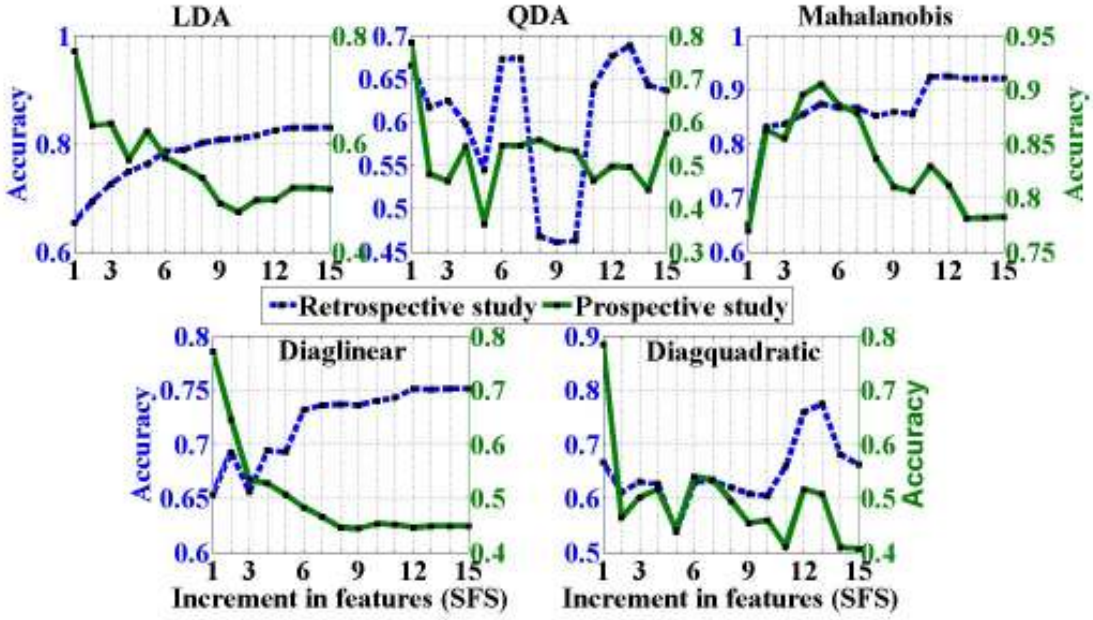


Figure 5.8: Comparison of retrospective and prospective results for OVR configuration, using five different classifiers and SFS

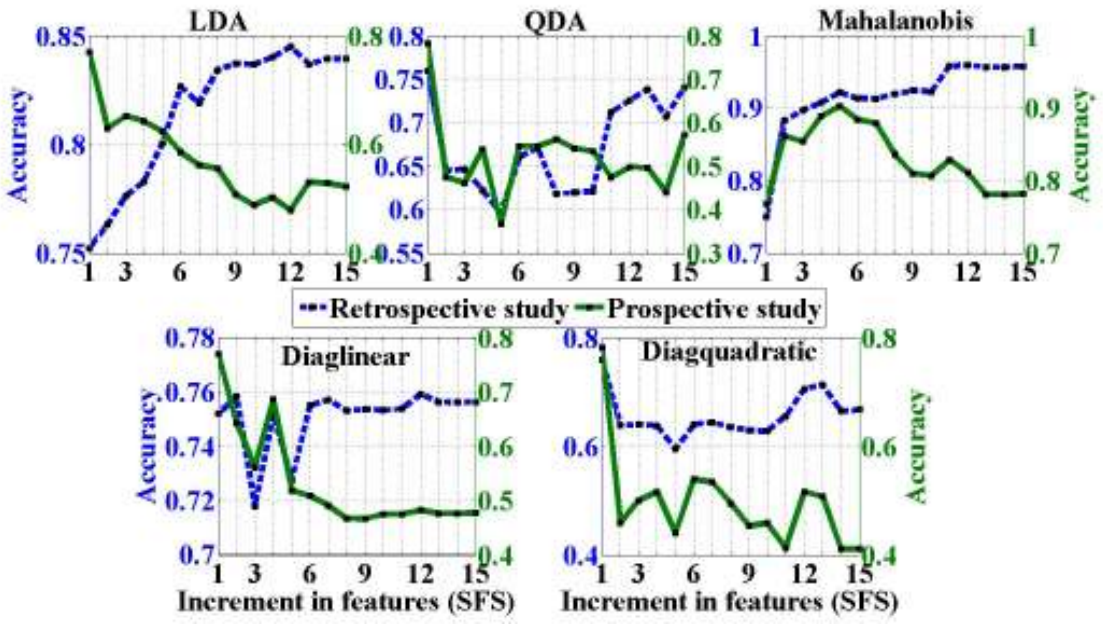


Figure 5.9: Comparison of retrospective and prospective results for OVO configuration, using five different classifiers and SFS

It can be seen that the Mahalanobis classifier in an OVR setting produced the best results for both retrospective (~87%) and prospective (~90%) study, using the top five features. Although the prospective results were around 77% when using the Naïve Bayes (diaglinear and diagquadratic) classifiers with the top feature, the retrospective results were limited to below 70%. The LDA and QDA also showed a mediocre performance, as seen in Figure 5.8.

Likewise in an OVO setting, the Mahalanobis classifier again outperformed others by producing the best results for both retrospective (~92%) and prospective (~90%) studies.

Overall, it was noted that by using the top five features in an OVO setting with Mahalanobis classifier gave the best classification accuracy for both retrospective and prospective studies.

Earlier, a multiclass classification was addressed using two decision tree architectures. The exploration found that the best features-classifier combinations obtained from retrospective study did not work on the prospective study. From the alternate configuration, using the top five features in an OVO setting with the Mahalanobis classifier could achieve good classification accuracy for both retrospective (~92%) and prospective (~90%) studies. These features (given in Appendix C) are *IQR*, *Hyper-flatness*, *Kurtosis*, *DFA*, *Variance*, *Wavelet entropy*, *Hurst exponent*, *Hyper-Skewness*, and *Average spectral power*. A correlation analysis was carried out (given in Appendix C), as done previously, but it found that few features are correlated (e.g. IQR and Variance, DFA and Hurst exponent etc.). Since good

classification accuracy is being obtained despite using these correlated features, they are proposed as suitable for classification. The results are also an improvement over those reported in Chapter 4.

5.5 Analysis of Filtered Signals

This section focuses on the information contained within the stochastic part of the plant electrical signals by applying high pass filtering to the raw signals to remove inconsistent trends or drift. The same set of 15 features was extracted as reported earlier and used for classification. A comparison is presented of the classification performance of the raw and filtered signals by exploring whether there was any improvement in the classification process when using the detrended, random, part rather than the raw signal containing small local fluctuations superimposed on relatively larger change in the trends.

To filter the raw signals, focusing only on the stochastic part, the optimum filter parameters must be determined. Since no prior filtering criteria exist for processing these signals, the optimal filter characteristics for the data available to us had to be explored [198].

5.5.1 Designing an optimum filter for the removal of drift from signals

In general, it is not known which range of frequencies contain useful information about the external stimulus. The aim was to calibrate the pre-stimulus parts (due to different amplitude level for each experiment being different at the onset) and adjust them to a common view, so that changes in the post-stimulus part could be quantified.

Standard analysis of other bio-signals, such as EEG analysis, shows that the vital cognition-related biological responses are generally in the high frequency bands. Thus a search was attempted to find an optimum frequency band for plants' electrical signal responses, based on plant-specific and experimental condition-specific characteristics. The filter was initially designed to have minimum discrimination in the unstimulated signals – even if they came from different plants, channels, etc. It was found that the strong low frequency components, which often contain the drift and artefact (in time domain), were not consistent in different channels and different plants. The presence of these low frequency drifts aid in discriminating the pre-stimulus parts of the signals for different plants and channels as shown in Figure 5.10.

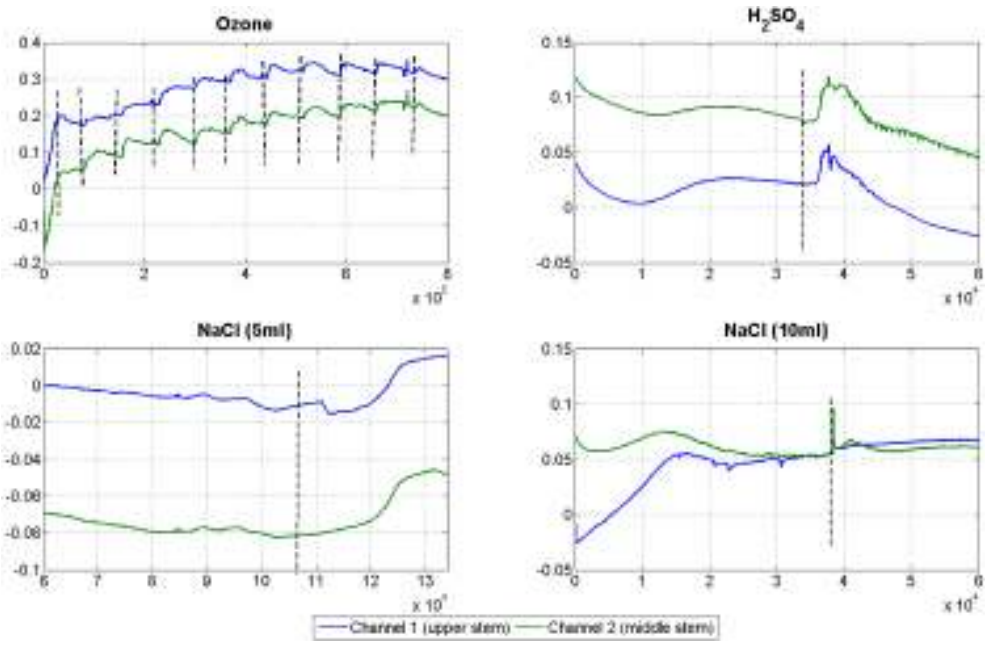


Figure 5.10: Plant response due to four different external stimuli
(vertical lines indicate the stimulus application time)

This made the task of discriminating the stimulated plant responses quite difficult because the true response may often get buried under strong low frequency drifts. These inconsistent low frequency drifts had to be removed to achieve more consistency, so that the optimum frequency bandwidth could be addressed for the identification of the externally applied stimulus. Therefore, the filter had to detrend the raw signals by pre-processing before extracting other informative statistical features from the detrended signal, which might contain high frequency information that were worth preserving [178].

A comparison was carried out of the optimal settings of four classes of digital IIR filter. The approach assumed that some statistical characteristics of the unstimulated plant signal (after filtering) should be similar, e.g. zero mean, uniform variance and energy, as commonly employed for processing other bio-potentials [178]. Ideally, the pre-stimulus part of different signals should form clusters in some statistically meaningful feature space, with their centroids lying as close as possible. Therefore, the IIR filters were designed so that the pre-stimulus part of the signals overlapped and formed a common reference for different plants and recording electrodes. By doing this, any change in the electrical response due to the application of an external stimulus could be analysed as a deviation from this reference or the filtered background signal. Since the signals show nonlinear input-output relationships and strong non-stationary behaviour [116], [142], the use of wavelets was a natural choice for estimating the frequency domain response. The performance of the IIR filters were judged by

an objective function which used non-stationary time-frequency domain decomposition using wavelets. The use of wavelets was not as filters but to provide selection criteria for the IIR filters with much fewer tuning parameters compared to Finite Impulse Response (FIR) filters.

Choice of the filter was crucial, because no prior knowledge of the frequency spectrum for plant's true electrical response exists. Traditionally, band-pass filters were used for most biological signals, like ECG, EEG, EMG [178], to eliminate the effect of low frequency drift/artefact and high frequency measurement noise. As the sampling rate in the experiments is 10 Hz, according to the Nyquist criterion, there is no frequency component above 5 Hz, which is relatively low for biological signal processing [178]. Therefore, instead of removing the spectrum from both sides as in bandpass filter, a high pass filter was chosen since the higher side of the spectrum (5 Hz) was almost insensitive to measurement noise because of the low sampling rate. In contrast, choosing the filter's cut-off frequency at the lower end had more impact on shaping the frequency spectrum and the time domain response of the random part of the signals.

This is especially important since a lower value of the filter cut-off frequency (ω_c) may allow significant amounts of artefact/drift to go into the plant signal, but a higher value of ω_c may remove some significant information from the frequency spectrum. The main challenge was in balancing this trade-off between losing significant information and not allowing low frequency drifts and artefacts to contaminate the spectrum. Figure 5.11 shows, as an example, the effect of applying a digital high-pass Butterworth filter with a cut-off frequency of 1 Hz for two plants with two channels.

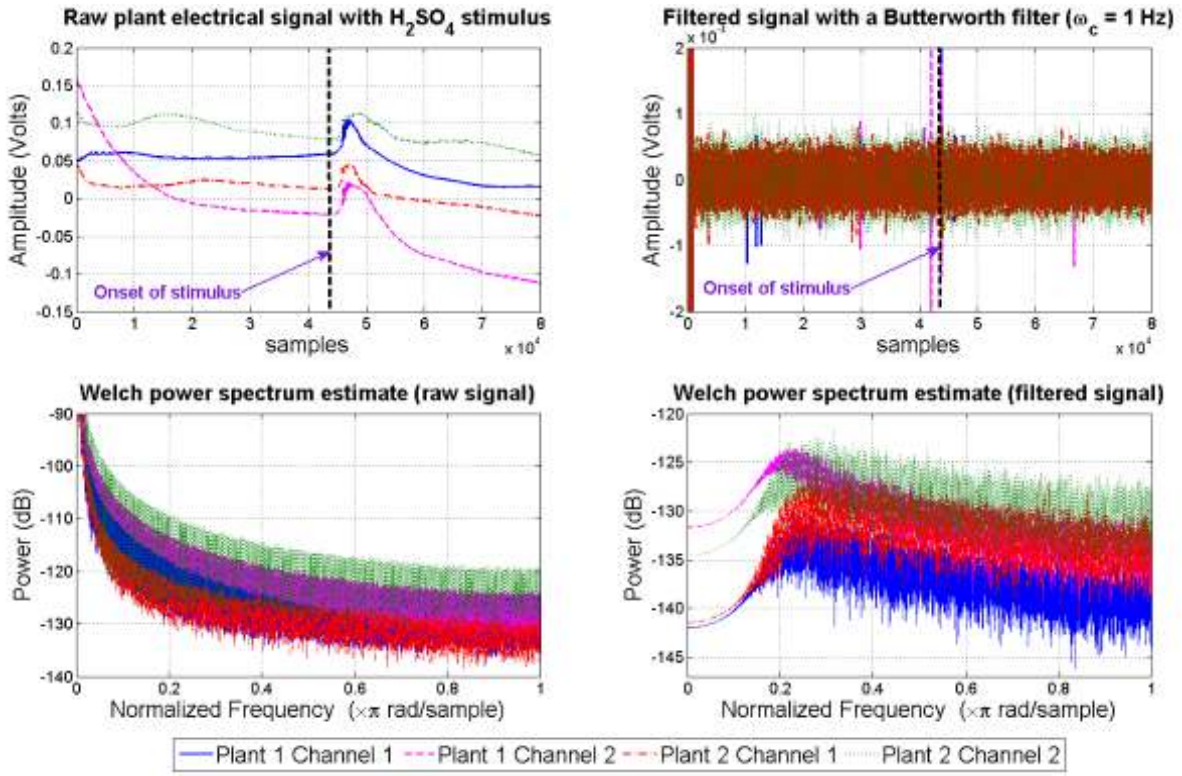


Figure 5.11: Time and frequency domain representation of the raw and filtered signal using a Butterworth filter with ω_c = 1 Hz.

Due to the low frequency drifts present in the time domain, it is evident from Figure 5.11 that the frequency spectrum of the raw signals has high power at low frequencies, which could mask any underlying biological response of the plant that presents at relatively higher frequencies. Clearly, the IIR filter forces the signal to have almost zero mean and uniform variance in both the pre- and post-stimulus part and therefore may help in characterizing the stimulus in terms of other higher order statistical features.

Although the time domain representation looks similar for the filtered signals, their frequency responses were different. Especially the gain and ripple in the low frequency region, cut-off frequency as well as type of digital IIR filter which needed to be optimally tuned, using some criterion (distribution of energy along different wavelet was chosen as a basis for the pre-stimulus signals) ensured similar statistical behaviour of the pre-stimulus part.

The objective function for the optimization-based tuning of IIR filters was chosen as the energy contents of different nodes in *wavelet packet decomposition* for both pre- and post-stimulus parts of the signal, acquired under similar laboratory settings. The IIR filter parameters were tuned so that it produced almost overlapping clusters of the distribution of energy along different wavelet basis for the pre-stimulus signals, but non-overlapping clusters for the post-stimulus parts. The wavelet decomposition of the IIR filtered signal was

restricted up to level 2, to keep the number of basis vectors small (Q = 4 in this case), and to make the analysis consistent and computationally efficient. After applying an IIR filter (which required further optimization), the signal was segmented in smaller non-overlapping windows of M = 256, 512 and 1024 samples. The wavelet energy in all four nodes at level 2 for segmented signals of length M was projected into a 4-dimensional feature space. Some/all of the filter parameters were then optimized, such as filter order, cut-off frequency, passband, and stopband ripples (N, ω_c, R_p and/or R_s) for the four IIR filter variants, so that it enforced the centroids (in the feature space) of the pre-stimulus signals in different experiments (D), to lie as close as possible. The objective function (J) was framed as the sum of the Euclidean distances of the centroids (C_i^d) under different experiments from the mean of all these centroids (μ_{C_i}), as shown in (5.7).

$$J = \sqrt{\sum_{i=1}^{Q}\sum_{d=1}^{D}\left(C_i^d - \mu_{C_i}\right)^2},\ \mu_{C_i} = (1/D)\sum_{d=1}^{D} C_i^d\ \forall i = 1,\cdots,Q,$$
$$C_i^d = (1/M)\sum_{m=1}^{M} E_{im}\ \forall d = 1,\cdots,D. \tag{5.7}$$

The objective function in (5.7) was minimized (using the Nelder-Mead Simplex algorithm in MATLAB) with an initial guess of $\omega_c^0 = 1\,\text{Hz}, N^0 = 7, R_p^0 = 0.5\,\text{dB}, R_s^0 = 80\,\text{dB}$.

Parameters of four IIR) filters were then tuned – *Butterworth*, *Chebyshev type-I* and *II*, and *Elliptic* – to explore which filter structure minimized the distance between the centroids of the clusters, formed using distribution of wavelet packet energy along different wavelet basis, for the pre-stimulus part.

The optimized parameters of the IIR filters are shown in Table 5.5 for the four different structures and three data segmentation sizes (M = 256, 512 and 1024). It was found that the *Chebyshev type-II* filter yielded the minimum cost function (J_{min}) for the 256 non-overlapping samples of data segmentation. For the 512 and 1024 samples of data segmentation, the *Chebyshev type-II* filter outperformed the other three filter structures in minimizing the discrimination of the pre-stimulus information, under different experimental conditions. The ripples in passband and stopband are designated as R_p and R_s respectively.

Table 5.5: Optimum IIR filter settings for different filters

Filter Type	Segment Size	J_{min}	ω_c (Hz)	N	R_p (dB)	R_s (dB)
Butterworth	256	17.48	1.50	4	-	-
	512	19.77	1.50	4	-	-
	1024	24.41	1.50	5	-	-
Chebyshev Type I	256	16.71	1.43	3	1.00	-
	512	19.73	1.37	6	0.96	-
	1024	23.62	1.50	4	1.00	-
Chebyshev Type II	*256*	*11.64*	*0.77*	*6*	-	*100*
	512	*12.55*	0.77	6	-	100
	1024	*13.50*	1.34	6	-	70.19
Elliptic	256	17.58	1.45	6	0.43	60.00
	512	18.61	1.37	4	1.00	60.00
	1024	23.57	1.50	4	1.00	60.00

5.5.2 Pre-processing, feature extraction and classification

The same datasets used in Section 5.3.1 were used again. This time, the signals were filtered using the filter parameters from Table 5.5. About 73% of the total datasets were used for retrospective study (using LOOCV) while the remaining 27% set aside for prospective study. The same datasets were set aside for prospective study as reported earlier, except that they were filtered. Figure 5.12 shows the steps followed for pre-processing and classification.

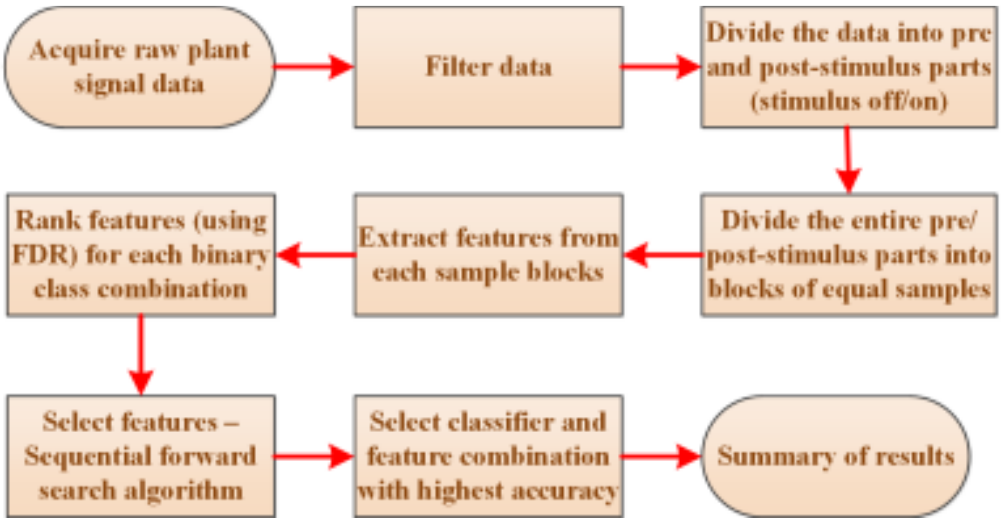


Figure 5.12: Steps for classification of environmental stimuli from plant electrical signal

The entire filtered signal was divided into blocks of 1024 samples. The same set of 15 features as reported earlier, were extracted and these features were then normalized so that their values lay between {0} and {1}.

The classification methodology adopted for filtered signals was the same as for the raw signals. OVR and OVO decision tree structures were constructed, with five classifier variants of *LDA*, *QDA*, *Diaglinear*, *Diagquadratic*, and the *Mahalanobis distance*.

The retrospective study was first addressed to find the best feature-classifier combinations using LOOCV. Thereafter, the same feature-classifier combinations was tested on unseen data.

5.6 Filtered signals – Results and discussion

Figure 5.13 shows the univariate histogram plots of the features, extracted from the filtered data, which helped visualize the class separation and overlaps for individual features. In all cases, the individual feature did not allow a straightforward separation of the classes.

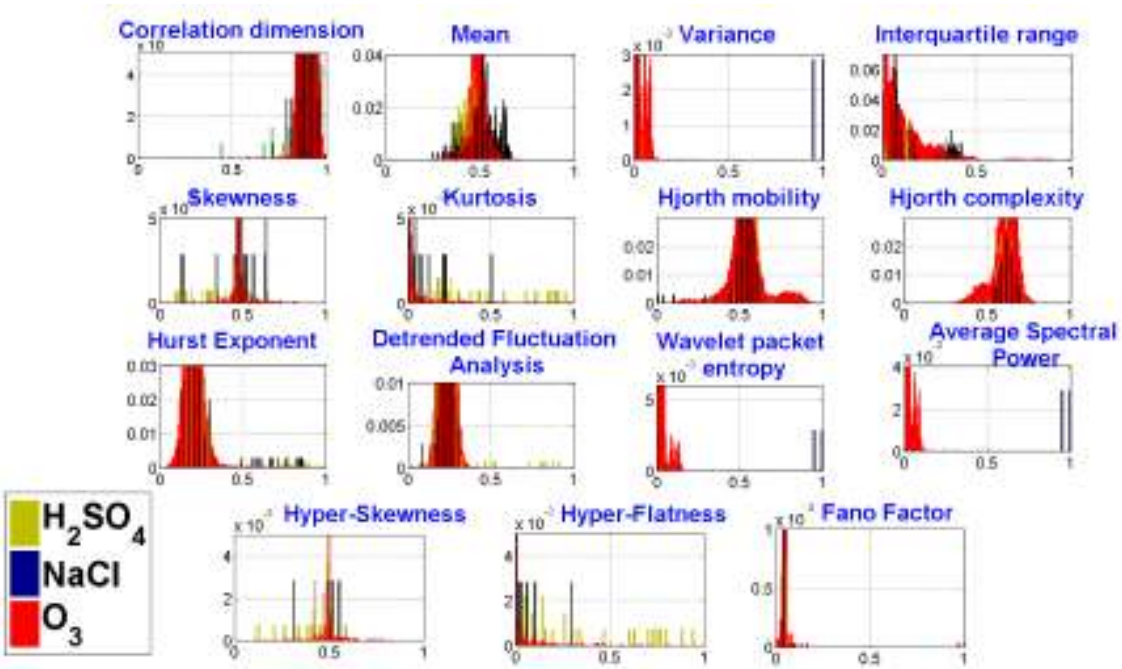


Figure 5.13: Histogram plots for 15 features (computed from filtered data), showing separation of classes

5.6.1 Visualization of class separability on the LDA basis

The separability of the three stimuli in LDA basis was investigated, as shown in Figure 5.14. H_2SO_4 and NaCl looks to be separated to a good extent. This separation also appears to be better than the separation of the other two binary stimuli combinations exactly as found from

the raw signals. However, the separation could only be known quantitatively with the results of the classification.

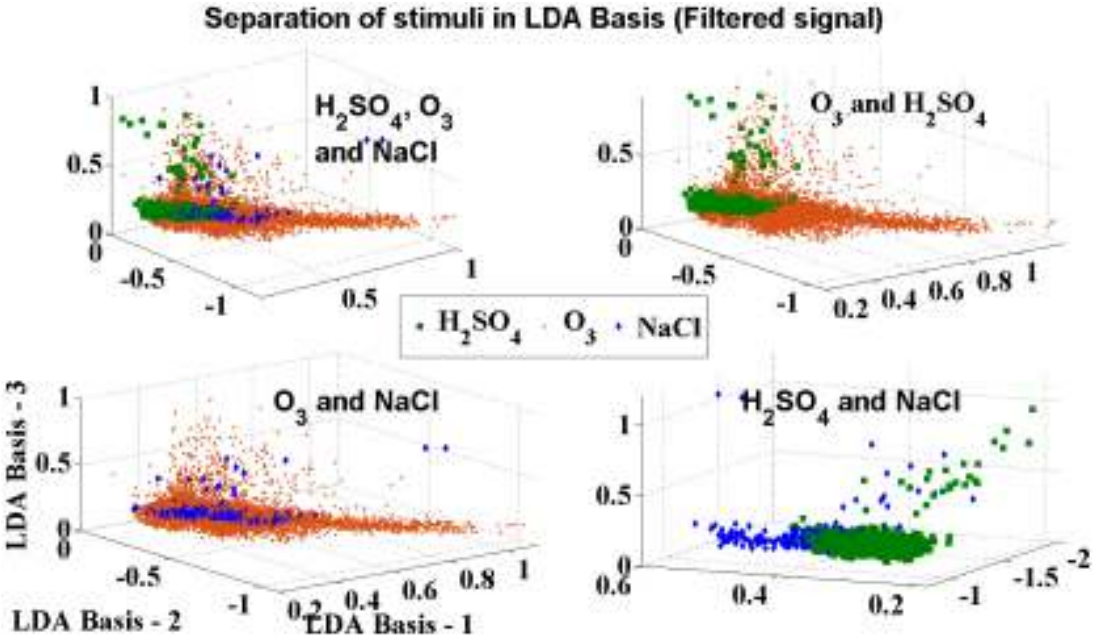


Figure 5.14: LDA Basis showing separation of three stimuli using filtered signal

5.6.2 Retrospective study using filtered plant signals

The classification accuracy obtained using the SFS algorithm on ranked features with all five variants of discriminant analysis classifiers is shown in Figure 5.15 (OVR) and Figure 5.16 (OVO). The classifier and feature combination giving the best accuracy can be easily identified (as the maxima of the curves) for each binary classification scenario. These results were obtained by carrying out *LOOCV* on the 73% training dataset for the retrospective study. Table 5.6 shows the top features and classifiers with the best results for different binary classification scenarios (in percentage accuracy). For every such scenario, the features were ranked using *FDR*, and hence not all scenarios will have the same features at the same ranking. The features, as ranked by FDR, are provided in Appendix C.

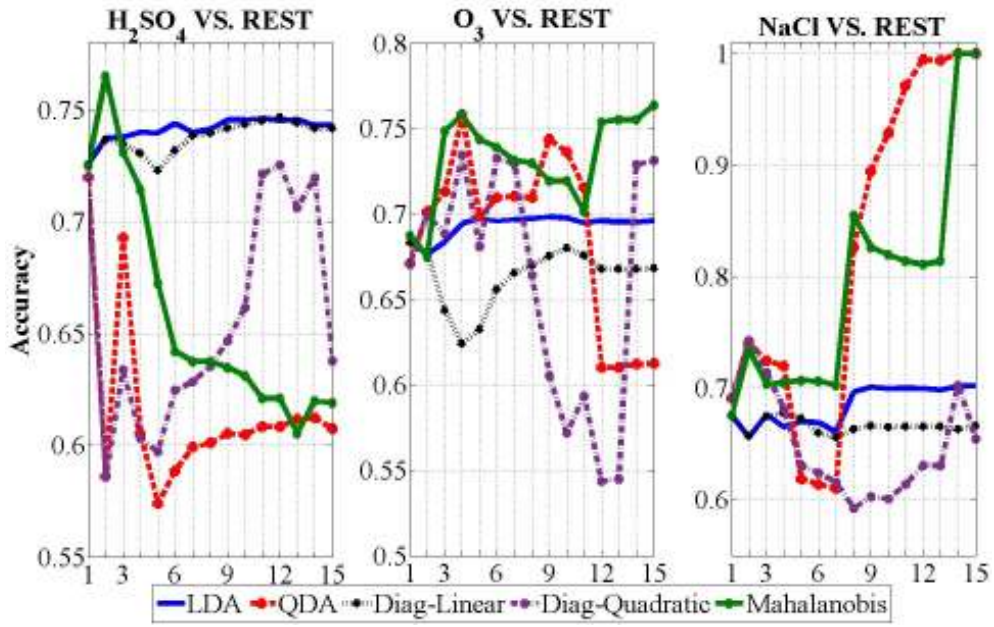


Figure 5.15: Accuracy vs. increment in features (SFS) for OVR setting (using filtered signal)

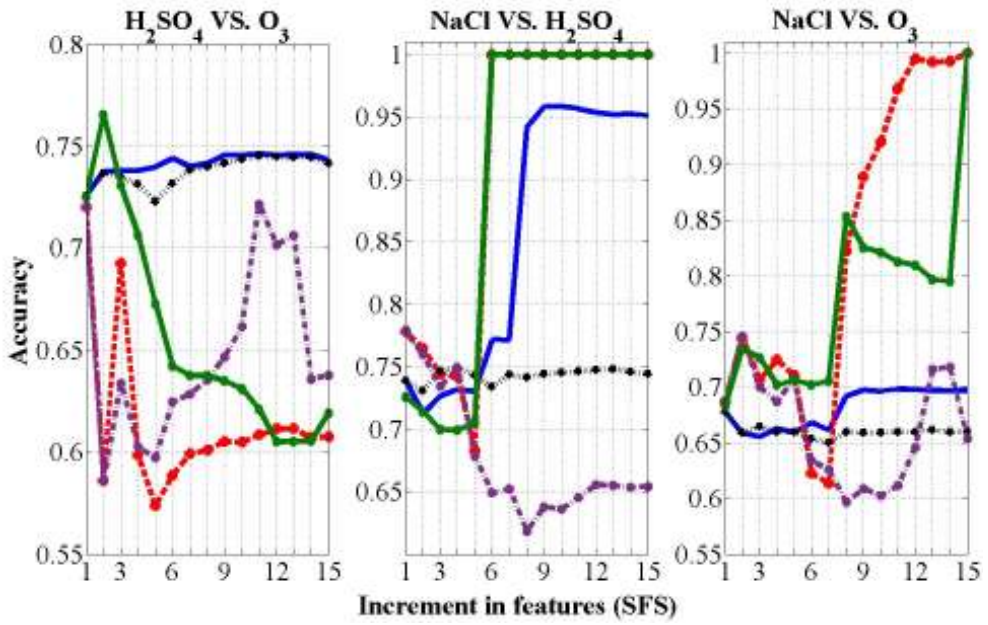


Figure 5.16: Accuracy vs. increment in features (SFS) for OVO setting (using filtered signal)

Table 5.6 shows that 100% classification was achieved between NaCl and H_2SO_4 using the top 6 features and a QDA/Mahalanobis classifier. This result is significant since both stimuli are administered through the soil (and hence the uptake is through the roots).

A similar result was obtained for NaCl and O_3 although using the top 15 features, along with a QDA/Mahalanobis classifier. Here, O_3 was administered by a spray all over the plant, so the absorption was mainly through the leaves. The need for a 15-dimensional feature space to distinguish between the two stimuli probably means that how the stimuli were applied to the

plant is irrelevant, i.e. whether through the root or leaf. However perhaps it also shows that the electrophysiological effect in the plant due to NaCl is different to a greater extent than the other two stimuli as it produced a 100% classification result.

The best result for O_3 and H_2SO_4 was approximately 76%, using the top two features and a Mahalanobis classifier, which was not as good as that obtained using NaCl as stimulus. This observation possibly points to the fact that the electrophysiological effect due to O_3 is not as separable from the effect of H_2SO_4.

Table 5.6: Best Classification Accuracy for the Filtered Signal

Scheme	**OVO**		**OVR**
Stimulus	**NaCl**	**O_3**	**Rest**
H_2SO_4	100% (top 6 features), *QDA* / *Mahalanobis*	76.53% (top 2 features), *Mahalanobis*	76.53% (top 2 features), *Mahalanobis*
NaCl	-	100% (top 15 features), *QDA* / *Mahalanobis*	100% (top 14 features), *QDA* / *Mahalanobis*
O_3	*	-	76.34% (top 15 features), *Mahalanobis*

5.6.3 Constructing the decision tree for the prospective study using results from the retrospective study

Table 5.6 shows that *NaCl vs. Rest* gave the best classification accuracy of 100%, using the top 14 features (as ranked using FDR) along with a *QDA/Mahalanobis distance* classifier. So, in the first node of OVR, class *A = NaCl* is the best separable stimulus from the rest. Hence, the decision tree will first test if an incoming feature vector belongs to NaCl or not. If it is tested to be *not being* NaCl, then the remaining binary classification to be evaluated in the second node would be *O_3 vs. H_2SO_4* (= *class B* and *class C* respectively in OVR structure) using the top 2 features and a Mahalanobis distance classifier.

In OVO configuration, three binary classifier settings were used. As can be seen from Table 5.6, *NaCl vs. O_3* and *NaCl vs. H_2SO_4* achieved 100% accuracy using the top 6 and top 15 features respectively, with either a QDA or Mahalanobis distance classifier. The *O_3 vs. H_2SO_4* achieved the best result of 76.53% using the top 2 features with a Mahalanobis distance classifier.

Following the retrospective study, it was noted that the decision tree in OVO configuration will test an incoming feature vector for three binary classification settings simultaneously.

The classifier setting of *NaCl vs.* O_3 would require the top 15 features with a QDA/Mahalanobis classifier. The next classifier setting of *NaCl vs.* H_2SO_4 would require the top 6 features with a QDA/Mahalanobis classifier. The last classifier setting for O_3 *vs.* H_2SO_4 would require the top 2 features with a Mahalanobis classifier. The voting of any two classifiers will confirm the presence of a particular stimulus. The ranked features are given in Appendix C.

The test feature matrix for the prospective study was exactly as previously. When the OVR/OVO configurations using the results from the retrospective study were tried on unseen data, the results were poor at classifying the three stimuli. These results are provided in Table 5.7. Observe that H_2SO_4 and O_3 were poorly detected and NaCl was mis-classified as over detected.

These results also included any variation of classifier found by the retrospective study (e.g. in Table 5.6, both *QDA* and *Mahalanobis distance* classifiers were responsible for finding the best results for *NaCl vs* O_3. Hence both these classifiers were tried for prospective study).

Table 5.7: Prospective study results using best settings obtained from retrospective study

Stimulus	**OVR**	**OVO**
	Correctly detected blocks / total blocks tested	
H_2SO_4	5 / 148	0 / 148
NaCl	10109 / 276	10114 / 276
O_3	2 / 9692	2 / 9692

5.6.4 Redesigning the decision tree for the prospective study

The best settings obtained during the retrospective study did not yield good results during the prospective study. Hence, the aimed was still to find those feature-classifier settings (for a decision tree architecture) which not only produced good results on known datasets, but also on unseen datasets. This would give a realistic overview of how good (or bad) the feature-classifier combination was. In order to proceed further, the two decision trees were tested on both retrospective and prospective study in the following manner:

- Keep the same classifier variant in every node (two nodes for OVR, three nodes for OVO)
- Increment the feature at every node simultaneously, using a *Sequential forward search* (SFS) method

As with the case for raw signals, SFS was employed and the correctly classified blocks from each of the three stimuli noted for each of the classifier variants. The accuracy was computed using the same formula as before.

The results obtained for both retrospective and prospective study in OVR configuration are shown in Figure 5.17. It is clear that using the top 3 to the top 7 features with a Mahalanobis distance classifier produced classification results better than 82% in the prospective study. Similarly, using the top 3 and above features with a Mahalanobis classifier produced a classification result better than 92% in the retrospective study. The best results were obtained using the top 4 features with a Mahalanobis distance classifier, which yielded ~93% during retrospective and ~89% during prospective study.

Among the Naïve Bayes classifiers, the diagquadratic classifier performed well using the top feature only, providing accuracy greater than 82% in both retrospective and prospective studies, whereas the diaglinear classifier provided only a mediocre performance of around 59% during the prospective study. Among other classifiers, QDA provided an accuracy better than 82% in retrospective and 83% in prospective study by using the top feature, whereas LDA produced a classification accuracy better than 80% and 59% in retrospective and prospective studies respectively.

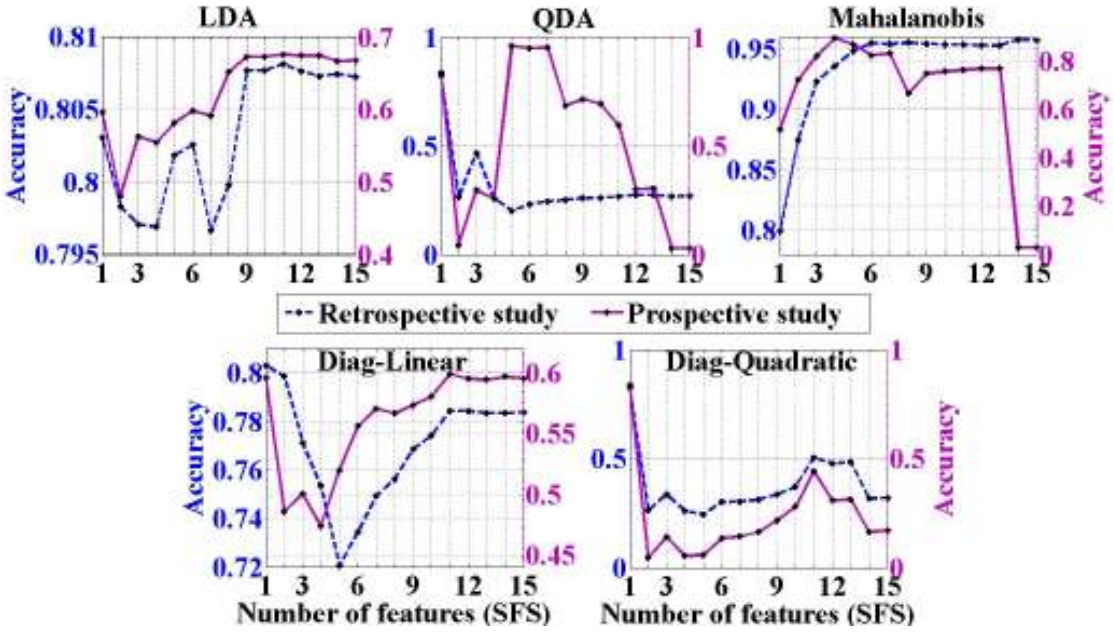


Figure 5.17: Retrospective vs. prospective study results for OVR configuration, using five different classifiers and SFS

The results obtained for both retrospective and prospective study in OVO configuration are shown in Figure 5.18. The best results for classification accuracy were better than 88% in retrospective and 83% in prospective study, and was obtained by using the top 5 features and a Mahalanobis classifier. The LDA produced the best classification accuracy of ~86% during the retrospective study and ~67% in the prospective study using the top eleven 11. The QDA produced a classification accuracy of ~85% and ~83% during retrospective and prospective studies respectively, using only the top feature. Among the Naïve Bayes classifiers, the diagquadratic classifier again performed well using only the top feature, providing accuracy of around 85% in the retrospective and ~83% in the prospective studies, whereas the diaglinear classifier provided a mediocre performance of around 60% in the prospective study and somewhat acceptable performance of ~75% in the retrospective study.

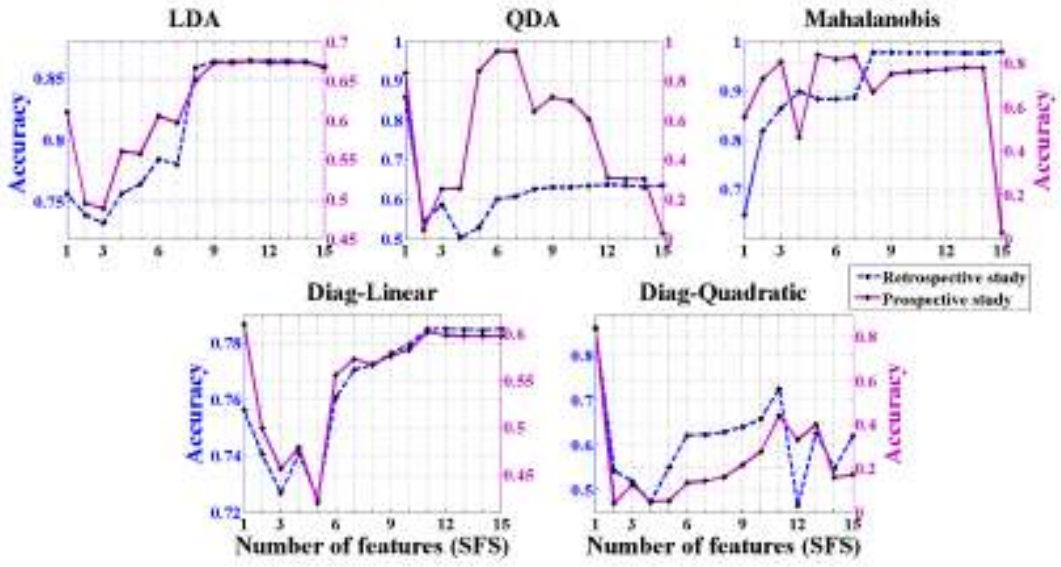


Figure 5.18: Retrospective vs. prospective study results for OVO configuration, using five different classifiers and SFS

In summary, using the top four features extracted from the stochastic part of the plant electrical response, with a Mahalanobis classifier, provided the best multiclass classification accuracy of ~93% in the retrospective study and ~89% in the prospective study in an OVR setting. The detailed list of features is provided in Appendix C.

5.7 Comparison of classification results between raw and filtered signals

The exploration using raw signals found that the best results were ~92% during the retrospective study and ~90% during the prospective study which were obtained using the top

five features with a Mahalanobis distance classifier in every node of a OVO structure. The features included *IQR*, *Hyper-flatness*, *Kurtosis*, *DFA*, *Variance*, *Hyper-skewness*, *Wavelet entropy*, *Hurst exponent*, and *Average spectral power*.

The exploration using filtered signals, where the same features were extracted from filtered signals, found the best results obtained were ~93% during the retrospective study and ~89% during the prospective study using top four features with a Mahalanobis distance classifier in an OVR setting. These features included *Mean*, *Wavelet entropy*, *IQR*, *Hjorth complexity*, *DFA*, and *Variance* (see Appendix C).

Although the classification results achieved using features extracted from raw signals were better than those extracted from filtered signals, the difference was only marginal. It is possible that the marginal increment arose from raw signal having the trend intact, which maybe added more information (captured by the features) to the classifiers. The results also showed that the trained models were not over-fitted to the training data and were generalized well enough to produce good results even on unseen data.

5.8 Summary

This chapter investigated the design of a decision tree classifier employing five different classifiers and 15 statistical features extracted from filtered plant electrical signals. Two multiclass strategies were used, OVO and OVR, along with *retrospective* and *prospective* testing to establish its generalization capability. Both raw and filtered signals were used for analysis and found that NaCl in general was the best separable stimulus compared to O_3 and H_2SO_4. Future work would be the implementation of the decision tree algorithm and the statistical features in hardware.

In both Chapter 4 and Chapter 5, the available time series were windowed and the segments were used for classification. With such segments, around 90% classification was achieved which was independently validated by using separate retained datasets (prospective study). It is possible that this windowing could lose some information and correlation, which might aid better classification.

6 Extraction of features for classification by considering the entire time series with trend

6.1 Introduction

Chapter 4 used segmentation methodology to segment the plant electrical signal response into windows of fixed length of 1000 samples to increase the number of data points with which various classifiers were trained for binary classification. This was reasonably successful provided classification accuracy of around 70%. Chapter 5 addressed multiclass classification using a custom designed decision tree. The same segmentation methodology was used on raw and filtered datasets and obtained an improved classification accuracy of around 88%. The raw signals provided better results than the filtered signals, possibly due the presence of trends in the raw signal which provided more information about the applied stimuli.

However, the windows were assumed to be independent and identically distributed (IID) samples, with no correlation between the segments because of the pre-processing steps which were carried out (subtraction of background or filtering). It is now necessary to consider whether some information could have been lost due to the segmentation method. This information, in the form of correlation between the samples throughout the entire duration of the time series, can possibly provide more information to the classifiers and hence improve the classification accuracy. The nature of the trend present in the signal could also be different for different stimuli. In other words, the shape of the plant's electrical response could vary due to different stimuli applied to the plants.

This chapter explores the possibility of performing classification using features which best capture the shape or the morphology of the raw plant electrical signals. Since the entire duration of the time series will be used, any long-range correlation present within the samples is not discarded.

To capture the morphology of the raw signals, only the post-stimulus parts were considered for feature extraction to see how the shape of the signal changes following application of the stimuli. One of the best ways to capture this trend or shape of the signal is through curve-fitting functions whose coefficients can describe the trend accurately. These coefficients can be used as the features for classification.

For continuity, the same set of classification algorithms, i.e. LDA, QDA, Diaglinear, Diagquadratic, and Mahalanobis distance, can be used with the curve fit coefficients as features.

A similar approach demonstrated the use of a Mahalanobis distance classifier for classification of object trajectory-based video motion clips, assuming that the clusters of the trajectory points are distributed normally in the coefficient feature space [199].

6.2 Methodology

Initially, the curve fit coefficients from four different models were used, i.e. Polynomial, Gaussian, Fourier, and Exponential, to capture the dynamics of the trend for the entire duration of the raw signal (post-stimulus part only) to the three stimuli, NaCl, H_2SO_4, and O_3. Thereafter, these coefficients were treated as *features* and normalized as previously. However, no ranking of features was required as the coefficients were taken together and not individually, which will be explained in detail later. Using the normalized coefficients as features, three sets of binary classifications were carried out to evaluate the results.

To ensure that the coefficients correctly captured the morphology of the signal, an R-squared value difference between the simulated response and actual response was calculated. A high R-squared value means that the simulated curve fit function is accurately following the trend present within the actual signal.

Figure 6.1 shows the methodology of binary classification using curve fit coefficients.

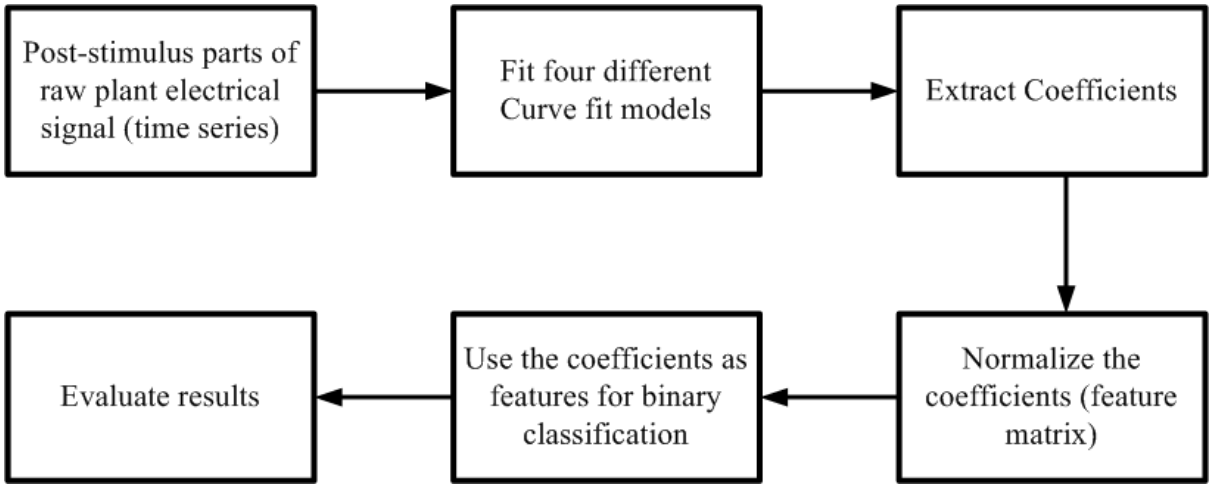


Figure 6.1: Classification using Curve fit coefficients

The reason for using binary classification was to enable a customised OVO decision tree to be designed, exactly as in Chapter 5, in order to carry out a prospective study on separate retained test datasets.

Figure 6.2 shows the four different curve fit functions chosen. For each of these curve fit types, a systematic variation in degree/order/terms was explored for extracting the coefficients, which were then used for classification of the stimuli. The range of variation of the parameters is given in Table 6.1.

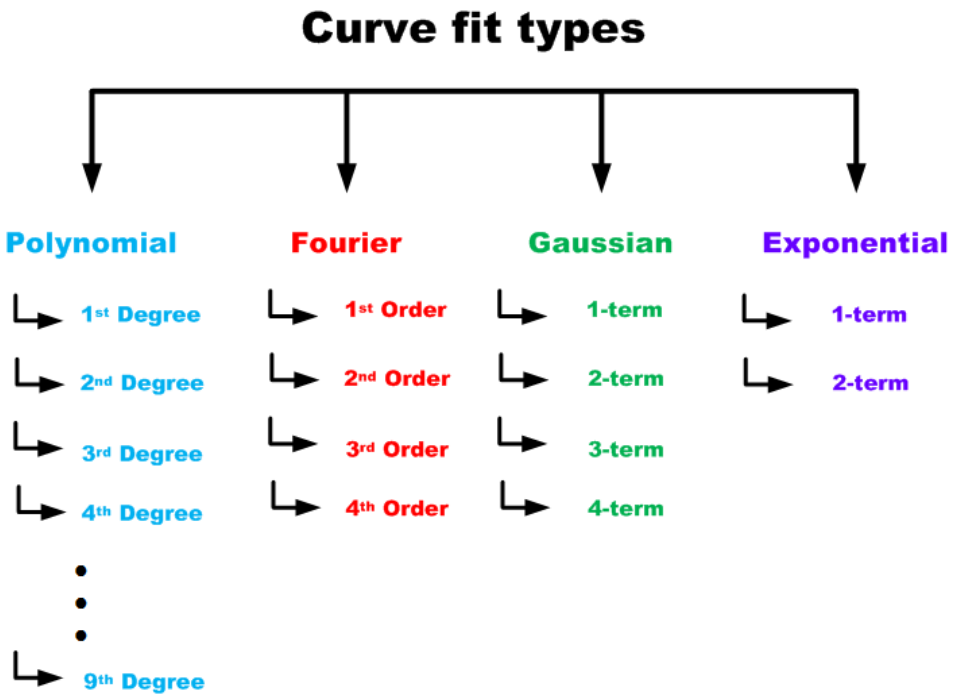


Figure 6.2: Four different curve fit types used to explore the coefficients as features for classification

6.2.1 Polynomial Curve Fit

Polynomial models for curve fits are given by equation (6.1), where n is the degree of the polynomial to be fitted to the time series, which results in $n+1$ order (number of coefficients) of the polynomial function. Here, the number of coefficients (feature space) is limited to 10 so $1 \leq n \leq 9$, since the number of time series were limited (total time series used were 412). This was done to avoid any form of over-fitting while training the classifiers [151].

$$\hat{y} = \sum_{i=1}^{n+1} P_i x^{n+1-i} \tag{6.1}$$

Polynomial fits are usually used for simple empirical models, especially for interpolation and extrapolation of data. Here, it was used for characterizing the data using a global fit. One of the main disadvantages of polynomial fitting is that the fit might be very good within a data range, but can diverge outside this range. Since the coefficients were used as features across different time series and they were used for classification, a good generic data fit was possible.

6.2.2 Gaussian Curve Fit

The Gaussian model fits peaks of the data and is defined by (6.2).

$$\hat{y} = \sum_{i=1}^{n} a_i e^{\left[-\left(\frac{x-b_i}{c_i}\right)^2\right]} \tag{6.2}$$

where a is the amplitude, b is the centroid/location of the data, c is related to the peak width and n is the number of peaks to fit. Here, $1 \le n \le 4$ to keep the number of coefficients down due to the limited number of time series available for classification.

6.2.3 Fourier Curve Fit

The Fourier series is a sum of sine and cosine functions which is used to describe a periodic signal, as described by (6.3).

$$\hat{y} = a_0 + \sum_{i=1}^{n} a_i \cos\left(j\omega x\right) + b_i \sin\left(j\omega x\right) \tag{6.3}$$

where a_0 models the intercept term in the data (which is a constant) and is associated with $i = 0$ in the cosine term. ω is the fundamental frequency of the signal and n is the number of terms (harmonics) used to fit the time series. Here, $1 \le n \le 4$ to keep the number of coefficients down due to the limited number of time series available for classification.

6.2.4 Exponential Model Curve Fit

The MATLAB curve fit capability provides one- or two-term exponential models which are given by (6.4).

$$\begin{cases} \hat{y} = ae^{bx} \\ \hat{y} = ae^{bx} + ce^{dx} \end{cases} \tag{6.4}$$

Exponential functions are usually used to define time-series when the rate of change of a quantity is proportional to the initial value of that quantity. If the coefficients *b* or *d* are negative, then the equation represents an exponential decay.

Table 6.1: Curve fit types and parameters

Fit type	Degree / No. of terms
Polynomial	$1 \leq n \leq 9$
Gaussian	$1 \leq n \leq 4$
Fourier	$1 \leq n \leq 4$
Exponential	$1 \leq n \leq 2$

Curve fiting was carried out using a custom MATLAB script, which executed the curve fit type and directly input the coefficients into dedicated Excel spreadsheets. This script is given in Appendix D.

6.3 Experimental datasets

The number of time series comprising the post-stimulus parts of the signals are given in Table 6.2. Few of these experimental datasets were new and not used in previous chapters, and are shown in column 3. Data from each channel was considered as a separate time series (two channels per plant were used for each experiment). Also, each experiment consisted of multiple exposure when using the O_3 stimulus. The signal response between two successive stimuli were considered as a separate time series as sufficient time (approx. 60 min) was allowed between two successive exposures.

Table 6.2: Number of time series used for each stimulus

Stimulus	Total number of time series	Number of new datasets included in total
NaCl	16	0
H_2SO_4	52	6
O_3	343	180

The entire duration of the post-stimulus part for each time series were used to fit the four different models.

6.4 Results and Discussion

This section presents the classification results obtained during the retrospective study (i.e. using LOOCV) for three different binary stimuli combinations. However, it is also necessary to look at the *Goodness of fit* of the various curve fit types, in order to understand what kind of result for classification can be expected. These are shown using box-plots of the *R-squared* values of the fit for the types of curves.

R-squared is used as a statistical measure to find out how close the actual data is to the fitted regression line. By definition,

$$R_{squared} = \frac{Explained\ variation}{Total\ variation} \tag{6.5}$$

and its value always ranges between 0 and 1 (i.e. 100%). A value of 0 indicates that the chosen model cannot explain any of the variability of the data centred on its mean, while 1 indicates that the chosen model explains all the variability of the data. As a consequence, a higher Rsquared value translates to a better fit for the model to the data [200]. The Rsquared values for each curve fitting type for the three stimuli are shown using Box-Plots to get a good overview of the *Goodness of fit*, as shown in Figure 6.3 to Figure 6.6. This way, any classification results can be related to the accuracy of the curve fit coefficients in capturing the morphology of the signal. Figure 6.3 shows the R-squared values for the Polynomial curve fit, ranging from 1st to 9th degree, for the three different stimuli.

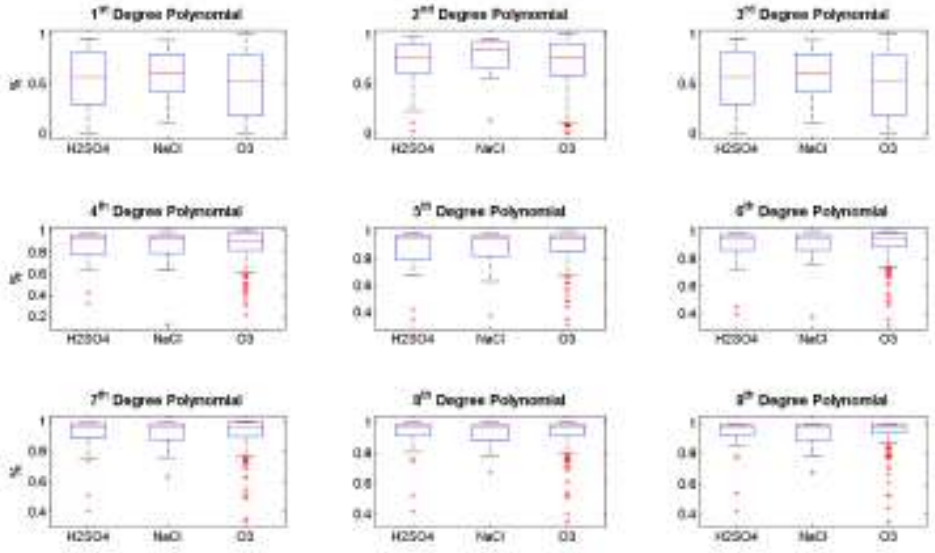


Figure 6.3: R-squared values for Polynomial curve fitting

Figure 6.3 clearly shows that for Polynomial degree 1 and 3, the median of the R-squared values lie around 0.5. For degree 2, although the median is higher than 0.5, the range of lowest to highest values (shown by the whiskers) are large for H_2SO_4 and O_3. This spread reduces from Polynomial 4th degree onwards, but also results in significant outliers for O_3 in particular.

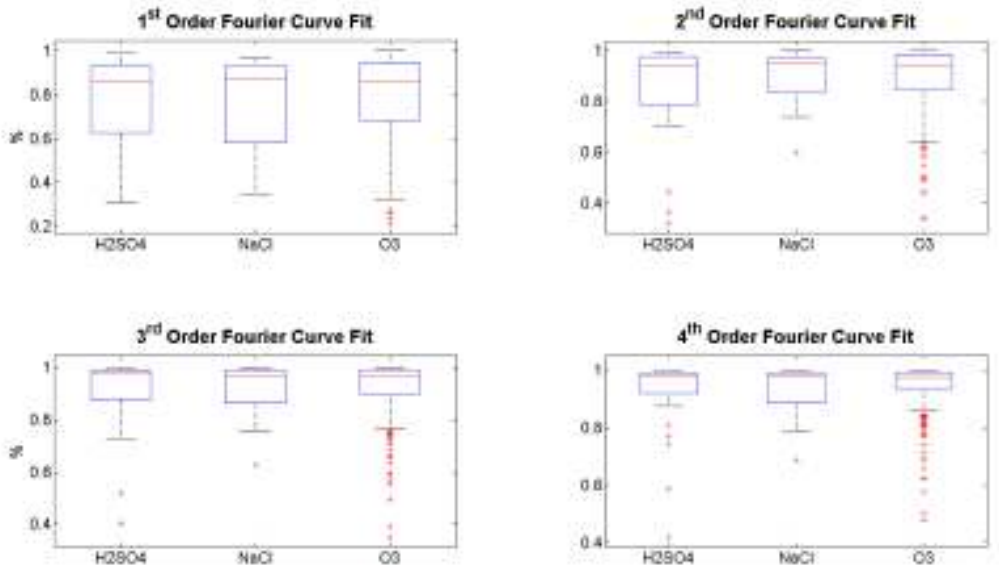


Figure 6.4: R-squared values for Fourier curve fitting

The R-squared values for Fourier curve fits in Figure 6.4 show that the median is above 0.8 with the interquartile range getting narrower as the order increments, denoting consistent fit. However the outliers are more visible for O_3 as the order increases.

For Gaussian and Exponential curve fits, the R-squared values as seen from Figure 6.5 and Figure 6.6 respectively, lie very close to 0 with incrementing terms. The negative R-squared values indicate that the chosen model fits worse than a horizontal line, and is therefore the wrong model to fit the data and does not follow the trend.

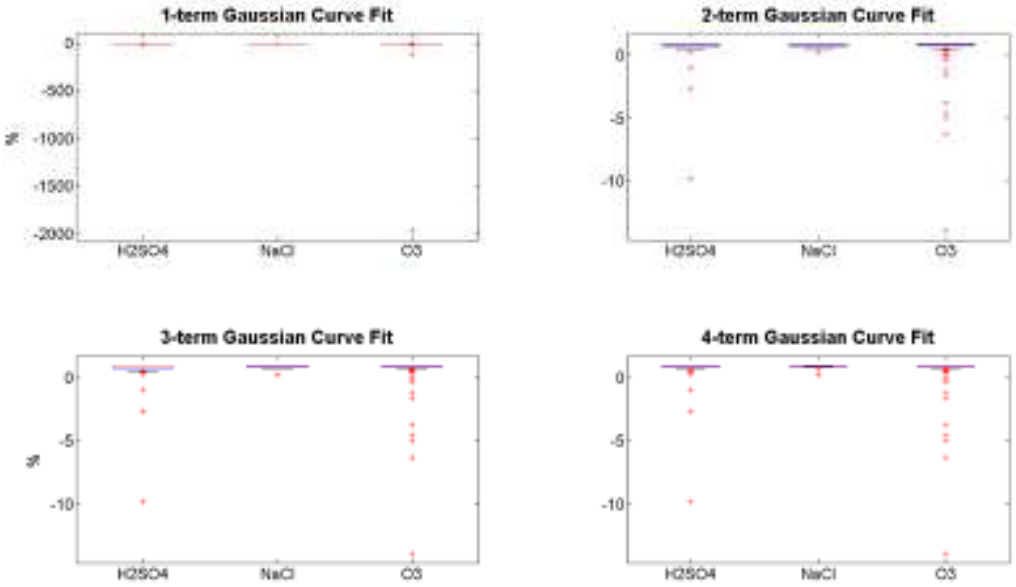


Figure 6.5: R-squared values for Gaussian curve fitting

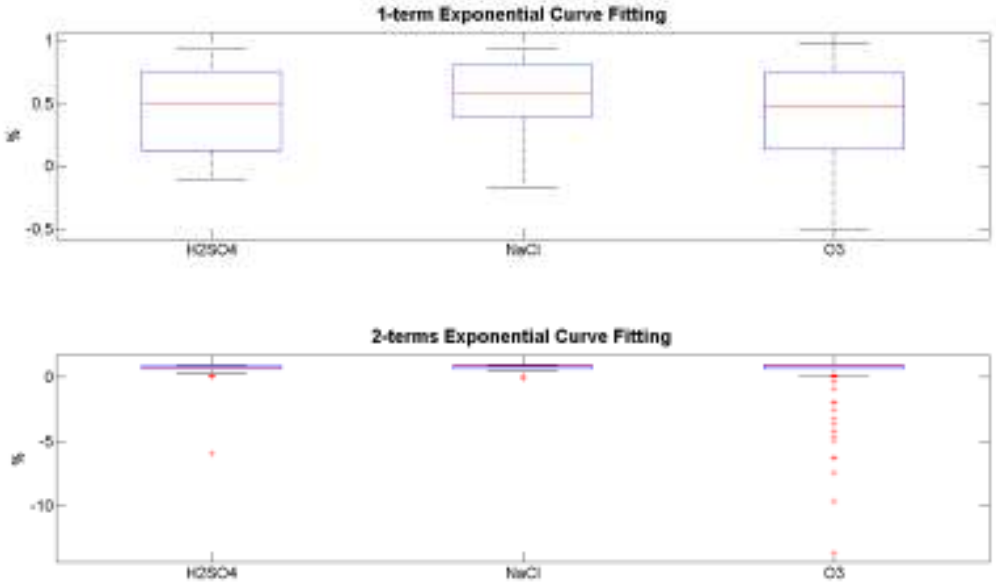


Figure 6.6: R-squared values for Exponential curve fitting

The binary classification results (three combinations using three stimuli) are now addressed, using each of the curve fit coefficients as features for classification. The results need to be evaluated by looking at Sensitivity, Specificity and Accuracy together (defined previously), to get a feel for the performance of the classifiers in detecting both classes (i.e. stimuli) rather than just Accuracy which can be artificially high if one of the two classes are misclassified

[151]. The Polynomial curve fit coefficients are shown in Figure 6.7 (a, b, c). Each figure shows the variation in degree of the curve fit type and for each classifier type (five variants). The figures also show the plots of the Accuracy, Sensitivity, Specificity, PPV, and NPV.

The best classification results were obtained by using Polynomial coefficients of degree 5 and above. These are consistent for all five classifiers used, although LDA/QDA outperforms other classifiers with a higher value for Sensitivity, Specificity and Accuracy.

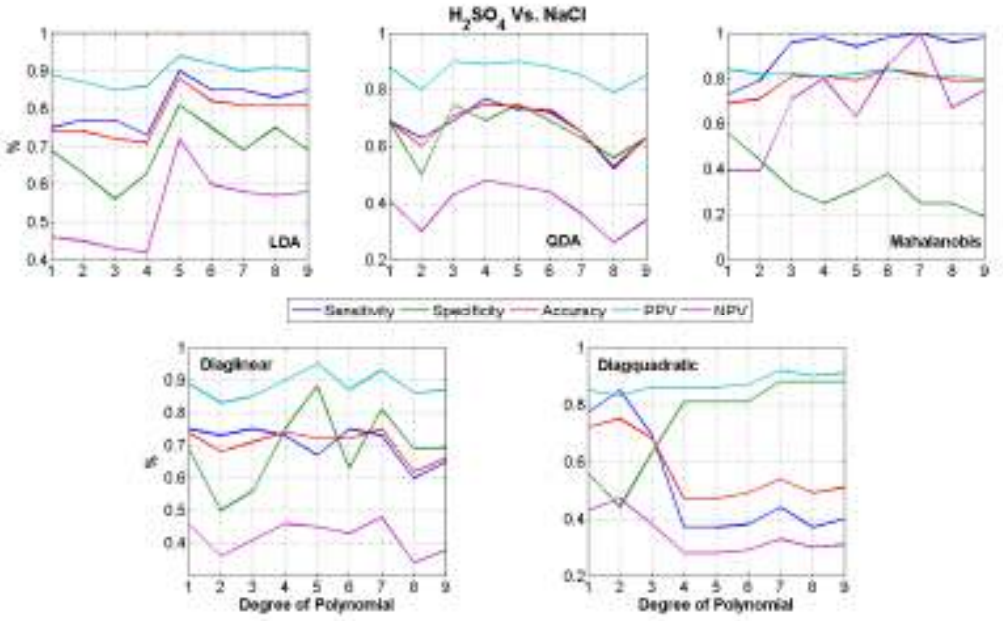


Figure 6.7 (a): Binary classification results using Polynomial Curve fit Coefficients

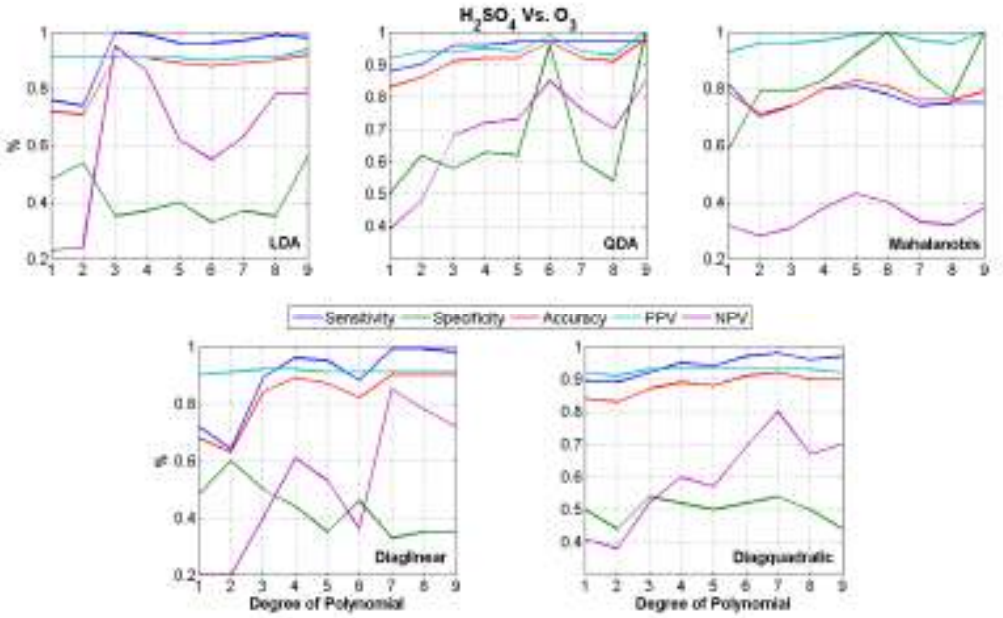


Figure 6.7 (b): Binary classification results using Polynomial Curve fit Coefficients

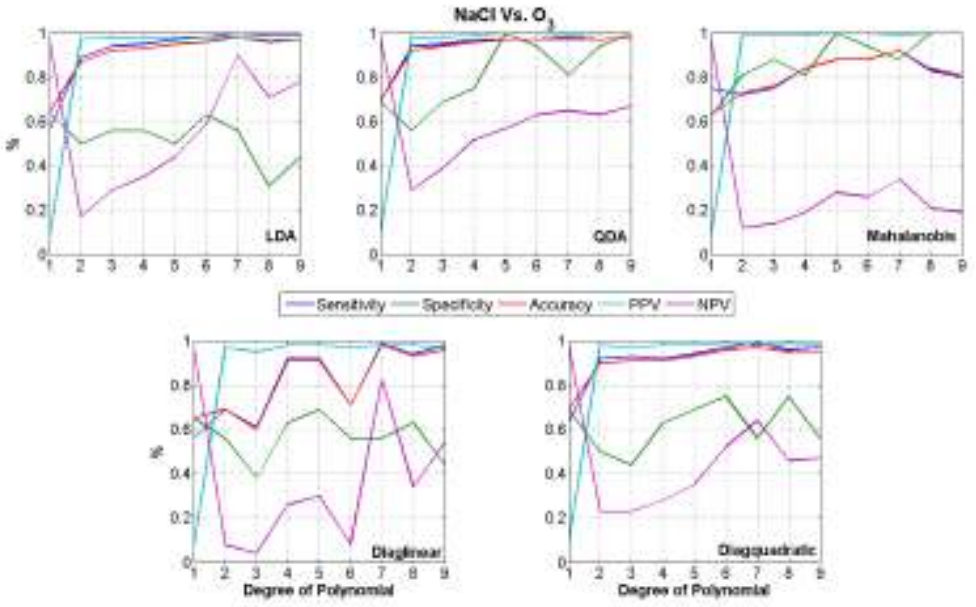


Figure 6.7 (c): Binary classification results using Polynomial Curve fit Coefficients

Figure 6.8 (a,b,c) shows the classification results using the coefficients of Fourier curve fit using different orders. Some of these parameters exhibit a higher value for 3rd or 4th orders, although the results were not as consistent as for the Polynomial curve fits. That is, the accuracy was found to be high in a few cases but either sensitivity or specificity were low. This signifies that classification of all the three stimuli would be difficult using just Fourier coefficients.

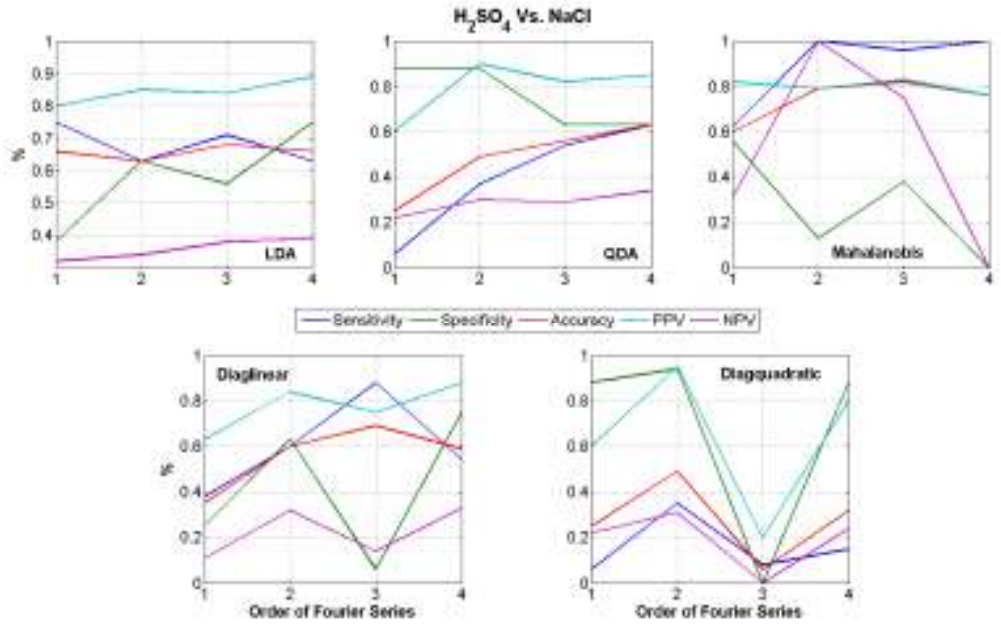


Figure 6.8 (a): Binary classification results using Fourier Curve fit Coefficients

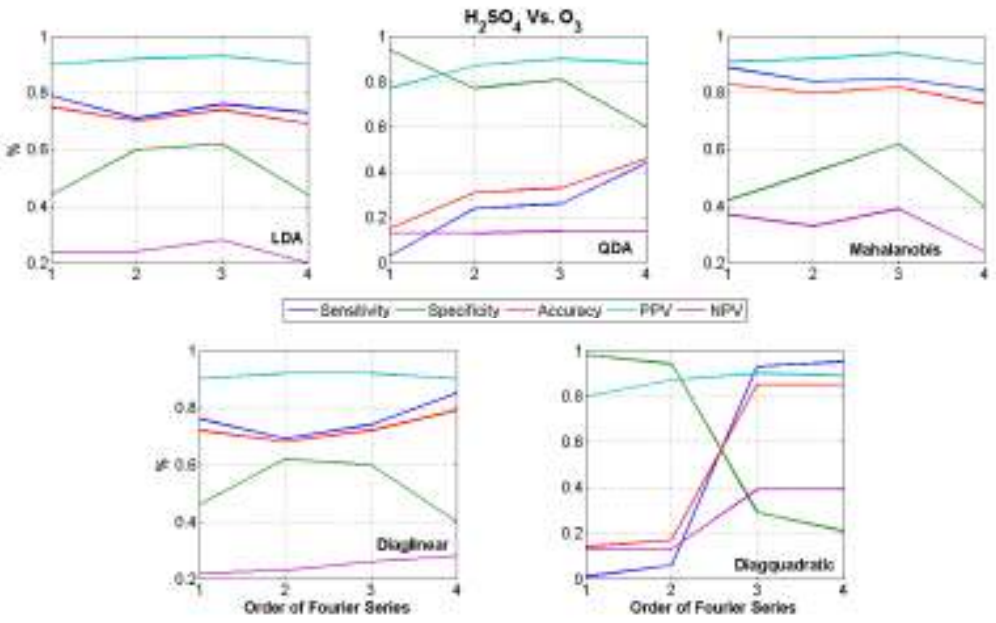


Figure 6.8 (b): Binary classification results using Fourier Curve fit Coefficients

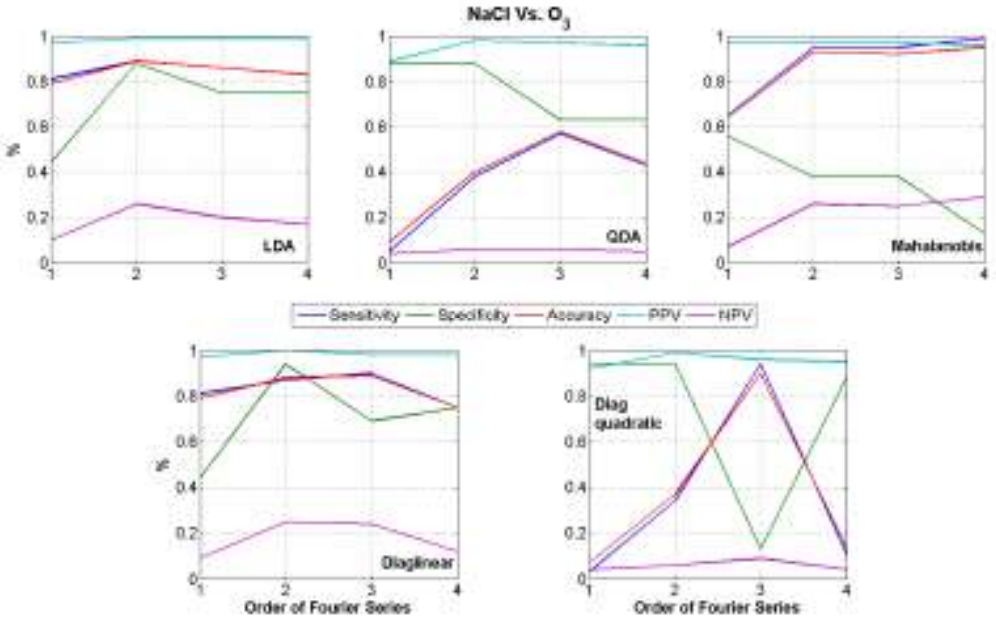


Figure 6.8 (c): Binary classification results using Fourier Curve fit Coefficients

Although the R-Squared values for Gaussian and Exponential curve fits were found to be poor, classification was still attempted using coefficients from these functions to determine how the poor fit contributed to the classification. Classification using Gaussian Curve fit coefficients are shown in Figure 6.9 (a, b, c). As expected, the classification results show inconsistent results, and were found to be somewhat better at 3rd and 4th terms.

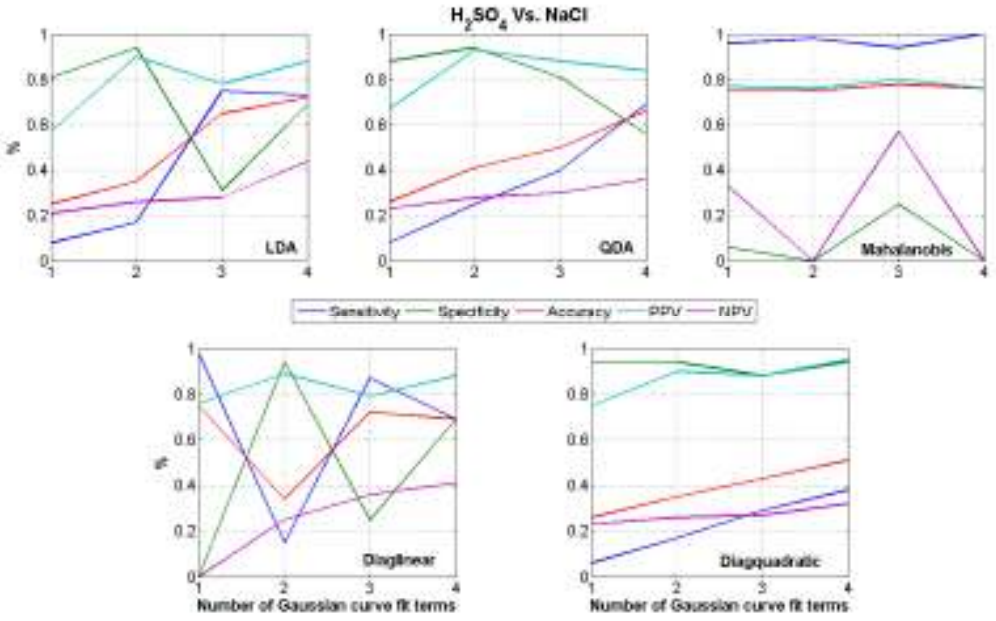


Figure 6.9 (a): Binary classification results using Gaussian Curve fit Coefficients

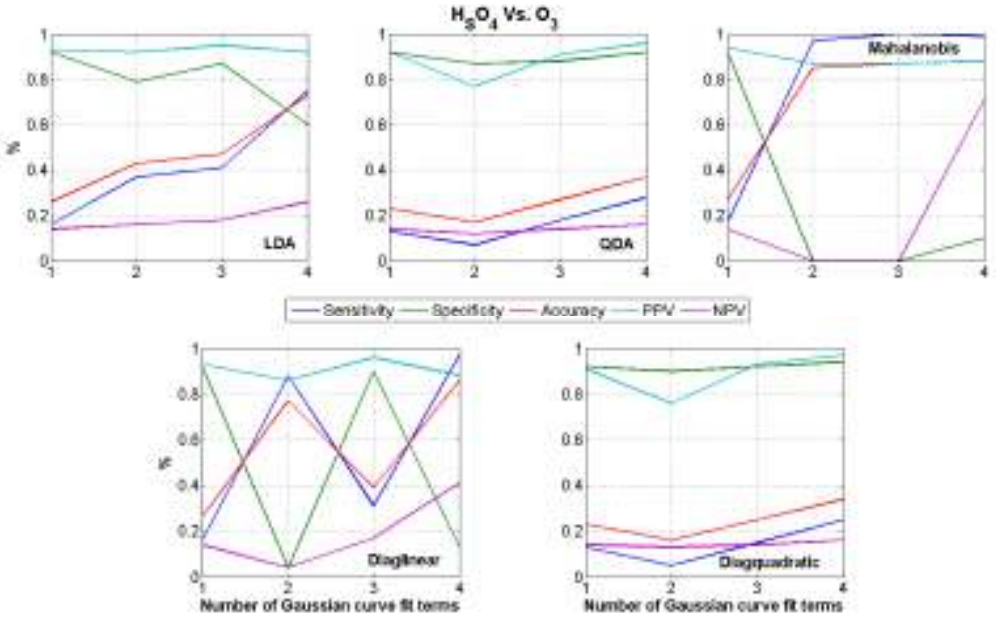


Figure 6.9 (b): Binary classification results using Gaussian Curve fit Coefficients

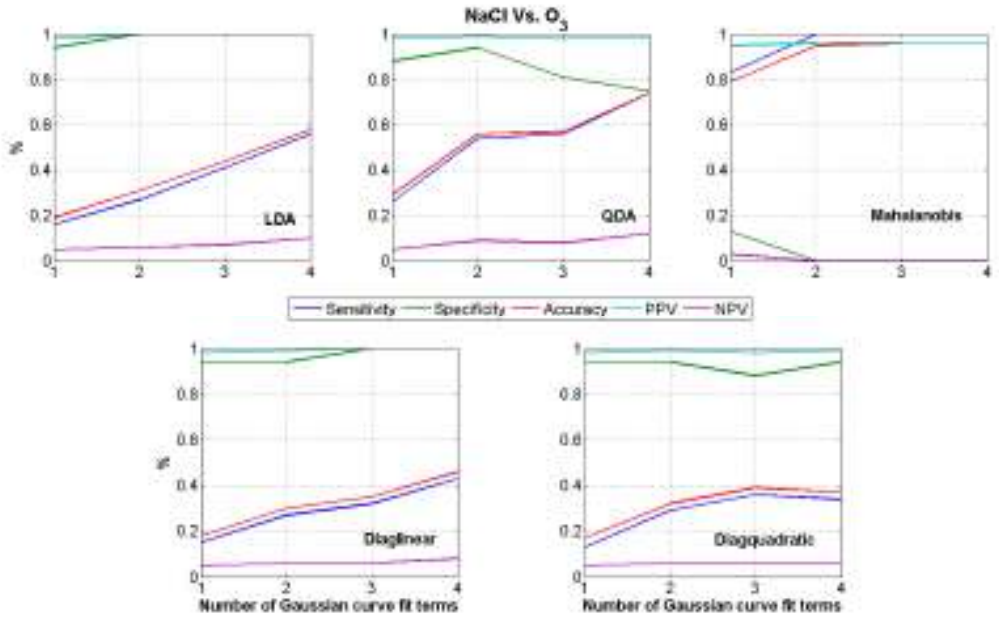


Figure 6.9 (c): Binary classification results using Gaussian Curve fit Coefficients

Figure 6.10 (a,b,c) shows the classification results using Exponential curve fit coefficients.

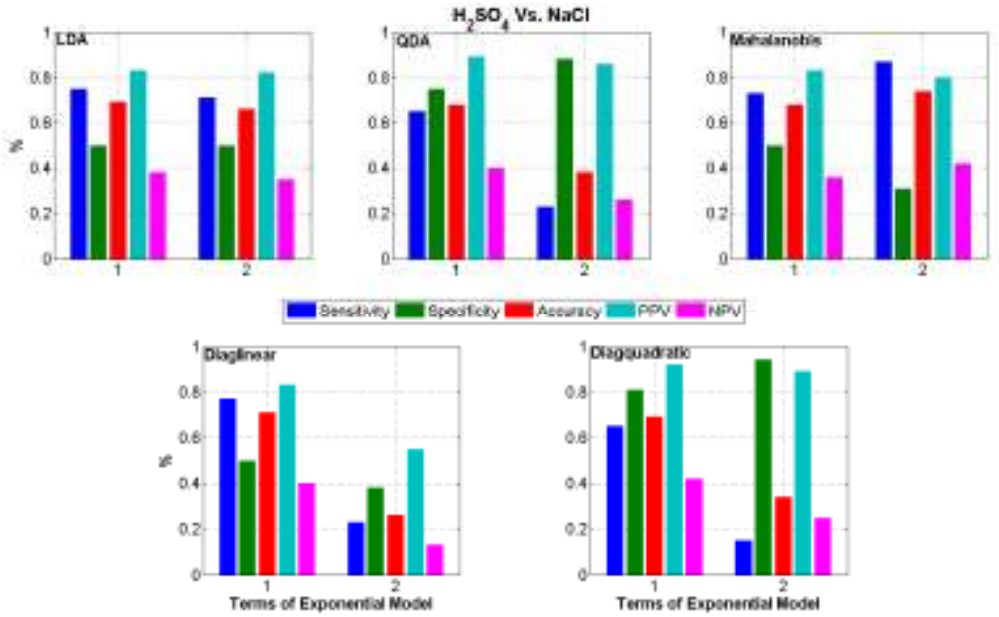


Figure 6.10 (a): Binary classification results using Exponential Curve fit Coefficients

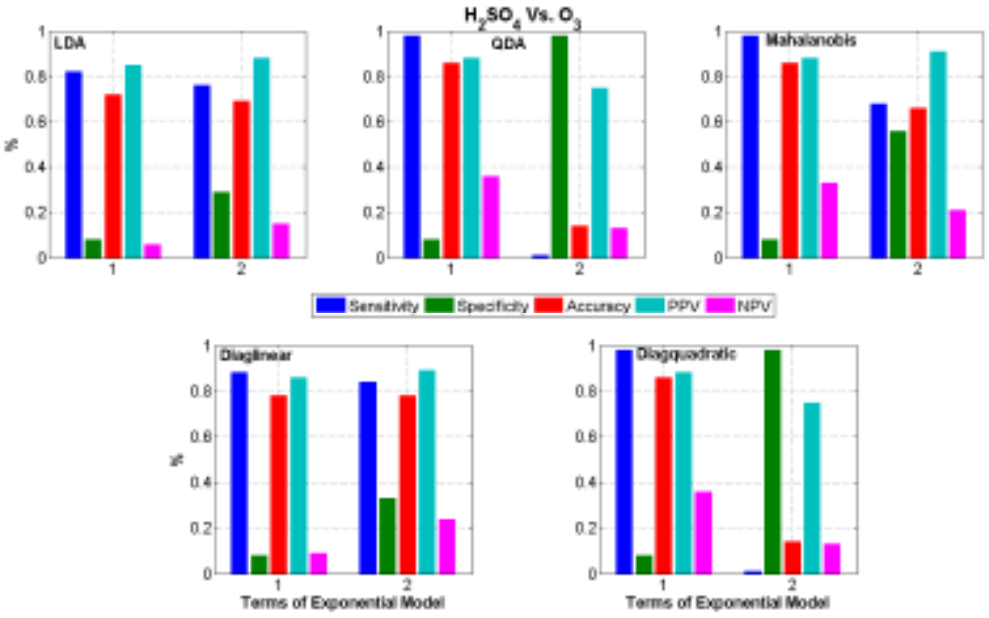


Figure 6.10 (b): Binary classification results using Exponential Curve fit Coefficients

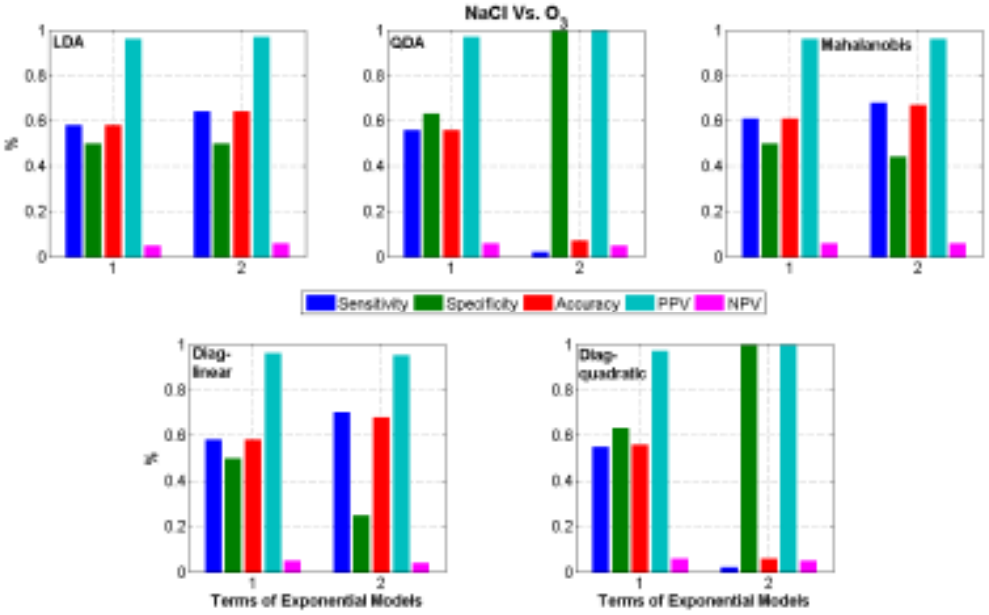


Figure 6.10 (c): Binary classification results using Exponential Curve fit Coefficients

The sensitivity, specificity and accuracy were poor for all three binary stimuli combinations when using coefficients from exponential curve fittings.

The best results from all these figures were obtained using Polynomial coefficients and is summarized in Table 6.3.

A separate prospective study was performed to test the effectiveness of the classification, using four time series for each stimulus that were retained for this purpose. Based on Table 6.3, an OVO decision tree was designed to test against the retained data (see Figure 6.11). Using this decision tree, the results obtained are given in Table 6.4.

Table 6.3: Best Binary Classification results (retrospective study) using Curve fit coefficients

Binary stimuli combination	**Fit type**	**Degree**	**Classifier**	**Classification results**		
				Sensitivity	*Specificity*	*Accuracy*
H_2SO_4 vs. NaCl	Polynomial	5^{th}	LDA	90%	81%	88%
H_2SO_4 vs. O_3	Polynomial	9^{th}	QDA	97%	100%	98%
NaCl vs. O_3	Polynomial	9^{th}	QDA	98%	100%	98%

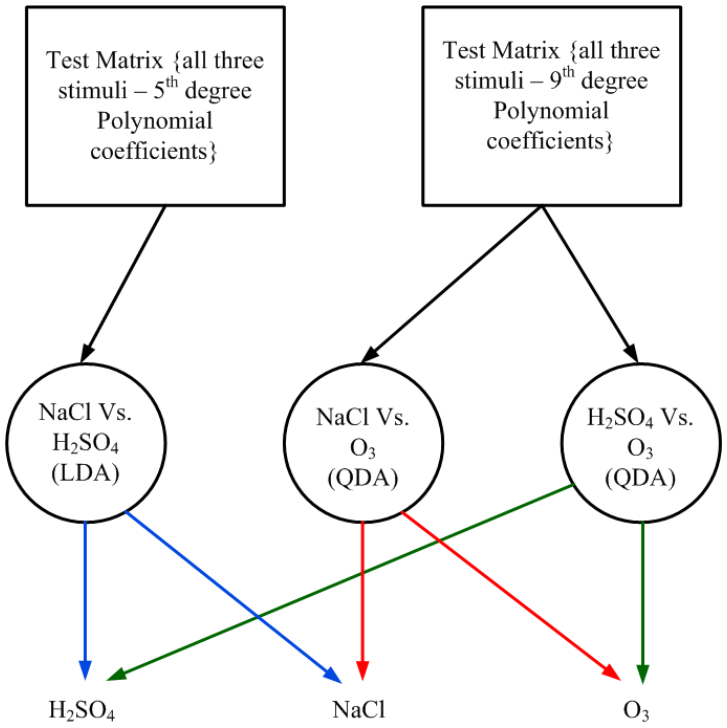


Figure 6.11: Prospective test method using *One Versus One* classification decision tree

The test matrix was designed by including the coefficients required for each binary stimuli combination given in Table 6.3, for all three stimuli. This way, it was easy to feed the coefficients from the curve fit function of desired degree to determine to which stimuli it belonged. The final classification of the test dataset was done using a *One versus One* Decision tree.

Table 6.4: Results from prospective study using retained data

Stimulus	Total retained time series	Number of correctly classified time series
H_2SO_4	4	2
NaCl	4	2
O_3	4	4

For the totally unseen datasets, only 2 out of 4 retained data were correctly classified for H_2SO_4 and NaCl, whereas all 4 retained datasets were correctly classified for O_3. This may be because there were more O_3 data (343 time series) for the classifiers to be trained on than H_2SO_4 (52 time series) and NaCl (16 time series).

Classification using polynomial curve fit coefficients produced good retrospective results and it also produced good prospective results for O_3 as a stimulus as the underlying trend in the data were better captured using Polynomial model. Although the *Goodness of fit* using the Fourier curve fit were closer to that of the Polynomial, the classification results were not as consistent.

6.5 Summary

This chapter explored the classification of stimuli applied to plants using Curve fit coefficients. It was carried out exactly as previously, i.e. using LOOCV on binary stimuli combinations. Coefficients from 5^{th} and 9^{th} degree Polynomial curve fit as features were found to produce the best classification results. These results were then used to design an OVO decision tree to test some retained data, of which H_2SO_4 and NaCl were classified only half the time. However, using the same decision tree, O_3 was detected every time. This can be attributed to less training data for the former two stimuli.

The conclusion is that coefficients of Polynomial curve fitting models can be used for classification of stimuli as they capture the entire trend of the raw plant electrical signal response. This raw signal trend adds more information to the signal than the stochastic part alone, explored in Chapter 5. Although only 12 time series were available to test the decision tree, the comparison is not exact with those results obtained using the segmentation method presented in Chapter 4 and Chapter 5 due to larger number of training and testing samples available. The exploration presented in this chapter provides another way to extract features from raw plant electrical signals, and resulted in around 88% classification accuracy in the

retrospective study. These results are much better than the binary classification results in Table 5.3. This was expected due to the longer duration of the time series used to extract features. However, the segmentation method does provide the insight that any segment from the entire plant electrical signal response provides enough information to the classifier about the stimuli affecting the plant.

Comparing the Classification accuracy (with high Sensitivity and Specificity values) with those reported in Section 5.4.2, the results obtained using curve fit coefficients were much better. Due to the limited data available for training, the prospective study resulted in only 50% classification accuracy for NaCl and H_2SO_4. Since more time series were available for training, a classification accuracy as high as 100% for O_3 was achieved.

7 Conclusions and future work

The goal this dissertation was to find an answer to the question – does plant electrical signal response has sufficient information about the time and type of stimuli which affects the plants? In order to explore this, a step by step approach was taken which has been reported in various chapters within this dissertation. This work establishes that plant electrical signal responses can be used to find out the *time* and the *type* of a stimulus applied to the plant. Two methodologies, *system identification* and *classification*, have been explored for extracting the information about the time and type of stimulus.

The exploration commenced by trying to find an appropriate model, using the electrical response data from 50 plants of four different species (*Laurus nobilis*, *Zamioculcas zamiifolia*, *Chrysanthemum*, and *Cucumis sativus*), to predict some characteristics of an incident light pulse stimulus. These characteristics included the instants of turning on/off and peak intensity of the light pulse. The rising and falling edges of the light pulse were successfully predicted with a forward and inverse modelling dynamical system framework. The best prediction for these characteristics were obtained by a set of NLHW models (Piecewise Linear and Sigmoid Network). The top 3 models showed good prediction ability for the 38 training datasets, and resulted in 44 (forward) and 42 (inverse) model settings. All these model settings were tried on 12 retained test datasets. The results from the tests narrowed down the model settings for both forward and inverse scenarios to the top three, which produced above 70% fit on an average. It was also found that if the shape of the signal was different from the signal that was used to train the models, then the prediction accuracies deteriorated. This system identification approach, can be further explored by using plant electrical response data to determine a variety of stimuli such as gas, chemicals in the soil, etc., thus paving the way towards conceptualizing plant-based environmental biosensors in future.

Although this method was employed to find out whether there was enough information contained within the plant's electrical signals about the *time* and *amplitude* of the applied light stimulus, it did not yield good results on independent datasets within the constraints (e.g. ignoring temperature and humidity, limited variation of parameters such as input-output units and poles-zeroes, etc.). These model constraints can be relaxed to explore whether the results of further studies yield better results in future.

The system identification modelling provided an important finding that it is possible to identify the time of application of a stimulus from the plant's electrical signal response. It also pointed towards plants being non-linear systems, when considering the input stimulus and output electrical signal response, and established a mathematical relationship between the stimulus the signal response.

The next methodology explored was binary classification for predicting the stimulus type from the plant electrical signal response. This was carried out using a small time duration of the plant's signal to extract simple statistical features with five different classifiers. To explore whether the signal could be used to predict the type of stimulus, more complex stimuli were used: Sulphuric acid (H_2SO_4), Sodium chloride (NaCl) solution, and Ozone (O_3). The classification covered 11 statistical features, extracted from segmented signals, followed by feature ranking and rigorous univariate and bivariate feature-based classification using five different classifiers.

The results of binary classification showed that external stimuli like H_2SO_4, O_3 and NaCl (5 ml and 10 ml) were successfully classified, using the adopted approach with 11 statistical features, capturing both the stationary and non-stationary behaviour of the signals. The classification yielded the best average accuracy of ~70% (across all stimuli and five different classifier variants by using *variance* and *skewness* as feature pairs), while the best individual accuracy was ~73.67% (across all stimuli and using *variance* and *IQR* as feature pairs with a *Diagquadratic* classifier). The most significant outcome was that the statistical features of the plant signal could successfully detect which stimulus caused the signal.

The third methodology explored was a decision tree based multiclass classification using raw signals, where two decision tree architecture based classification systems were designed. NaCl was found to be the best separable class in an OVR architecture, with O_3 and H_2SO_4 being the next two classes to be classified. The best results were achieved using the top five features in each of the nodes of the OVO with a *Mahalanobis distance* classifier. Using this configuration achieved good classification accuracy for both retrospective (~92%) and prospective (~90%) studies.

Similar exploration using filtered signals found that NaCl in general, was the best separable class compared with O_3 and H_2SO_4 in an OVR structure, providing accuracies of ~93%

during retrospective and ~89% during prospective study, using the top four features with a Mahalanobis distance classifier.

The most important findings were that raw signal produced marginally better results than filtered signals, and that the settings which provided good results during the retrospective study also worked well for the prospective study.

Based on the results obtained using small segments from both raw and filtered signals, another exploration was carried towards classification by extracting Curve fit coefficients from the entire duration of the raw signal data. This was attempted by fitting four curve types to the raw time series with various parameters (e.g. degrees, no. of terms, etc.). These curve fit coefficients were then used as features for classification resulting in above 90% accuracy in the retrospective study. A separate prospective study was carried out with some retained time series, which gave 100% accuracy for O_3, but only 50% for the remaining two stimuli. This may be because the classifiers were trained on more O_3 data (343 time series) than for H_2SO_4 (52 time series) and NaCl (16 time series).

One of the major limitations of the work presented in this dissertation, was the availability of sufficient data. Hence the next section outlines what could be potential future work which could be developed on the work presented in this dissertation.

7.1 Future work

The research presented in this dissertation provides a platform for exploring plant electrical signal responses to various stimuli as means to monitor environmental parameters. The exploration shows that information about both *time* and *type* of the applied stimulus are embedded within the electrical signals. Depending upon the requirement, either of the methodologies adopted in this dissertation may form the basis for further exploration. Based on the current exploration, constraints and results obtained, the following future projects may be taken further.

More laboratory based experiments need to be carried out:

Any machine learning based data analysis, will require sufficient data for training and testing. Hence one crucial element of any future work must involve more laboratory based experiments. Within these experiments, the following may be stressed on:

Identification of dominant stimuli

How can the dominant stimulus out of multiple stimuli be determined? This entails application of multiple types of stimuli to a single plant and recording its electrical responses. Such signals should be processed so that the most dominant stimulus could be isolated. This will help in realizing a more naturalistic plant electrical signal based classification system, where multiple environmental parameters simultaneously affect the plants.

Exploration of electrical signal response of different species

In this dissertation, the data from multiple species of plants have been combined. However, are there any particular species which are more responsive to a certain stimulus? Will the results be any better if only those species of plants are used to generate the data? One of the easiest ways to answer this will be to plot species wise data (i.e. features) on an LDA basis and see if - for the same stimulus, data from different species are separable. If they are found to be clearly distinguished, then perhaps the role of species in responding to different stimuli is a future direction.

Feature exploration

Although this work used bottom up approach, where up to 15 features were used, a greater number of features could be explored in a top down approach, i.e. starting with a larger set of features and narrowing them down to a few good features. The advantage of using a larger set of features is that more options are available to identify meaningful features which can improve the classification accuracy.

Kernel-based classifiers

Future work could be the use of features explored here with SVM classifiers to see whether any improvement of results occurs. Since a deliberate choice was made to keep the classifier simple and non-kernel-based, but increasing the features during the retrospective study, does reducing the features with kernel-based classifiers improve the results?

Lastly, future work must move on from laboratory based experiments to a more naturalistic setting for data collection. This will enable, a step by step realization of plant electrical signal response based environmental sensor which can be deployed in such naturalistic setting.

APPENDIX - A

Arduino Code for controlling Light

```
/*
  The program is designed to get a csv file containing schedule for light variation experiments to be carried out.
  ESP2866 (7) module is used to connect to the internet and download the csv from the specified server.
  PL1167 RF enabled RGBW bulb (9W) is controlled by an arduino with another PL1167 RF module.
  The colors on the RGBW bulb are varied according to the specified experiment.
  Each line of the csv describes one experiment in the following format.

  <COLOR,WAVE,DURATION>: W,S,10
  COLOR CAN BE, R (RED), G (GREEN), B (BLUE) or W (WHITE)
  WAVE CAN BE, S (Square), T (Triangle), ST (SawTooth).
  DURATION: numeric value in minutes
*/

/* * * * * * * * * * * * * * * * * * * * * * * * * * * *
 Code by Shre Kumar Chatterjee <skc105@ecs.soton.ac.uk>
 Version: 2.00
 Date: 12.08.2016
 * * * * * * * * * * * * * * * * * * * * * * * * * * * */
#include <SPI.h>
```

```
#include <EEPROM.h>
#include <PL1167.h>
#include <Lytwifi.h>
#include <string.h>
#include <SoftwareSerial.h>
#include <WiFiInterrupt.h>

#define PL1167_CS_PIN     10
#define BULB_ADDRESS_HIGH 0
#define BULB_ADDRESS_LOW  0

String CLR = "W";       //Color: Allowed Values (RED) R, (GREEN) G, (BLUE) B, WHITE (W)
String WAV = "S";       //Wave: Allowed Values (SQUARE) S, (TRIANGLE) T, (SAWTOOTH) ST
int DUR = 16;           //In minutes, (Used in case of Triange and Sawtooth waves ignored in case of Square Waves).
float onTime = 2;       //On time in mintues for square wave (Used only in case of Square Wave).
int noOfWaves = 4;      //No of on square waves. Total Time for experiment = noOfWaves*onTime+(noOfWaves-1)*offTime (Used only in case of Square Wave).
float offTime = (DUR-(noOfWaves*onTime))/(noOfWaves-1);    //Off time in mintues for square wave (Used only in case of Square Wave)
int expDone=0;

SoftwareSerial mySerial(5, 6); // RX, TX
LYTWiFi myNetWork(mySerial);
```

```
void setup() {
  // put your setup code here, to run once:
  Serial.begin(9600, SERIAL_8N1);
  myNetWork.vfInitialize(PL1167_CS_PIN);
  vfISRInit();
}

void loop() {
  if(expDone==0){
    myNetWork.vfSwitchOn(0, 0,  C_UNICAST);
    runExperiment();
    myNetWork.vfSwitchOn(0, 0,  C_UNICAST);
    myNetWork.vfSetBrightnessValue(0, 0, 0, C_UNICAST);
    expDone=1;
  }
  delay(6000);
}

void runExperiment() {
  String strDur = String(DUR);
  unsigned long time1 = millis();
  if (WAV.equalsIgnoreCase("S")) {
    Serial.println("Making Square Wave for " + strDur + " mins");
    makeMultiSquareWave( CLR,  noOfWaves,  onTime,  offTime);
  } else if (WAV.equalsIgnoreCase("T")) {
```

```
    Serial.println("Making Triangle Wave for " + strDur + " mins");
    makeTriangleWave(CLR, DUR);
  } else if (WAV.equalsIgnoreCase("ST")) {
    Serial.println("Making SawTooth Wave for " + strDur + " mins");
    makeSawToothWave(CLR, DUR);
  }
  unsigned long timeSpent = millis() - time1;

  Serial.print("All done in ");
  Serial.print((timeSpent / 1000));
  Serial.println("secs");
  Serial.println("");


  //myNetWork.vfSwitchOff(0, 0, C_UNICAST);
}

void makeMultiSquareWave(String CLR, int noOfWaves, float onTime, float offTime) {
  float totalTime = (noOfWaves * onTime) + (noOfWaves - 1) * offTime;
  Serial.print("Total Time ");
  Serial.println(totalTime);
  for (int i = 1; i <= noOfWaves; i++) {
    Serial.print("Turning On for Time ");
    Serial.print(onTime*60*1000);
    Serial.println("mili secs");
    makeSquareWave(CLR, onTime);
```

```
    Serial.print("Turning Off for Time ");
    Serial.print(offTime*60*1000);
    Serial.println("mili secs");
    delayXmins(offTime);
  }
}

void makeSquareWave(String CLR, float DUR) {
  if (CLR.equalsIgnoreCase("W")) {
    myNetWork.vfSetBrightnessValue(0, 0, 255, C_UNICAST);
    delayXmins(DUR);
    myNetWork.vfSetBrightnessValue(0, 0, 1, C_UNICAST);
  } else if (CLR.equalsIgnoreCase("R")) {
    myNetWork.vfSetRGBValues(0, 0, 255, 0, 0, C_UNICAST);
    delayXmins(DUR);
    myNetWork.vfSetRGBValues(0, 0, 1, 0, 0, C_UNICAST);
  } else if (CLR.equalsIgnoreCase("G")) {
    myNetWork.vfSetRGBValues(0, 0, 0, 255, 0, C_UNICAST);
    delayXmins(DUR);
    myNetWork.vfSetRGBValues(0, 0, 0, 1, 0, C_UNICAST);
  } else if (CLR.equalsIgnoreCase("B")) {
    myNetWork.vfSetRGBValues(0, 0, 0, 0, 255, C_UNICAST);
    delayXmins(DUR);
    myNetWork.vfSetRGBValues(0, 0, 0, 0, 1, C_UNICAST);
  }
}
```

```
void makeTriangleWave(String CLR, int DUR) {
  float timeToPeak = (float)DUR / (float)2;
  float stepSize = (float)timeToPeak / (float)255;
  Serial.println("Increasing/decreasing Light intensity in " + String(stepSize * 1000, 4) + "
mili sec steps");
  if (CLR.equalsIgnoreCase("W")) {
    myNetWork.vfSetBrightnessValue(0, 0, 1, C_UNICAST);
    for (int i = 0; i <= 255; i++) {
      myNetWork.vfSetBrightnessValue(0, 0, i, C_UNICAST);
      delayXmins(stepSize);
    }
    for (int i = 0; i <= 255; i++) {
      myNetWork.vfSetBrightnessValue(0, 0, 255 - i, C_UNICAST);
      delayXmins(stepSize);
    }
  } else if (CLR.equalsIgnoreCase("R")) {
    myNetWork.vfSetRGBValues(0, 0, 1, 0, 0, C_UNICAST);
    for (int i = 0; i <= 255; i++) {
      myNetWork.vfSetRGBValues(0, 0, i, 0, 0, C_UNICAST);
      delayXmins(stepSize);
    }
    for (int i = 0; i <= 255; i++) {
      myNetWork.vfSetRGBValues(0, 0, 255 - i, 0, 0, C_UNICAST);
      delayXmins(stepSize);
    }
```

```
  } else if (CLR.equalsIgnoreCase("G")) {
    myNetWork.vfSetRGBValues(0, 0, 0, 1, 0, C_UNICAST);
    for (int i = 0; i <= 255; i++) {
      myNetWork.vfSetRGBValues(0, 0, 0, i, 0, C_UNICAST);
      delayXmins(stepSize);
    }
    for (int i = 0; i <= 255; i++) {
      myNetWork.vfSetRGBValues(0, 0, 0, 255 - i, 0, C_UNICAST);
      delayXmins(stepSize);
    }
  } else if (CLR.equalsIgnoreCase("B")) {
    myNetWork.vfSetRGBValues(0, 0, 0, 0, 1, C_UNICAST);
    for (int i = 0; i <= 255; i++) {
      myNetWork.vfSetRGBValues(0, 0, 0, 0, i, C_UNICAST);
      delayXmins(stepSize);
    }
    for (int i = 0; i <= 255; i++) {
      myNetWork.vfSetRGBValues(0, 0, 0, 0, 255 - i, C_UNICAST);
      delayXmins(stepSize);
    }
  }
}

void makeSawToothWave(String CLR, int DUR) {
  myNetWork.vfSetRGBValues(0, 0, 0, 0, 0, C_UNICAST);
  float stepSize = (float)DUR / (float)255;
```

```
  Serial.println("Increasing Light intensity in " + String(stepSize * 1000, 4) + " mili sec steps");
  if (CLR.equalsIgnoreCase("W")) {
    for (int i = 0; i <= 255; i++) {
      myNetWork.vfSetBrightnessValue(0, 0, i, C_UNICAST);
      delayXmins(stepSize);
    }
    myNetWork.vfSetBrightnessValue(0, 0, 1, C_UNICAST);
  } else if (CLR.equalsIgnoreCase("R")) {
    for (int i = 0; i <= 255; i++) {
      myNetWork.vfSetRGBValues(0, 0, i, 0, 0, C_UNICAST);
      delayXmins(stepSize);
    }
    myNetWork.vfSetRGBValues(0, 0, 1, 0, 0, C_UNICAST);
  } else if (CLR.equalsIgnoreCase("G")) {
    for (int i = 0; i <= 255; i++) {
      myNetWork.vfSetRGBValues(0, 0, 0, i, 0, C_UNICAST);
      delayXmins(stepSize);
    }
    myNetWork.vfSetRGBValues(0, 0, 0, 1, 0, C_UNICAST);
  } else if (CLR.equalsIgnoreCase("B")) {
    for (int i = 0; i <= 255; i++) {
      myNetWork.vfSetRGBValues(0, 0, 0, 0, i, C_UNICAST);
      delayXmins(stepSize);
    }
    myNetWork.vfSetRGBValues(0, 0, 0, 0, 1, C_UNICAST);
```

```
  }
}

void delayXmins(int mins) {
  unsigned long longDelayInSeconds = mins * 60 * 1000; //two minutes;
  unsigned long p = 1;
  while (p < longDelayInSeconds) {
    delay(1);
    p = p + 1;
  }
}

void delayXmins(unsigned long mins) {
  unsigned long longDelayInSeconds = mins * 60 * 1000; //two minutes;
  unsigned long  p = 1;
  while (p < longDelayInSeconds) {
    delay(1);
    p = p + 1;
  }
}

void delayXmins(double mins) {
  unsigned long longDelayInSeconds = mins * 60 * 1000; //two minutes;
  unsigned long p = 1;
  while (p < longDelayInSeconds) {
    delay(1);
```

```
    p = p + 1;
  }
}
void delayXmins(float mins) {
  float longDelayInSeconds = mins * 60 * 1000; //two minutes;
  float p = 1;
  while (p < longDelayInSeconds) {
    delay(1);
    p = p + 1;
  }
}
```

APPENDIX - B

MATLAB Code for System Identification

```
%This automated script outputs the results of System Identification Linear
%Models (ARX, ARMAX, OE and BJ) into a dedicated Excel sheet, for both
%Forward and Inverse relationships.
% skc105@ecs.soton.ac.uk;
% 14/09/2016

% ASSIGNING MULTIPLE PROCESSORS FOR COMPUTATION
isOpen = matlabpool('size') > 0
if(~isOpen)
    matlabpool open local 2 %Keep the number of processors below 4...
end



clc
% Input_1 = LightPulse;
% Output_1 = smooth_D1;
tic
Ts = 0.001; %Sampling rate of 1KHz
Data_1 = LightPulse;
Data_2 = Data1_smooth;
```

```
ze_fwd = iddata(Data_2,Data_1,Ts); % Creating iddata structure involving input and output,
% data = iddata(y,u,Ts) WHERE y=output,u=input
ze_inv = iddata(Data_1,Data_2,Ts); % Creating iddata structure involving input and output

ze_fwd.Tstart
ze_inv.Tstart



% NN1 = struc(1:5,1:5,0); %Using 'struc' to create a matrix of possible model orders for
na,nb,nk

%Estimating the transfer function for the data

% Opt = tfestOptions('Display','on'); % Choosing to view a progress report by setting the
Display option to on in the option set created by the 'tfestOptions' command
%
% np = 2;
% ioDelay = 0;
% mtf = tfest(Ze,np,[],ioDelay,Opt); % Estimating the transfer function.
filename = 'System_Identification_November_2016_Data16a';



%for N=5:5:40
for i=1:1:10
```

```
    j=i;
    k=i;
    l=1;
%ARX Model parameter estimation

% a(i,:)={i j k 1};

opt_arx = arxOptions('Focus','prediction');
opt_armax = armaxOptions('Focus','prediction');
opt_oe = oeOptions('Focus','prediction');
opt_bj=bjOptions('Focus','prediction');

marx_fwd_arx = arx(ze_fwd,[i j l],opt_arx);
marx_inv_arx = arx(ze_inv,[i j l],opt_arx);


marx_fwd_armax = armax(ze_fwd,[i j k l],opt_armax);
marx_inv_armax = armax(ze_inv,[i j k l],opt_armax);

marx_fwd_oe = oe(ze_fwd,[i j l],opt_oe);
marx_inv_oe = oe(ze_inv,[i j l],opt_oe);

marx_fwd_bj = bj(ze_fwd,[i i i i l],opt_bj);
marx_inv_bj = bj(ze_inv,[i i i i l],opt_bj);

present(marx_fwd_bj) % Displays model parameters with uncertainty information
```

```
% yp = compare(ze,marx);

% sys = arx(ze,[3 2 1]);

[y_fwd_arx,fit_fwd_arx(i,1),x0_fwd_arx] = compare(ze_fwd,marx_fwd_arx);
[y_inv_arx,fit_inv_arx(i,1),x0_inv_arx] = compare(ze_inv,marx_inv_arx);

[y_fwd_armax,fit_fwd_armax(i,1),x0_fwd_armax] = compare(ze_fwd,marx_fwd_armax);
[y_inv_armax,fit_inv_armax(i,1),x0_inv_armax] = compare(ze_inv,marx_inv_armax);

[y_fwd_oe,fit_fwd_oe(i,1),x0_fwd_oe] = compare(ze_fwd,marx_fwd_oe);
[y_inv_oe,fit_inv_oe(i,1),x0_inv_oe] = compare(ze_inv,marx_inv_oe);

[y_fwd_bj,fit_fwd_bj(i,1),x0_fwd_bj] = compare(ze_fwd,marx_fwd_bj);
[y_inv_bj,fit_inv_bj(i,1),x0_inv_bj] = compare(ze_inv,marx_inv_bj);

xlswrite(filename,fit_fwd_arx,'Forward_Models','B3'); %B3
xlswrite(filename,fit_inv_arx,'Inverse_Models','B3');

xlswrite(filename,fit_fwd_armax,'Forward_Models','B14'); %B14
xlswrite(filename,fit_inv_armax,'Inverse_Models','B14');

xlswrite(filename,fit_fwd_oe,'Forward_Models','B25'); %B25
xlswrite(filename,fit_inv_oe,'Inverse_Models','B25');
```

```
xlswrite(filename,fit_fwd_bj,'Forward_Models','B36'); %B36
xlswrite(filename,fit_inv_bj,'Inverse_Models','B36');

end

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
setpref('Internet','SMTP_Server','smtp.ecs.soton.ac.uk');
setpref('Internet','E_mail','skc105@ecs.soton.ac.uk');
sendmail('skc105@ecs.soton.ac.uk','Linear Models complete,Data16a') % ADD DATASET
NAME
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ Non-Linear Models

for i=1:1:10

   j=i;
   k=i;

InputNL_deadzone = 'deadzone';
OutputNL_deadzone = 'deadzone';

InputNL_saturation = 'saturation';
OutputNL_saturation = 'saturation';

InputNL_wavenet = wavenet('NumberOfUnits','auto');
OutputNL_wavenet = wavenet('NumberOfUnits','auto');
```

```
InputNL_poly1d = poly1d('Degree',i); %Degree is 1:10, represented by No.
%of Units in toolbox
OutputNL_poly1d = poly1d('Degree',i); %Degree is 1:10, represented by No.
%of Units in toolbox

marx_fwd_nlhw_deadzone = nlhw(ze_fwd,[i j 1],InputNL_deadzone,OutputNL_deadzone);
marx_inv_nlhw_deadzone = nlhw(ze_inv,[i j 1],InputNL_deadzone,OutputNL_deadzone);

marx_fwd_nlhw_saturation = nlhw(ze_fwd,[i j 1],InputNL_saturation,OutputNL_saturation);
marx_inv_nlhw_saturation = nlhw(ze_inv,[i j 1],InputNL_saturation,OutputNL_saturation);

marx_fwd_nlhw_wavenet = nlhw(ze_fwd,[i j 1],InputNL_wavenet,OutputNL_wavenet);
marx_inv_nlhw_wavenet = nlhw(ze_inv,[i j 1],InputNL_wavenet,OutputNL_wavenet);

marx_fwd_nlhw_poly1d_5 = nlhw(ze_fwd,[5 5 1],InputNL_poly1d,OutputNL_poly1d);
marx_inv_nlhw_poly1d_5 = nlhw(ze_inv,[5 5 1],InputNL_poly1d,OutputNL_poly1d);

marx_fwd_nlhw_poly1d_10 = nlhw(ze_fwd,[10 10 1],InputNL_poly1d,OutputNL_poly1d);
marx_inv_nlhw_poly1d_10 = nlhw(ze_inv,[10 10 1],InputNL_poly1d,OutputNL_poly1d);

[y_fwd_deadzone,fit_fwd_deadzone(i,1),x0_fwd_deadzone]                    =
compare(ze_fwd,marx_fwd_nlhw_deadzone);
[y_inv_deadzone,fit_inv_deadzone(i,1),x0_inv_deadzone]                    =
compare(ze_inv,marx_inv_nlhw_deadzone);
```

```
[y_fwd_saturation,fit_fwd_saturation(i,1),x0_fwd_saturation] =
compare(ze_fwd,marx_fwd_armax);
[y_inv_saturation,fit_inv_saturation(i,1),x0_inv_saturation] =
compare(ze_inv,marx_inv_armax);

[y_fwd_wavenet,fit_fwd_wavenet(i,1),x0_fwd_wavenet] =
compare(ze_fwd,marx_fwd_nlhw_wavenet);
[y_inv_wavenet,fit_inv_wavenet(i,1),x0_inv_wavenet] =
compare(ze_inv,marx_inv_nlhw_wavenet);

[y_fwd_polyid_5,fit_fwd_poly1d_5(i,1),x0_fwd_polyid_5] =
compare(ze_fwd,marx_fwd_nlhw_poly1d_5);
[y_inv_polyid_5,fit_inv_poly1d_5(i,1),x0_inv_polyid_5] =
compare(ze_inv,marx_inv_nlhw_poly1d_5);

[y_fwd_polyid_10,fit_fwd_poly1d_10(i,1),x0_fwd_polyid_10] =
compare(ze_fwd,marx_fwd_nlhw_poly1d_10);
[y_inv_polyid_10,fit_inv_poly1d_10(i,1),x0_inv_polyid_10] =
compare(ze_inv,marx_inv_nlhw_poly1d_10);



end

xlswrite(filename,fit_fwd_deadzone,'Forward_Models','B47'); %B47
xlswrite(filename,fit_inv_deadzone,'Inverse_Models','B47');

xlswrite(filename,fit_fwd_saturation,'Forward_Models','B58'); %B58
```

```
xlswrite(filename,fit_inv_saturation,'Inverse_Models','B58');

xlswrite(filename,fit_fwd_wavenet,'Forward_Models','B69'); %B69
xlswrite(filename,fit_inv_wavenet,'Inverse_Models','B69');

xlswrite(filename,fit_fwd_poly1d_5,'Forward_Models','B80'); %B80
xlswrite(filename,fit_inv_poly1d_5,'Inverse_Models','B80');

xlswrite(filename,fit_fwd_poly1d_10,'Forward_Models','B91'); %B91
xlswrite(filename,fit_inv_poly1d_10,'Inverse_Models','B91');

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~

setpref('Internet','SMTP_Server','smtp.ecs.soton.ac.uk');
setpref('Internet','E_mail','skc105@ecs.soton.ac.uk');
sendmail('skc105@ecs.soton.ac.uk','NL Models Phase-1 complete,Data16a') % ADD
DATASET NAME

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~

% matlabpool close
clc
```

```
fit_fwd_pwlinear = zeros(10,8);
fit_inv_pwlinear = zeros(10,8);
fit_fwd_sigmoid = zeros(10,8);
fit_inv_sigmoid = zeros(10,8);

for N=1:8
    for i=1:1:10
    j=i;

% InputNL_pwlinear = pwlinear('NumberOfUnits',N*5); %N=5:5:40
% OutputNL_pwlinear = pwlinear('NumberOfUnits',N*5); %N=5:5:40
%
% InputNL_sigmoid = sigmoidnet('NumberOfUnits',N*5); %N=5:5:40
% OutputNL_sigmoid = sigmoidnet('NumberOfUnits',N*5);%N=5:5:40

marx_fwd_nlhw_pwlinear = nlhw(ze_fwd,[i j 1],pwlinear('NumberOfUnits',N*5),pwlinear('NumberOfUnits',N*5));
marx_inv_nlhw_pwlinear = nlhw(ze_inv,[i j 1],pwlinear('NumberOfUnits',N*5),pwlinear('NumberOfUnits',N*5));

marx_fwd_nlhw_sigmoid = nlhw(ze_fwd,[i j 1],sigmoidnet('NumberOfUnits',N*5),sigmoidnet('NumberOfUnits',N*5));
marx_inv_nlhw_sigmoid = nlhw(ze_inv,[i j 1],sigmoidnet('NumberOfUnits',N*5),sigmoidnet('NumberOfUnits',N*5));

[y_fwd_pwlinear,fit_fwd_pwlinear(i,N),x0_fwd_pwlinear] = compare(ze_fwd,marx_fwd_nlhw_pwlinear);
```

```
[y_inv_pwlinear,fit_inv_pwlinear(i,N),x0_inv_pwlinear] =
compare(ze_inv,marx_inv_nlhw_pwlinear);

[y_fwd_sigmoid,fit_fwd_sigmoid(i,N),x0_fwd_sigmoid] =
compare(ze_fwd,marx_fwd_nlhw_sigmoid);
[y_inv_sigmoid,fit_inv_sigmoid(i,N),x0_inv_sigmoid] =
compare(ze_inv,marx_inv_nlhw_sigmoid);

    end
end
% end


xlswrite(filename,fit_fwd_pwlinear,'Forward_Models','B102'); %B102
xlswrite(filename,fit_inv_pwlinear,'Inverse_Models','B102');

xlswrite(filename,fit_fwd_sigmoid,'Forward_Models','B113'); %B113
xlswrite(filename,fit_inv_sigmoid,'Inverse_Models','B113');

matlabpool close

toc

%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ SEND EMAIL AFTER
SIMULATION
%IS COMPLETE ~~~~~~~~~~~~~~
```

```
setpref('Internet','SMTP_Server','smtp.ecs.soton.ac.uk');

setpref('Internet','E_mail','skc105@ecs.soton.ac.uk');

sendmail('skc105@soton.ac.uk','All simulation complete, Data16a') % ADD DATASET NAME
```

APPENDIX - C

1. **Acronyms**

C. Dimension = Correlation Dimension

DFA = Detrended Fluctuation Analysis

H. Complexity = Hjorth Complexity

H. Mobility = Hjorth Mobility

H. Exponent = Hurst Exponent

H. Skewness = Hyper Skewness

H. Flatness = Hyper Flatness

IQR = Interquartile range

PSD = Average Spectral Power

W. Entropy = Wavelet Entropy

Correlation between features (computed from Raw signals)

	Corr. Dim	Mean	Var	IQR	Skew	Kurtosis	H. Mob	H. comp	Hurst exp.	DFA	Wentropy	PSD	H. skew	H. flat	Fano factor
Corr. Dim	1.00	0.19	-0.09	-0.13	0.01	-0.09	0.04	0.15	-0.13	-0.20	-0.32	-0.26	-0.04	-0.09	-0.02
Mean	0.19	1.00	0.02	-0.01	-0.14	-0.33	-0.02	0.05	0.07	0.00	0.24	0.31	-0.30	-0.34	0.00
Var	-0.09	0.02	1.00	0.76	-0.04	-0.04	-0.01	-0.04	0.05	0.16	0.04	0.05	-0.05	-0.05	0.17
IQR	-0.13	-0.01	0.76	1.00	-0.17	-0.07	-0.04	-0.07	0.26	0.49	0.01	0.01	-0.20	-0.09	0.11
Skew	0.01	-0.14	-0.04	-0.17	1.00	0.24	0.02	0.00	-0.22	-0.22	0.04	0.03	0.68	0.29	0.00
Kurtosis	-0.09	-0.33	-0.04	-0.07	0.24	1.00	-0.06	-0.31	-0.19	-0.19	0.10	0.01	0.44	0.97	0.00
H. Mob	0.04	-0.02	-0.01	-0.04	0.02	-0.06	1.00	0.37	-0.05	-0.06	-0.09	-0.07	0.01	-0.05	0.01
H. comp	0.15	0.05	-0.04	-0.07	0.00	-0.31	0.37	1.00	-0.04	-0.07	-0.26	-0.16	-0.06	-0.30	-0.03
Hurst exp.	-0.13	0.07	0.05	0.26	-0.22	-0.19	-0.05	-0.04	1.00	0.87	0.07	0.07	-0.30	-0.19	0.00
DFA	-0.20	0.00	0.16	0.49	-0.22	-0.19	-0.06	-0.07	0.87	1.00	0.08	0.07	-0.24	-0.19	0.00
Wentropy	-0.32	0.24	0.04	0.01	0.04	0.10	-0.09	-0.26	0.07	0.08	1.00	0.97	0.10	0.12	0.00
PSD	-0.26	0.31	0.05	0.01	0.03	0.01	-0.07	-0.16	0.07	0.07	0.97	1.00	0.07	0.04	0.00
H. skew	-0.04	-0.30	-0.05	-0.20	0.68	0.44	0.01	-0.06	-0.30	-0.24	0.10	0.07	1.00	0.52	0.00
H. flat	-0.09	-0.34	-0.05	-0.09	0.29	0.97	-0.05	-0.30	-0.19	-0.19	0.12	0.04	0.52	1.00	0.00
Fano factor	-0.02	0.00	0.17	0.11	0.00	0.00	0.01	-0.03	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Feature Ranking (for Raw Plant Signals) Under OVR Scheme

Feature rank	*NaCl vs. Rest*	*H_2SO_4 vs. Rest*	*O_3 vs. Rest*
1	IQR	DFA	W. Entropy
2	H. Flatness	W. Entropy	DFA
3	Kurtosis	H. Exponent	H. Skewness
4	DFA	H. Skewness	H. Exponent
5	Variance	PSD	PSD
6	C. Dimension	IQR	IQR
7	W. Entropy	H. Complexity	Skewness
8	PSD	H. Flatness	H. Complexity
9	H. Mobility	Kurtosis	Mean
10	H. Complexity	Skewness	C. Dimension
11	H. Skewness	Mean	Variance
12	H. Exponent	Variance	H. Mobility
13	Skewness	C. Dimension	H. Flatness
14	Fano factor	H. Mobility	Kurtosis
15	Mean	Fano factor	Fano factor

Feature Ranking (for Raw Plant Signals) Under OVO Scheme

Feature rank	*NaCl vs. O_3*	*NaCl vs. H_2SO_4*	*H_2SO_4 vs. O_3*
1	IQR	DFA	DFA
2	H. Flatness	Kurtosis	W. Entropy
3	Kurtosis	H. Flatness	H. Exponent
4	Variance	H. Skewness	H. Skewness
5	DFA	IQR	PSD
6	C. Dimension	H. Exponent	IQR
7	W. Entropy	H. Mobility	H. Complexity
8	PSD	C. Dimension	H. Flatness
9	H. Mobility	H. Complexity	Kurtosis
10	H. Complexity	Variance	Skewness
11	Skewness	W. Entropy	Mean
12	H. Skewness	Mean	Variance
13	H. Exponent	Skewness	C. Dimension
14	Fano factor	PSD	H. Mobility
15	Mean	Fano factor	Fano factor

Feature Ranking (for Filtered Plant Signals) Under OVR Scheme

Feature rank	*NaCl vs. Rest*	*H_2SO_4 vs. Rest*	*O_3 vs. Rest*
1	Mean	Mean	Mean
2	H. Complexity	W. Entropy	W. Entropy
3	IQR	IQR	IQR
4	DFA	PSD	PSD
5	Variance	Variance	H. Mobility
6	H. Exponent	H. Mobility	H. Exponent
7	Kurtosis	H. Exponent	Kurtosis
8	W. Entropy	Skewness	H. Skewness
9	H. Mobility	Kurtosis	H. Flatness
10	H. Skewness	H. Skewness	Skewness
11	H. Flatness	H. Flatness	H. Complexity
12	Fano factor	C. Dimension	Variance
13	C. Dimension	H. Complexity	C. Dimension
14	PSD	DFA	Fano factor
15	Skewness	Fano factor	DFA

Feature Ranking (for Filtered Plant Signals) Under OVO Scheme

Feature rank	*NaCl vs. O_3*	*NaCl vs. H_2SO_4*	*H_2SO_4 vs. O_3*
1	Mean	Mean	Mean
2	H. Complexity	IQR	W. Entropy
3	DFA	H. Mobility	IQR
4	IQR	H. Complexity	Variance
5	H. Exponent	Variance	PSD
6	Variance	PSD	H. Mobility
7	Kurtosis	DFA	H. Exponent
8	W. Entropy	W. Entropy	Skewness
9	H. Mobility	H. Flatness	Kurtosis
10	H. Skewness	Skewness	H. Skewness
11	H. Flatness	H. Skewness	H. Flatness
12	Fano factor	Kurtosis	H. Complexity
13	Skewness	Fano factor	C. Dimension
14	C. Dimension	C. Dimension	Fano factor
15	PSD	H. Exponent	DFA

APPENDIX – D

MATLAB Code for Extracting Curve Fit Coefficients

```
% This script is written to automate Curve Fit on Raw Plant Electrical
Signals
%skc105@ecs.soton.ac.uk

% function[percentfit,coeffecients]=CurveFit_Shre(data,'fittype')

data = Ozone_Cabbage_4_Stimulus_5_Post_Ch4;
[rows,columns] = size(data);
j=1:1:rows;
j = transpose(j);

tic

cc = 'H416'; %Cell on which to write the COEFFICIENTS
rsq = 'F416'; %Cell to write RSQUARED values
RR = 'G416';  %Cell to write RMSE values

filename = 'Curve_Fitting_Gaussian';

% newOptions = fitoptions(fitOptions,Name,Value)
% f = fit(x,y,'exp1');
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
%Available options for Curve Fittings
% 'weibul' - There are no fit settings to configure
% 'exp1' or 'exp2' for Single term or two term Exponential

% x = linspace(0,4*pi,10);
% y = sin(x);

[gfit_1,gof1] = fit(j,data,'gauss1');
gfit_Coeffs_1 = coeffvalues(gfit_1);
xlswrite(filename,gfit_Coeffs_1,'Post Stimulus-1st Order',cc);
xlswrite(filename,gof1.rsquare,'Post Stimulus-1st Order',rsq);
xlswrite(filename,gof1.rmse,'Post Stimulus-1st Order',RR);

[gfit_2,gof2] = fit(j,data,'gauss2');
gfit_Coeffs_2 = coeffvalues(gfit_2);
xlswrite(filename,gfit_Coeffs_2,'Post Stimulus-2nd Order',cc);
xlswrite(filename,gof2.rsquare,'Post Stimulus-2nd Order',rsq);
xlswrite(filename,gof2.rmse,'Post Stimulus-2nd Order',RR);

[gfit_3,gof3] = fit(j,data,'gauss3');
gfit_Coeffs_3 = coeffvalues(gfit_3);
xlswrite(filename,gfit_Coeffs_3,'Post Stimulus-3rd Order',cc);
xlswrite(filename,gof3.rsquare,'Post Stimulus-3rd Order',rsq);
xlswrite(filename,gof3.rmse,'Post Stimulus-3rd Order',RR);

[gfit_4,gof4] = fit(j,data,'gauss4');
gfit_Coeffs_4 = coeffvalues(gfit_4);
xlswrite(filename,gfit_Coeffs_4,'Post Stimulus-4th Order',cc);
xlswrite(filename,gof4.rsquare,'Post Stimulus-4th Order',rsq);
```

```
xlswrite(filename,gof4.rmse,'Post Stimulus-4th Order',RR);

clc

% xlswrite(filename,gfit_Coeffs,'Post Stimulus-2nd Order',cc);
%
% xlswrite(filename,gfit_Coeffs,'Post Stimulus-3rd Order',cc);
%
% xlswrite(filename,gfit_Coeffs,'Post Stimulus-4th Order',cc);


% xlswrite(filename,p{1,5},'Post Stimulus-5th Order',cc);
% xlswrite(filename,p{1,6},'Post Stimulus-6th Order',cc);
% xlswrite(filename,p{1,7},'Post Stimulus-7th Order',cc);
% xlswrite(filename,p{1,8},'Post Stimulus-8th Order',cc);
% xlswrite(filename,p{1,9},'Post Stimulus-9th Order',cc);

filename_f = 'Curve_Fitting_Fourier';

% newOptions = fitoptions(fitOptions,Name,Value)
% f = fit(x,y,'exp1');
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
%Available options for Curve Fittings
% 'weibul' - There are no fit settings to configure
% 'exp1' or 'exp2' for Single term or two term Exponential

% x = linspace(0,4*pi,10);
% y = sin(x);

[f_fit_1,f_gof1] = fit(j,data,'fourier1');
ffit_Coeffs_1 = coeffvalues(f_fit_1);
xlswrite(filename_f,ffit_Coeffs_1,'Post Stimulus-1st Order',cc);
xlswrite(filename_f,f_gof1.rsquare,'Post Stimulus-1st Order',rsq);
xlswrite(filename_f,f_gof1.rmse,'Post Stimulus-1st Order',RR);

[f_fit_2,f_gof2] = fit(j,data,'fourier2');
ffit_Coeffs_2 = coeffvalues(f_fit_2);
xlswrite(filename_f,ffit_Coeffs_2,'Post Stimulus-2nd Order',cc);
xlswrite(filename_f,f_gof2.rsquare,'Post Stimulus-2nd Order',rsq);
xlswrite(filename_f,f_gof2.rmse,'Post Stimulus-2nd Order',RR);

[f_fit_3,f_gof3] = fit(j,data,'fourier3');
ffit_Coeffs_3 = coeffvalues(f_fit_3);
xlswrite(filename_f,ffit_Coeffs_3,'Post Stimulus-3rd Order',cc);
xlswrite(filename_f,f_gof3.rsquare,'Post Stimulus-3rd Order',rsq);
xlswrite(filename_f,f_gof3.rmse,'Post Stimulus-3rd Order',RR);

[f_fit_4,f_gof4] = fit(j,data,'fourier4');
ffit_Coeffs_4 = coeffvalues(f_fit_4);
xlswrite(filename_f,ffit_Coeffs_4,'Post Stimulus-4th Order',cc);
xlswrite(filename_f,f_gof4.rsquare,'Post Stimulus-4th Order',rsq);
xlswrite(filename_f,f_gof4.rmse,'Post Stimulus-4th Order',RR);

clc
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

filename_e = 'Curve_Fitting_Exponential';

% newOptions = fitoptions(fitOptions,Name,Value)
% f = fit(x,y,'exp1');
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
%Available options for Curve Fittings
% 'weibul' - There are no fit settings to configure
% 'exp1' or 'exp2' for Single term or two term Exponential

% x = linspace(0,4*pi,10);
% y = sin(x);

[exp_gfit_1,exp_gof1] = fit(j,data,'exp1');
efit_Coeffs_1 = coeffvalues(exp_gfit_1);
xlswrite(filename_e,efit_Coeffs_1,'Post Stimulus-1st Order',cc);
xlswrite(filename_e,exp_gof1.rsquare,'Post Stimulus-1st Order',rsq);
xlswrite(filename_e,exp_gof1.rmse,'Post Stimulus-1st Order',RR);

[exp_gfit_2,exp_gof2] = fit(j,data,'exp2');
efit_Coeffs_2 = coeffvalues(exp_gfit_2);
xlswrite(filename_e,efit_Coeffs_2,'Post Stimulus-2nd Order',cc);
xlswrite(filename_e,exp_gof2.rsquare,'Post Stimulus-2nd Order',rsq);
xlswrite(filename_e,exp_gof2.rmse,'Post Stimulus-2nd Order',RR);

clc

toc

% xlswrite(filename,gfit_Coeffs,'Post Stimulus-2nd Order',cc);
%
% xlswrite(filename,gfit_Coeffs,'Post Stimulus-3rd Order',cc);
%
% xlswrite(filename,gfit_Coeffs,'Post Stimulus-4th Order',cc);


% xlswrite(filename,p{1,5},'Post Stimulus-5th Order',cc);
% xlswrite(filename,p{1,6},'Post Stimulus-6th Order',cc);
% xlswrite(filename,p{1,7},'Post Stimulus-7th Order',cc);
% xlswrite(filename,p{1,8},'Post Stimulus-8th Order',cc);
% xlswrite(filename,p{1,9},'Post Stimulus-9th Order',cc);
```

References

[1] D. J. Nowak, S. Hirabayashi, A. Bodine, and E. Greenfield, “Tree and forest effects on air quality and human health in the United States,” *Environmental Pollution*, vol. 193, pp. 119–129, 2014.

[2] A. Z. Abbasi, N. Islam, Z. A. Shaikh, and others, “A review of wireless sensors and networks’ applications in agriculture,” *Computer Standards \& Interfaces*, vol. 36, no. 2, pp. 263–270, 2014.

[3] A. Zerger, R. V. Rossel, D. Swain, T. Wark, R. N. Handcock, V. Doerr, G. Bishop-Hurley, E. Doerr, P. Gibbons, and C. Lobsey, “Environmental sensor networks for vegetation, animal and soil sciences,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 12, no. 5, pp. 303–316, 2010.

[4] J. K. Hart and K. Martinez, “Environmental Sensor Networks: A revolution in the earth system science?,” *Earth-Science Reviews*, vol. 78, no. 3, pp. 177–191, 2006.

[5] “http://www.un.org/en/events/forestsday/.” [Online]. Available: http://www.un.org/en/events/forestsday/.

[6] “www.cancer.org.” [Online]. Available: www.cancer.org.

[7] M. Guarnieri and J. R. Balmes, “Outdoor air pollution and asthma,” *The Lancet*, vol. 383, no. 9928, pp. 1581–1592, 2014.

[8] J. Bartra, J. Mullol, A. Del Cuvillo, I. Dávila, M. Ferrer, I. Jáuregui, J. Montoro, J. Sastre, and A. Valero, “Air pollution and allergens,” *J Investig Allergol Clin Immunol*, vol. 17, no. Suppl 2, pp. 3–8, 2007.

[9] P. D. Sly, “Traffic-related air pollution: an avoidable exposure to improve respiratory health,” *Thorax*, p. thoraxjnl–2014, 2014.

[10] “http://climatechange-foodsecurity.org/trop_ozone.html.” [Online]. Available: http://climatechange-foodsecurity.org/trop_ozone.html.

[11] Z. Feng, E. Paoletti, A. Bytnerowicz, and H. Harmens, “Ozone and plants,” *Environmental Pollution*, vol. 30, p. 1e2, 2015.

[12] C. Cabot, J. V. Sibole, J. Barceló, and C. Poschenrieder, “Lessons from crop plants struggling with salinity,” *Plant Science*, vol. 226, pp. 2–13, 2014.

[13] L. Wang, Z. Chen, H. Shang, J. Wang, and P. Zhang, “Impact of simulated acid rain on soil microbial community function in Masson pine seedlings,” *Electronic Journal of Biotechnology*, vol. 17, no. 5, pp. 199–203, 2014.

[14] G. Rusu-Zagar, C. Rusu-Zagar, I. Iorga, and A. Iorga, “Air Pollution Particles PM 10, PM 2, 5 and the Tropospheric Ozone Effects on Human Health,” *Procedia-Social and*

Behavioral Sciences, vol. 92, pp. 826–831, 2013.

[15] M. S. Gustin, R. Fine, M. Miller, D. Jaffe, and J. Burley, "The Nevada Rural Ozone Initiative (NVROI): Insights to understanding air pollution in complex terrain," *Science of The Total Environment*, 2015.

[16] R. van Zelm, M. A. Huijbregts, H. A. den Hollander, H. A. van Jaarsveld, F. J. Sauter, J. Struijs, H. J. van Wijnen, and D. van de Meent, "European characterization factors for human health damage of PM 10 and ozone in life cycle impact assessment," *Atmospheric Environment*, vol. 42, no. 3, pp. 441–453, 2008.

[17] M. R. Heal, C. Heaviside, R. M. Doherty, M. Vieno, D. S. Stevenson, and S. Vardoulakis, "Health burdens of surface ozone in the UK for a range of future scenarios," *Environment international*, vol. 61, pp. 36–44, 2013.

[18] S. Sousa, M. Alvim-Ferraz, and F. Martins, "Health effects of ozone focusing on childhood asthma: what is now known-a review from an epidemiological point of view," *Chemosphere*, vol. 90, no. 7, pp. 2051–2058, 2013.

[19] Ó. Borrego-Hernández, J. A. Garc'\ia-Reynoso, M. M. Ojeda-Ram'\irez, and M. Suárez-Lastra, "Retrospective health impact assessment for ozone pollution in Mexico City from 1991 to 2011," *Atmósfera*, vol. 27, no. 3, pp. 261–271, 2014.

[20] J. K. Vanos, C. Hebbern, and S. Cakmak, "Risk assessment for cardiovascular and respiratory mortality due to air pollution and synoptic meteorology in 10 Canadian cities," *Environmental Pollution*, vol. 185, pp. 322–332, 2014.

[21] U. Braukmann and D. Böhme, "Salt pollution of the middle and lower sections of the river Werra (Germany) and its impact on benthic macroinvertebrates," *Limnologica-Ecology and Management of Inland Waters*, vol. 41, no. 2, pp. 113–124, 2011.

[22] V. R. Kelly, G. M. Lovett, K. C. Weathers, S. E. Findlay, D. L. Strayer, D. J. Burns, and G. E. Likens, "Long-term sodium chloride retention in a rural watershed: legacy effects of road salt on streamwater concentration," *Environmental science \& technology*, vol. 42, no. 2, pp. 410–415, 2007.

[23] D. W. Ostendorf, E. S. Hinlein, C. Rotaru, and D. J. DeGroot, "Contamination of groundwater by outdoor highway deicing agent storage," *Journal of Hydrology*, vol. 326, no. 1, pp. 109–121, 2006.

[24] R. J. Sibert, C. M. Koretsky, and D. A. Wyman, "Cultural meromixis: Effects of road salt on the chemical stratification of an urban kettle lake," *Chemical Geology*, vol. 395, pp. 126–137, 2015.

[25] S. M. Green, R. Machin, and M. S. Cresser, "Effect of long-term changes in soil chemistry induced by road salt applications on N-transformations in roadside soils," *Environmental pollution*, vol. 152, no. 1, pp. 20–31, 2008.

[26] B. Blasius and R. Merritt, “Field and laboratory investigations on the effects of road salt (NaCl) on stream macroinvertebrate communities,” *Environmental Pollution*, vol. 120, no. 2, pp. 219–231, 2002.

[27] M. Meriano, N. Eyles, and K. W. Howard, “Hydrogeological impacts of road salt from Canada’s busiest highway on a Lake Ontario watershed (Frenchman’s Bay) and lagoon, City of Pickering,” *Journal of contaminant hydrology*, vol. 107, no. 1, pp. 66–81, 2009.

[28] E. V. Novotny, D. Murphy, and H. G. Stefan, “Increase of urban lake salinity by road deicing salt,” *Science of the Total Environment*, vol. 406, no. 1, pp. 131–144, 2008.

[29] E.-L. Thunqvist, “Regional increase of mean chloride concentration in water due to the application of deicing salt,” *Science of the Total Environment*, vol. 325, no. 1, pp. 29–37, 2004.

[30] K. W. Howard and H. Maier, “Road de-icing salt as a potential constraint on urban growth in the Greater Toronto Area, Canada,” *Journal of contaminant hydrology*, vol. 91, no. 1, pp. 146–170, 2007.

[31] D. D. Williams, N. E. Williams, and Y. Cao, “Road salt contamination of groundwater in a major metropolitan area and development of a biological index to monitor its impact,” *Water Research*, vol. 34, no. 1, pp. 127–138, 2000.

[32] A. H. Schweiger, V. Audorff, and C. Beierkuhnlein, “Salt in the wound: The interfering effect of road salt on acidified forest catchments,” *Science of The Total Environment*, vol. 532, pp. 595–604, 2015.

[33] S. Sudalma, P. Purwanto, and L. W. Santoso, “The Effect of SO 2 and NO 2 from Transportation and Stationary Emissions Sources to SO 4 2- and NO 3- in Rain Water in Semarang,” *Procedia Environmental Sciences*, vol. 23, pp. 247–252, 2015.

[34] S. V. Krupa, “Sampling and physico-chemical analysis of precipitation: a review,” *Environmental Pollution*, vol. 120, no. 3, pp. 565–594, 2002.

[35] C. González and B. Aristizábal, “Acid rain and particulate matter dynamics in a mid-sized Andean city: The effect of rain intensity on ion scavenging,” *Atmospheric Environment*, vol. 60, pp. 164–171, 2012.

[36] X. Zhang, H. Jiang, J. Jin, X. Xu, and Q. Zhang, “Analysis of acid rain patterns in northeastern China using a decision tree method,” *Atmospheric environment*, vol. 46, pp. 590–596, 2012.

[37] F. C. Menz and H. M. Seip, “Acid rain in Europe and the United States: an update,” *Environmental Science \& Policy*, vol. 7, no. 4, pp. 253–265, 2004.

[38] T. Wang, L. Jin, Z. Li, and K. Lam, “A modeling study on acid rain and recommended emission control strategies in China,” *Atmospheric Environment*, vol. 34, no. 26, pp.

4467–4477, 2000.

[39] J. Chen, W.-H. Wang, T.-W. Liu, F.-H. Wu, and H.-L. Zheng, “Photosynthetic and antioxidant responses of Liquidambar formosana and Schima superba seedlings to sulfuric-rich and nitric-rich simulated acid rain,” *Plant Physiology and Biochemistry*, vol. 64, pp. 41–51, 2013.

[40] S. Chen, X. Shen, Z. Hu, H. Chen, Y. Shi, and Y. Liu, “Effects of simulated acid rain on soil CO 2 emission in a secondary forest in subtropical China,” *Geoderma*, vol. 189, pp. 65–71, 2012.

[41] W. Tsujita, S. Kaneko, T. Ueda, H. Ishida, and T. Moriizumi, “Sensor-based air-pollution measurement system for environmental monitoring network,” in *TRANSDUCERS, Solid-State Sensors, Actuators and Microsystems, 12th International Conference on, 2003*, 2003, vol. 1, pp. 544–547.

[42] F. Pierce and T. Elliott, “Regional and on-farm wireless sensor networks for agricultural systems in Eastern Washington,” *Computers and electronics in agriculture*, vol. 61, no. 1, pp. 32–43, 2008.

[43] A.-J. Garcia-Sanchez, F. Garcia-Sanchez, and J. Garcia-Haro, “Wireless sensor network deployment for integrating video-surveillance and data-monitoring in precision agriculture over distributed crops,” *Computers and Electronics in Agriculture*, vol. 75, no. 2, pp. 288–303, 2011.

[44] J. L. Riquelme, F. Soto, J. Suardiaz, P. Sánchez, A. Iborra, and J. Vera, “Wireless sensor networks for precision horticulture in Southern Spain,” *Computers and Electronics in Agriculture*, vol. 68, no. 1, pp. 25–35, 2009.

[45] N. Wang, N. Zhang, and M. Wang, “Wireless sensors in agriculture and food industry—Recent development and future perspective,” *Computers and electronics in agriculture*, vol. 50, no. 1, pp. 1–14, 2006.

[46] N. W. Quinn, R. Ortega, P. J. Rahilly, and C. W. Royer, “Use of environmental sensors and sensor networks to develop water and salinity budgets for seasonal wetland real-time water quality management,” *Environmental Modelling \& Software*, vol. 25, no. 9, pp. 1045–1058, 2010.

[47] A. G. Volkov, *Plant electrophysiology: theory and methods*. Springer, 2006.

[48] J. B. Sanderson, “Note on the electrical phenomena which accompany irritation of the leaf of Dionaea muscipula,” *Proceedings of the Royal Society of London*, vol. 21, no. 139–147, pp. 495–496, 1872.

[49] C. Darwin and S. F. Darwin, *Insectivorous plants*, 2nd ed. London: J. Murray, 1888.

[50] E. Davies, “New functions for electrical signals in plants,” *New Phytologist*, vol. 161, no. 3, pp. 607–610, 2004.

[51] J. C. Bose, *The physiology of photosynthesis*. Long Mans, Green And Co; London, 1924.

[52] B. G. Pickard, “Action potentials in higher plants,” *The Botanical Review*, vol. 39, no. 2, pp. 172–201, 1973.

[53] J. Schroeder, R. Hedrich, and J. Fernandez, “Potassium-selective single channels in guard cell protoplasts of Vicia faba,” *Nature*, vol. 312, no. 5992, pp. 361–362, 1984.

[54] X. Yan, Z. Wang, L. Huang, C. Wang, R. Hou, Z. Xu, and X. Qiao, “Research progress on electrical signals in higher plants,” *Progress in Natural Science*, vol. 19, no. 5, pp. 531–541, 2009.

[55] R. Wayne, “The excitability of plant cells: with a special emphasis on characean internodal cells,” *The Botanical Review*, vol. 60, no. 3, pp. 265–367, 1994.

[56] M. A. Hall and others, *Plant structure, function and adaptation.* Macmillan Press Ltd., 1976.

[57] A. M. Hetherington and F. I. Woodward, “The role of stomata in sensing and driving environmental change,” *Nature*, vol. 424, no. 6951, pp. 901–908, 2003.

[58] F. Tardieu and W. J. Davies, “Stomatal response to abscisic acid is a function of current plant water status,” *Plant physiology*, vol. 98, no. 2, pp. 540–545, 1992.

[59] A. Pask, J. Pietragalla, D. Mullan, and M. Reynolds, *Physiological breeding II: a field guide to wheat phenotyping*. CIMMYT, 2012.

[60] J. Riikonen, L. Syrjälä, I. Tulva, P. Mänd, E. Oksanen, M. Poteri, and E. Vapaavuori, “Stomatal characteristics and infection biology of< i> Pyrenopeziza betulicola</i> in< i> Betula pendula</i> trees grown under elevated CO< sub> 2</sub> and O< sub> 3</sub>,” *Environmental Pollution*, vol. 156, no. 2, pp. 536–543, 2008.

[61] M. R. G. Roelfsema and R. Hedrich, “Studying guard cells in the intact plant: modulation of stomatal movement by apoplastic factors,” *New Phytologist*, vol. 153, no. 3, pp. 425–431, 2008.

[62] A. De Angeli, S. Thomine, J. M. Frachisse, G. Ephritikhine, F. Gambale, and H. Barbier-Brygoo, “Anion channels and transporters in plant cell membranes,” *FEBS letters*, vol. 581, no. 12, pp. 2367–2374, 2007.

[63] L.-M. Fan, Z. Zhao, and S. M. Assmann, “Guard cells: a dynamic signaling model,” *Current opinion in plant biology*, vol. 7, no. 5, pp. 537–546, 2004.

[64] S. Zimmermann and H. Sentenac, “Plant ion channels: from molecular structures to physiological functions,” *Current opinion in plant biology*, vol. 2, no. 6, pp. 477–482, 1999.

[65] Z.-M. Pei, V. M. Baizabal-Aguirre, G. J. Allen, and J. I. Schroeder, “A transient outward-rectifying K+ channel current down-regulated by cytosolic Ca2+ in Arabidopsis thaliana guard cells,” *Proceedings of the National Academy of Sciences*, vol. 95, no. 11, pp. 6548–6553, 1998.

[66] M. R. McAinsh, C. Brownlee, and A. M. Hetherington, “Abscisic acid-induced elevation of guard cell cytosolic Ca2+ precedes stomatal closure,” *Letters to Nature*, vol. 343, 1990.

[67] J. I. Schroeder, G. J. Allen, V. Hugouvieux, J. M. Kwak, and D. Waner, “Guard cell signal transduction,” *Annual review of plant biology*, vol. 52, no. 1, pp. 627–658, 2001.

[68] P. H. Raven, R. F. Evert, and S. E. Eichhorn, *Biology of plants*. Macmillan, 2005.

[69] M. J. Chrispeels, N. M. Crawford, and J. I. Schroeder, “Proteins for transport of water and mineral nutrients across the membranes of plant cells,” *The Plant Cell*, vol. 11, no. 4, pp. 661–675, 1999.

[70] E. Krol and K. Trebacz, “Ways of ion channel gating in plant cells,” *Annals of Botany*, vol. 86, no. 3, pp. 449–469, 2000.

[71] M. Tester, “Tansley Review No. 21 Plant ion channels: whole-cell and single channel studies,” *New Phytologist*, vol. 114, no. 3, pp. 305–340, 2006.

[72] B. J. Atwell, P. E. Kriedemann, and C. G. Turnbull, *Plants in action: adaptation in nature, performance in cultivation*. Palgrave MacMillan, 1999.

[73] N. Masih and P. C. Misra, “Ca< sup> 2+</sup> uptake and plasma membrane depolarization associated with blue light-sensitive exogenous NADH oxidation by< i> Cuscuta</i> protoplasts,” *Journal of plant physiology*, vol. 158, no. 1, pp. 29–34, 2001.

[74] J. I. Schroeder and P. Thuleau, “Ca2+ Channels in Higher Plant Cells.,” *The Plant Cell*, vol. 3, no. 6, p. 555, 1991.

[75] H. Miedema, J. H. F. Bothwell, C. Brownlee, and J. M. Davies, “Calcium uptake by plant cells-channels and pumps acting in concert,” *Trends in plant science*, vol. 6, no. 11, pp. 514–519, 2001.

[76] X. Zhang, Z. Shen, J. Sun, Y. Yu, S. Deng, Z. Li, C. Sun, J. Zhang, R. Zhao, X. Shen, and others, “NaCl-elicited, vacuolar Ca 2+ release facilitates prolonged cytosolic Ca 2+ signaling in the salt response of Populus euphratica cells,” *Cell calcium*, vol. 57, no. 5, pp. 348–365, 2015.

[77] W.-G. Choi, M. Toyota, S.-H. Kim, R. Hilleary, and S. Gilroy, “Salt stress-induced Ca2+ waves are associated with rapid, long-distance root-to-shoot signaling in plants,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 17, pp. 6497–6502,

2014.

[78] J. R. Dinneny and M. F. Yanofsky, “Vascular patterning: xylem or phloem?,” *Current biology*, vol. 14, no. 3, pp. R112–R114, 2004.

[79] A. J. van Bel, M. Knoblauch, A. C. Furch, and J. B. Hafke, “(Questions) n on phloem biology. 1. Electropotential waves, Ca 2+ fluxes and cellular cascades along the propagation pathway,” *Plant science*, vol. 181, no. 3, pp. 210–218, 2011.

[80] A. G. Volkov, “Green plants: electrochemical interfaces,” *Journal of Electroanalytical Chemistry*, vol. 483, no. 1, pp. 150–156, 2000.

[81] J. Fromm and S. Lautner, “Electrical signals and their physiological significance in plants,” *Plant, cell & environment*, vol. 30, no. 3, pp. 249–257, 2007.

[82] A. G. Volkov and D. R. A. Ranatunga, “Plants as environmental biosensors,” *Plant signaling \& behavior*, vol. 1, no. 3, pp. 105–115, 2006.

[83] R. Stahlberg, R. E. Cleland, and E. Van Volkenburgh, “Slow Wave Potentials-a Propagating Electrical Signal,” *Communication in Plants: Neuronal Aspects of Plant Life*, p. 291, 2006.

[84] H. Dziubinska, K. Trebacz, and T. Zawadzki, “Transmission route for action potentials and variation potentials in Helianthus annuus L.,” *Journal of plant physiology*, vol. 158, no. 9, pp. 1167–1172, 2001.

[85] D.-J. Zhao, Y. Chen, Z.-Y. Wang, L. Xue, T.-L. Mao, Y.-M. Liu, Z.-Y. Wang, and L. Huang, “High-resolution non-contact measurement of the electrical activity of plants in situ using optical recording,” *Scientific reports*, vol. 5, 2015.

[86] D.-J. Zhao, Z.-Y. Wang, J. Li, X. Wen, A. Liu, L. Huang, X.-D. Wang, R.-F. Hou, and C. Wang, “Recording extracellular signals in plants: A modeling and experimental study,” *Mathematical and Computer Modelling*, vol. 58, no. 3, pp. 556–563, 2013.

[87] S. Lautner, T. E. E. Grams, R. Matyssek, and J. Fromm, “Characteristics of electrical signals in poplar and responses in photosynthesis,” *Plant Physiology*, vol. 138, no. 4, pp. 2200–2209, 2005.

[88] “Patch clamp technique.” [Online]. Available: http://plantsinaction.science.uq.edu.au/edition1/?q=content/4-2-3-patch-clamping.

[89] Y. Qin, L. Huang, A. Liu, D. Zhao, Z. Wang, Y. Liu, and T. Mao, “Visualization of synchronous propagation of plant electrical signals using an optical recoding method,” *Mathematical and Computer Modelling*, 2011.

[90] M. R. G. Roelfsema, R. Steinmeyer, M. Staal, and R. Hedrich, “Single guard cell recordings in intact plants: light-induced hyperpolarization of the plasma membrane,” *The Plant Journal*, vol. 26, no. 1, pp. 1–13, 2001.

[91] L. A. Gurovich and P. Hermosilla, “Electric signalling in fruit trees in response to water applications and light-darkness conditions,” *Journal of plant physiology*, vol. 166, no. 3, pp. 290–300, 2009.

[92] A. G. Volkov, T. C. Dunkley, S. A. Morgan, D. Ruff, Y. L. Boyce, and A. J. Labady, “Bioelectrochemical signaling in green plants induced by photosensory systems,” *Bioelectrochemistry*, vol. 63, no. 1, pp. 91–94, 2004.

[93] I. Ruberti, G. Sessa, A. Ciolfi, M. Possenti, M. Carabelli, and G. Morelli, “Plant adaptation to dynamically changing environment: the shade avoidance response,” *Biotechnology Advances*, vol. 30, no. 5, pp. 1047–1058, 2012.

[94] L. Huché-Thélier, L. Crespel, J. Le Gourrierec, P. Morel, S. Sakr, and N. Leduc, “Light signaling and plant responses to blue and UV radiations—Perspectives for applications in horticulture,” *Environmental and Experimental Botany*, 2015.

[95] P. Oyarce and L. Gurovich, “Electrical signals in avocado trees: Responses to light and water availability conditions,” *Plant signaling \& behavior*, vol. 5, no. 1, pp. 34–41, 2010.

[96] L. Tian, Q. Meng, L. Wang, J. Dong, and H. Wu, “Research on the Effect of Electrical Signals on Growth of Sansevieria under Light-Emitting Diode (LED) Lighting Environment,” *PloS one*, vol. 10, no. 6, p. e0131838, 2015.

[97] R. F. Rachel Casiday, “Acid Rain - Inorganic Reactions Experiment,” 1998.

[98] T. S. Amy and P. H. C. Jessica, “THe Effect of Acid Rain on Plants,” 2002.

[99] O. Klimenko and N. Klimenko, “Nutrient concentrations in leaves of peach trees subjected to acid rain,” in *Plant Nutrition*, Springer, 2001, pp. 926–927.

[100] J. Mwesigwa, D. J. Collins, and A. G. Volkov, “Electrochemical signaling in green plants: effects of 2, 4-dinitrophenol on variation and action potentials in soybean,” *Bioelectrochemistry*, vol. 51, no. 2, pp. 201–205, 2000.

[101] T. Shvetsova, J. Mwesigwa, and A. G. Volkov, “Plant electrophysiology: FCCP induces action potentials and excitation waves in soybean,” *Plant Science*, vol. 161, no. 5, pp. 901–909, 2001.

[102] A. G. Volkov, D. J. Collins, and J. Mwesigwa, “Plant electrophysiology: pentachlorophenol induces fast action potentials in soybean,” *Plant Science*, vol. 153, no. 2, pp. 185–190, 2000.

[103] A. Labady, D. Thomas, T. Shvetsova, and A. G. Volkov, “Plant bioelectrochemistry: effects of CCCP on electrical signaling in soybean,” *Bioelectrochemistry*, vol. 57, no. 1, pp. 47–53, 2002.

[104] M. Jerominek and R. Claßen-Bockhoff, “Electrical Signals in Prayer Plants (Marantaceae)? Insights into the Trigger Mechanism of the Explosive Style

Movement," 2015.

[105] A. B. Stephan and J. I. Schroeder, "Plant salt stress status is transmitted systemically via propagating calcium waves," *Proceedings of the National Academy of Sciences*, vol. 111, no. 17, pp. 6126–6127, 2014.

[106] A. G. Volkov, R. D. Lang, and M. I. Volkova-Gugeshashvili, "Electrical signaling in Aloe vera induced by localized thermal stress," *Bioelectrochemistry*, vol. 71, no. 2, pp. 192–197, 2007.

[107] P. Oyarce and L. Gurovich, "Evidence for the transmission of information through electric potentials in injured avocado trees," *Journal of plant physiology*, vol. 168, no. 2, pp. 103–108, 2011.

[108] Z. Y. Wang, Q. Leng, L. Huang, L. L. Zhao, Z. L. Xu, R. F. Hou, and C. Wang, "Monitoring system for electrical signals in plants in the greenhouse and its applications," *Biosystems Engineering*, vol. 103, no. 1, pp. 1–11, 2009.

[109] L. R'\ios-Rojas, F. Tapia, and L. A. Gurovich, "Electrophysiological assessment of water stress in fruit-bearing woody plants," *Journal of plant physiology*, vol. 171, no. 10, pp. 799–806, 2014.

[110] L. Jingxia and D. Weimin, "Study and evaluation of plant electrical signal processing method," in *Image and Signal Processing (CISP), 2011 4th International Congress on*, 2011, vol. 5, pp. 2788–2791.

[111] L. Jingxia and D. Weimin, "Analysis of electric signal of plant based on lifting wavelet and correlation," in *Multimedia and Signal Processing (CMSP), 2011 International Conference on*, 2011, vol. 1, pp. 223–227.

[112] Lan-zhou Wang, L. Hai-xia, and L. Qiao, "Studies on the plant electric wave signal by the wavelet analysis," in *Journal of Physics: Conference Series*, 2007, vol. 48, no. 1, p. 1367.

[113] J. Lu and W. Ding, "The feature extraction of plant electrical signal based on wavelet packet and neural network," in *Automatic Control and Artificial Intelligence (ACAI 2012), International Conference on*, 2012, pp. 2119–2122.

[114] Y. Liu, Z. Junmei, L. Xiaoli, K. Jiangming, and Y. Kai, "The Research of Plants' Water Stress Acoustic Emission Signal Processing Methods," in *Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on*, 2011, vol. 3, pp. 922–925.

[115] L. Wang and Q. Li, "Weak electrical signals of the jasmine processed by RBF neural networks forecast," in *Biomedical Engineering and Informatics (BMEI), 2010 3rd International Conference on*, 2010, vol. 7, pp. 3095–3099.

[116] L. Huang, Z.-Y. Wang, L.-L. Zhao, D. Zhao, C. Wang, Z.-L. Xu, R.-F. Hou, and X.-J. Qiao, "Electrical signal measurement in plants using blind source separation with

independent component analysis," *Computers and Electronics in Agriculture*, vol. 71, pp. S54–S59, 2010.

[117] E. Darko, P. Heydarizadeh, B. Schoefs, and M. R. Sabzalian, "Photosynthesis under artificial light: the shift in primary and secondary metabolism," *Phil. Trans. R. Soc. B*, vol. 369, no. 1640, p. 20130243, 2014.

[118] R. Bula, R. Morrow, T. Tibbitts, D. Barta, R. Ignatius, and T. Martin, "Light-emitting diodes as a radiation source for plants," *HortScience*, vol. 26, no. 2, pp. 203–205, 1991.

[119] V. Martineau, M. Lefsrud, M. T. Naznin, and D. A. Kopsell, "Comparison of light-emitting diode and high-pressure sodium light treatments for hydroponics growth of Boston lettuce," *HortScience*, vol. 47, no. 4, pp. 477–482, 2012.

[120] R. L. Chang, L. Ghamsari, A. Manichaikul, E. F. Hom, S. Balaji, W. Fu, Y. Shen, T. Hao, B. \O Palsson, K. Salehi-Ashtiani, and others, "Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism," *Molecular systems biology*, vol. 7, no. 1, p. 518, 2011.

[121] B. Schoefs, "Chlorophyll and carotenoid analysis in food products. Properties of the pigments and methods of analysis," *Trends in food science \& technology*, vol. 13, no. 11, pp. 361–371, 2002.

[122] A. Schwartz and E. Zeiger, "Metabolic energy for stomatal opening. Roles of photophosphorylation and oxidative phosphorylation," *Planta*, vol. 161, no. 2, pp. 129–136, 1984.

[123] G. D. Goins, N. C. Yorio, M. M. Sanwo-Lewandowski, and C. S. Brown, "Life cycle experiments with Arabidopsis grown under red light-emitting diodes (LEDs).," *Life support \& biosphere science: international journal of earth space*, vol. 5, no. 2, pp. 143–149, 1997.

[124] M. Beilby and V. Shepherd, "Modeling the current-voltage characteristics of charophyte membranes. II. The effect of salinity on membranes of Lamprothamnium papulosum," *The Journal of membrane biology*, vol. 181, no. 2, pp. 77–89, 2001.

[125] V. Sukhov and V. Vodeneev, "A mathematical model of action potential in cells of vascular plants," *Journal of Membrane Biology*, vol. 232, no. 1–3, pp. 59–67, 2009.

[126] H. Mummert and D. Gradmann, "Action potentials inAcetabularia: Measurement and simulation of voltage-gated fluxes," *The Journal of membrane biology*, vol. 124, no. 3, pp. 265–273, 1991.

[127] M. Beilby, R. Keynes, and N. Walker, "Cl $\^-$ Channels in Chara [and Discussion]," *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, vol. 299, no. 1097, pp. 435–445, 1982.

[128] V. Sukhov, E. Akinchits, L. Katicheva, and V. Vodeneev, “Simulation of Variation Potential in Higher Plant Cells,” *The Journal of membrane biology*, pp. 1–10, 2013.

[129] D. Gradmann and J. Hoffstadt, “Electrocoupling of ion transporters in plants: interaction with internal ion concentrations,” *The Journal of membrane biology*, vol. 166, no. 1, pp. 51–59, 1998.

[130] B. Stankovi’c, D. L. Witters, T. Zawadzki, and E. Davies, “Action potentials and variation potentials in sunflower: an analysis of their relationships and distinguishing characteristics,” *Physiologia Plantarum*, vol. 103, no. 1, pp. 51–58, 1998.

[131] L. Ljung, “System identification,” in *Signal Analysis and Prediction*, Springer, 1998, pp. 163–173.

[132] J. P. Norton, *An introduction to identification*. Dover Publications, 2009.

[133] T. Söderström and P. Stoica, *System identification*. Prentice-Hall, Inc., 1988.

[134] L. Ljung, “System identification toolbox,” *The Matlab user’s guide*, 2012.

[135] T. Hastie, R. Tibshirani, and J. Friedman, “The elements of statistical learning. 2001,” *NY Springer*, 2001.

[136] S. Das, S. Mukherjee, I. Pan, and A. Gupta, “Identification of the core temperature in a fractional order noisy environment for thermal feedback in nuclear reactors,” in *Students’ Technology Symposium (TechSym), 2011 IEEE*, 2011, pp. 180–186.

[137] “http://labjack.com/support/ei-1040/datasheet.” [Online]. Available: http://labjack.com/support/ei-1040/datasheet.

[138] “http://sine.ni.com/nips/cds/view/p/lang/it/nid/201986.” [Online]. Available: http://sine.ni.com/nips/cds/view/p/lang/it/nid/201986.

[139] “http://www.ni.com/labview.” [Online]. Available: http://www.ni.com/labview.

[140] P. Il’\ik, V. Hlavá\vcková, P. Krch\vnák, and J. Nau\vs, “A low-noise multi-channel device for the monitoring of systemic electrical signal propagation in plants,” *Biologia Plantarum*, vol. 54, no. 1, pp. 185–190, 2010.

[141] L. Wang and J. Ding, “Processing on information fusion of weak electrical signals in plants,” in *Information and Computing (ICIC), 2010 Third International Conference on*, 2010, vol. 2, pp. 21–24.

[142] E. F. Cabral, P. C. Pecora, A. I. C. Arce, A. R. B. Tech, and E. J. X. Costa, “The oscillatory bioelectrical signal from plants explained by a simulated electrical model and tested using Lempel-Ziv complexity,” *Computers and Electronics in Agriculture*, vol. 76, no. 1, pp. 1–5, 2011.

[143] “Lux to PAR.” [Online]. Available: http://www.apogeeinstruments.co.uk/conversion-ppf-to-lux/.

[144] “Authometion STARTER.” [Online]. Available: https://authometion.com/shop/en/15-for-maker.

[145] “Authometion Library.” [Online]. Available: https://authometion.com/shop/en/home/5-starter-kit-plus.html.

[146] “BH1750FVI nodemcu.” [Online]. Available: https://github.com/nodemcu/nodemcu-firmware/blob/master/lua_modules/bh1750/bh1750.lua .

[147] “Thingspeak - Light Sensor Data (Shre Chatterjee).” [Online]. Available: https://thingspeak.com/channels/176749.

[148] S. K. Chatterjee, S. Ghosh, S. Das, V. Manzella, A. Vitaletti, E. Masi, L. Santopolo, S. Mancuso, and K. Maharatna, “Forward and Inverse Modelling Approaches for Prediction of Light Stimulus from Electrophysiological Response in Plants,” *Measurement*, vol. 53, pp. 101–116, 2014.

[149] “http://www.wpiinc.com/blog/2013/05/01/product-information/data-trax-software-for-labscribe/.” [Online]. Available: http://www.wpiinc.com/blog/2013/05/01/product-information/data-trax-software-for-labscribe/.

[150] “http://www.sepra.it/products-linea-generatori-serie-steril250mgo3h-da-aria-6.html.” [Online]. Available: http://www.sepra.it/products-linea-generatori-serie-steril250mgo3h-da-aria-6.html.

[151] C. M. Bishop and others, *Pattern recognition and machine learning*, vol. 1. Springer New York, 2006.

[152] F.-L. Chung, T.-C. Fu, V. Ng, and R. W. Luk, “An evolutionary approach to pattern-based time series segmentation,” *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 5, pp. 471–489, 2004.

[153] P. de Chazal and R. B. Reilly, “A patient-adapting heartbeat classifier using ECG morphology and heartbeat interval features,” *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2535–2543, 2006.

[154] X. Jiang, L. Zhang, Q. Zhao, and S. Albayrak, “ECG arrhythmias recognition system based on independent component analysis feature extraction,” in *TENCON 2006-2006 IEEE Region 10 Conference*, 2006, pp. 1–4.

[155] M. Sinn, K. Keller, and B. Chen, “Segmentation and classification of time series using ordinal pattern distributions,” *The European Physical Journal Special Topics*, vol. 222, no. 2, pp. 587–598, 2013.

[156] P. De Chazal, M. O’Dwyer, and R. B. Reilly, “Automatic classification of heartbeats using ECG morphology and heartbeat interval features,” *IEEE Transactions on*

Biomedical Engineering, vol. 51, no. 7, pp. 1196–1206, 2004.

[157] G. K. Prasad and J. Sahambi, "Classification of ECG arrhythmias using multi-resolution analysis and neural networks," in *TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region*, 2003, vol. 1, pp. 227–231.

[158] S. Barro, R. Ruiz, D. Cabello, and J. Mira, "Algorithmic sequential decision-making in the frequency domain for life threatening ventricular arrhythmias and imitative artefacts: a diagnostic system," *Journal of biomedical engineering*, vol. 11, no. 4, pp. 320–328, 1989.

[159] C. Ye, B. V. Kumar, and M. T. Coimbra, "Heartbeat classification using morphological and dynamic features of ECG signals," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 10, pp. 2930–2941, 2012.

[160] M. Lovri'c, M. Milanovi'c, and M. Stamenkovi'c, "Algorithmic Methods For Segmentation of Time Series: An overview," *JCEBI*, vol. 1, no. 1, pp. 31–53, 2014.

[161] V. Manzella, C. Gaz, A. Vitaletti, E. Masi, L. Santopolo, S. Mancuso, D. Salazar, and J. de las Heras, "Plants as sensing devices: the PLEASED experience," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, 2013, p. 76.

[162] D. Kugiumtzis and A. Tsimpiris, "Measures of analysis of time series (MATS): a MATLAB toolkit for computation of multiple measures on time series data bases," *Journal of Statistical Software*, vol. 33, 2010.

[163] B. Hjorth, "Time domain descriptors and their relation to a particular model for generation of EEG activity," *CEAN-Computerized EEG analysis*, pp. 3–8, 1975.

[164] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalography and clinical neurophysiology*, vol. 29, no. 3, pp. 306–310, 1970.

[165] B. Hjorth, "The physical significance of time domain descriptors in EEG analysis," *Electroencephalography and clinical neurophysiology*, vol. 34, no. 3, pp. 321–325, 1973.

[166] J.-M. Lee, D.-J. Kim, I.-Y. Kim, K.-S. Park, and S. I. Kim, "Detrended fluctuation analysis of EEG in sleep apnea using MIT/BIH polysomnography data," *Computers in Biology and Medicine*, vol. 32, no. 1, pp. 37–47, 2002.

[167] A. K. Goli'nska, "Detrended Fluctuation Analysis (DFA) in biomedical signal processing: selected examples," *Logical, Statistical and Computer Methods in Medicine*, vol. 29(42), 2012.

[168] N. Kannathal, U. R. Acharya, C. Lim, and P. Sadasivan, "Characterization of EEG—A comparative study," *Computer methods and Programs in Biomedicine*, vol. 80, no. 1, pp. 17–23, 2005.

[169] J. Mielniczuk and P. Wojdy\l\lo, “Estimation of Hurst exponent revisited,” *Computational Statistics \& Data Analysis*, vol. 51, no. 9, pp. 4510–4525, 2007.

[170] R. Q. Quiroga, O. A. Rosso, E. Ba\csar, and M. Schürmann, “Wavelet entropy in event-related potentials: a new method shows ordering of EEG oscillations,” *Biological Cybernetics*, vol. 84, no. 4, pp. 291–299, 2001.

[171] O. A. Rosso, S. Blanco, J. Yordanova, V. Kolev, A. Figliola, M. Schürmann, and E. Ba\csar, “Wavelet entropy: a new tool for analysis of short duration brain electrical signals,” *Journal of neuroscience methods*, vol. 105, no. 1, pp. 65–75, 2001.

[172] L. Zunino, D. Perez, M. Garavaglia, and O. Rosso, “Wavelet entropy of stochastic processes,” *Physica A: Statistical Mechanics and its Applications*, vol. 379, no. 2, pp. 503–512, 2007.

[173] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications: with R examples*. Springer, 2010.

[174] C. Chen, L. Pau, and P. Wang, “Statistical pattern recognition,” 1973.

[175] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras, *Introduction to Pattern Recognition: A Matlab Approach: A Matlab Approach*. Academic Press, 2010.

[176] C. Sammut and G. I. Webb, *Encyclopedia of machine learning*. Springer, 2010.

[177] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley \& Sons, 2012.

[178] L. Sörnmo and P. Laguna, *Bioelectrical signal processing in cardiac and neurological applications*. Academic Press, 2005.

[179] S. K. Chatterjee, S. Das, K. Maharatna, E. Masi, L. Santopolo, S. Mancuso, and A. Vitaletti, “Exploring strategies for classification of external stimuli using statistical features of the plant electrical response,” *Journal of The Royal Society Interface*, vol. 12, no. 104, p. 20141225, 2015.

[180] “PLEASED FP7 Project.” [Online]. Available: http://pleased-fp7.eu.

[181] N. Uniyal, H. Eskandari, P. Abolmaesumi, S. Sojoudi, P. Gordon, L. Warren, R. Rohling, S. Salcudean, and M. Moradi, “Ultrasound RF time series for classification of breast lesions.,” *IEEE Transactions on Medical Imaging*, vol. 34, no. 2, pp. 652 – 661, 2014.

[182] B. Lindgren, A. V. Johansson, and Y. Tsuji, “Universality of probability density distributions in the overlap region in high Reynolds number turbulent boundary layers,” *Physics of Fluids (1994-present)*, vol. 16, no. 7, pp. 2587–2591, 2004.

[183] M. C. Teich, “Fractal character of the auditory neural spike train,” *Biomedical Engineering, IEEE Transactions on*, vol. 36, no. 1, pp. 150–160, 1989.

[184] K. J. Blinowska and J. Zygierewicz, *Practical Biomedical Signal Analysis Using MATLAB*. CRC Press, 2011.

[185] E. L. Allwein, R. E. Schapire, and Y. Singer, “Reducing multiclass to binary: A unifying approach for margin classifiers,” *The Journal of Machine Learning Research*, vol. 1, pp. 113–141, 2001.

[186] A. Fernández, V. López, M. Galar, M. J. Del Jesus, and F. Herrera, “Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches,” *Knowledge-based systems*, vol. 42, pp. 97–110, 2013.

[187] S. Hashemi, Y. Yang, Z. Mirzamomen, and M. Kangavari, “Adapted one-versus-all decision trees for data stream classification,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, no. 5, pp. 624–637, 2009.

[188] R. Debnath, N. Takahide, and H. Takahashi, “A decision based one-against-one method for multi-class support vector machine,” *Pattern Analysis and Applications*, vol. 7, no. 2, pp. 164–175, 2004.

[189] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, “An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes,” *Pattern Recognition*, vol. 44, no. 8, pp. 1761–1776, 2011.

[190] A. Rocha and S. Klein Goldenstein, “Multiclass from binary: Expanding one-versus-all, one-versus-one and ecoc-based approaches,” *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 25, no. 2, pp. 289–302, 2014.

[191] J. A. Sáez, M. Galar, J. Luengo, and F. Herrera, “Tackling the problem of classification with noisy data using Multiple Classifier Systems: Analysis of the performance and robustness,” *Information Sciences*, vol. 247, pp. 1–20, 2013.

[192] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches,” *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 4, pp. 463–484, 2012.

[193] A. Maratea, A. Petrosino, and M. Manzo, “Adjusted F-measure and kernel scaling for imbalanced data learning,” *Information Sciences*, vol. 257, pp. 331–341, 2014.

[194] C. X. Ling, J. Huang, and H. Zhang, “AUC: a better measure than accuracy in comparing learning algorithms,” in *Advances in Artificial Intelligence*, Springer, 2003, pp. 329–341.

[195] M. Sokolova, N. Japkowicz, and S. Szpakowicz, “Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation,” in *AI 2006: Advances*

in Artificial Intelligence, Springer, 2006, pp. 1015–1021.

[196] V. Garcia, R. A. Mollineda, and J. S. Sánchez, "Index of balanced accuracy: A performance measure for skewed class distributions," in *Pattern Recognition and Image Analysis*, Springer, 2009, pp. 441–448.

[197] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, "Empowering difficult classes with a similarity-based aggregation in multi-class classification problems," *Information Sciences*, vol. 264, pp. 135–157, 2014.

[198] S. Das, B. J. Ajiwibawa, S. K. Chatterjee, S. Ghosh, K. Maharatna, S. Dasmahapatra, A. Vitaletti, E. Masi, and S. Mancuso, "Drift removal in plant electrical signals via IIR filtering using wavelet energy," *Computers and Electronics in Agriculture*, vol. 118, pp. 15–23, 2015.

[199] A. Naftel and S. Khalid, "Classifying spatiotemporal object trajectories using unsupervised learning in the coefficient feature space," *Multimedia Systems*, vol. 12, no. 3, pp. 227–238, 2006.

[200] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.