UNIVERSITY OF NAIROBI

RAPID ASSESSMENT OF CALCIUM CARBIDE RIPENED BANANAS USING MACHINE-LEARNING ASSISTED LASER RAMAN SPECTROSCOPY

BY

ODONGO KENNEDY ONYANGO

I56/12383/2018

A thesis submitted for examination in partial fulfillment of the requirements for the award of the degree of Master of Science in Physics of the University of Nairobi.

DECLARATION

I declare that this thesis is my original work and has not been submitted elsewhere for examination, award of degree or publication. Where other people's work or my own work has been used, this has been properly acknowledged and referenced in accordance with the University of Nairobi's requirements.

Signature……………………Date…25th August 2021……

ODONGO KENNEDY ONYANGO

I56/12383/2018
Department of Physics
University of Nairobi

This thesis has been submitted for examination with our approval as research supervisors:

Signature Date

Dr. M. I. Kaniu
Department of Physics
University of Nairobi
P.O Box 30197-00100,
Nairobi, Kenya.
ikaniu@uonbi.ac.ke

…………………………… 25th August 2021……

Signature Date

Dr. J. M. Wanjohi
Department of Chemistry
University of Nairobi
P.O Box 30197-00100,
Nairobi, Kenya.
jwanjohi@uonbi.ac.ke

………………….… 25th August 2021……

DEDICATION

‘There is no chance, no destiny, no fate that
can hinder or control the firm resolve
of a determined soul’
Ella Wheeler

Dedicated to my wife, Amina and my two adorable sons, Aaron and Ryan for their unwavering support, understanding and patience.

ACKNOWLEDGEMENTS

The work presented in this thesis has presented a challenging but rewarding learning and research experience. I give thanks to God almighty for giving me strength to pursue this and seeing me through the end. Special thanks go to my supervisor, Dr. M. I. Kaniu of the Department of Physics, University of Nairobi (UoN) for proposing this interesting research topic and introducing me to the science of data analytics. I thank Dr. Kaniu for his patience, encouragement and guidance throughout the course of this project. In the same context I would also like to thank Dr. J. M. Wanjohi of the Department of Chemistry, UoN for his expert advice and help on different aspects of the research. Much appreciation to Mr. Dickson Omucheni, the chief technologist, Department of Physics, UoN for the invaluable lessons in spectroscopy.

I am grateful to my employer, The Teachers Service Commission, for giving me a paid study leave for two years to pursue this course. Without their approval, I would not have had ample time to undertake these studies at UoN.

I would be unfair if I failed to acknowledge the great support accorded to me by my colleagues at the Physics Laser Lab, Chiromo campus. In particular, Charles Ndung'u and Robert Ikedi made my stay here interesting and enjoyable. Their discussions and suggestions provided useful insights in this research. I am also grateful to my wife, Amina, for the moral support and continuous encouragement to complete the course.

ABSTRACT

Fruit ripening is usually a natural process in which fruits undergo various chemical and physical changes before they become palatable. New artificial ways of fruit ripening have been developed as a result of recent breakthroughs in agricultural technology mainly to meet market demands and to deal with the logistics of storage and transportation. However, this practice has become a concern because of the human health risks resulting from the uncontrolled use of ripening agents which contain toxic elements. For instance industrial grade calcium carbide has impurities of arsenic and phosphorus as well as other heavy metals. High intake of these elements is known to cause neurological disorders such has cerebral edema and memory loss as well as carcinogenic disorders like cancer of the colon, lungs and peptic ulcers. Hardly is there a method capable of rapidly and non-invasively assessing artificial ripeners in fruits (ARF) with reliability and accuracy. The wet chemical techniques conventionally used such as the various forms of chromatography are time consuming, destructive, costly and involve laborious sample preparations. This work aims at developing a rapid and non-invasive technique for assessing calcium carbide ripened bananas using machine-learning (ML) assisted laser Raman spectroscopy (LRS). In this study, Raman spectra was recorded from naturally and carbide ripened banana samples using a 785 nm laser for excitation. The bananas were ripened using calcium carbide with concentrations ranging from 0.240 g/L to 2.0 g/L. Exploratory analysis using PCA revealed that clustering of the carbide ripened samples was due to the presence of sulfur, acetylene, calcium hydroxide and phosphine impurities contained in CaC_2. These molecules have Raman bands centered at 480 cm^{-1} (S-S bond stretching), 612 cm^{-1} (C-H asymmetric bending), 780 cm^{-1} (O-H bending) and 979 cm^{-1} (P-H stretching) respectively. Classification and quantification of CaC_2 concentrations used in ripening was achieved using the following ML algorithms: support vector machine, artificial neural networks and random forest. High correct classification accuracies were realized (> 85 %) in the ML classification models. Furthermore, the performance of the regression models showed good performance as indicated by high R^2 values (>0.95) and the low RMSEP values (<0.34g/L) when predicting test data sets. Banana samples collected from local markets around Nairobi were found to have been ripened by CaC_2 (up to 1.30 g/L) using the optimized LRS conditions and ML models developed in this work. Therefore, ML-assisted LRS allows for rapid and direct assessment of artificial ripeners in fruits. The findings of this study will aid in the development of spectral libraries for use in food safety analysis procedures involving fruits.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial neural networks
ANOVA	Analysis of variance
ARF	Artificial ripeners in fruits
BP-ANN	Back propagated - artificial neural network
CaC_2	Calcium carbide
CCD	Charged-coupled device
EDXRF	Energy dispersive X-Ray fluorescence
FWHM	Full width half maximum
GC	Gas chromatography
HPLC	High performance liquid chromatography
ICP-AES	Inductively coupled plasma – atomic emission spectrometry
IUPAC	International union of pure and applied chemistry
LOD	Limit of detection
LOQ	Limit of quantification
LRS	Laser Raman spectroscopy
MIR	Mid-infrared
ML	Machine-learning
MS	Mass spectrometry
MSE	Mean square error
mtry	number of variables tried at each split
ND	Neutral density
NIR	Near-infrared
NRF	Naturally ripened fruits
ntree	number of trees in the forest
OOB error	Out-of-bag error
PCA	Principal component analysis
RBK	Radial basis kernel
RF	Random forest
RMSEC	Root mean square error of calibration
RMSEP	Root mean square error of prediction
SEM	Standard error of mean
SNR	Signal-to-noise ratio
SVM	Support vector machine
SVR	Support vector regression

CHAPTER 1

INTRODUCTION

1.1 Background to the Study

Fruit ripening is ordinarily a natural process involving a series of physiological changes in odour, colour and quality of the fruit (Adeniji *et al.*, 2010). The time taken for fruits to ripen naturally varies across different fruits. Thus farmers and retailers ripen fruits artificially using chemicals and other ripening agents, primarily, to meet market demands and achieve uniform ripeness of their fruits. Artificial ripening is also done to deal with logistics of transportation and distribution as ripe fruits cannot be stored in transit for long; farmers harvest fruits whilst still raw and ripen them later (Dhembare, 2013).

The use of artificial ripeners on fruits dates back to ancient China whereby pears were ripened artificially by placing them in closed spaces and burning incense in the chamber (Mehnaz *et al.*, 2013). In the 1920s, researchers observed that unsaturated hydrocarbon gases such as ethylene were responsible for ripening and that plants were able to produce ethylene by themselves (Kendrick, 2009). These observations led to growth of a variety of chemicals and ways of artificially ripening of fruits. Notably, the traditional chemical-free fruit ripening techniques hardly posed any human health risks.

At present, artificial ripening agents commonly used include ethylene gas, calcium carbide, ethephon (2-chloroethylphosphonic acid), ethylene glycol (1,2-thanediol), carbon monoxide, potassium sulphate and oxytocin (Singal *et al.*, 2012). Whereas the natural ripening process is usually initiated with the production of ethylene within mature fruits, artificial ripening agents like ethephon, methanol, and ethylene glycol produce ethylene for accelerating the process in a manner similar to the ethylene produced naturally by fruits (Nagel, 1989). The practice is mostly prevalent during post-harvest stages in the food chain, particularly, during transportation and storage. Fruits which are more prone to artificial ripening include bananas, apples, mangoes, tomatoes and avocados (Dhembare, 2013). These fruits are targeted owing to their widespread demand.

The uncontrolled use of hazardous ripening agents particularly in developing countries poses a great concern to human health. Several studies have shown the detrimental nature to human health of these ripening agents as they cause memory loss, cerebral edema, colonic and lung cancer among others (Kesse *et al*., 2019; Lakade *et al*., 2018; Chandel *et al*., 2018; Kathirvelan *et al*., 2017). As these ripeners could have direct and indirect health hazards, it is imperative to determine their elemental compositions and assess their safety levels within the artificially ripened fruits.

1.1.1 Use of Calcium Carbide as an Artificial Ripener and Associated Health Risks

Among the many chemicals used to ripen fruits artificially calcium carbide (CaC_2) the most preferred due to its fast action, ease of use and availability. Hydrolyzed CaC_2 liberates acetylene which functions as ethylene analogue to influence the ripening of fruits (Bari *et al*., 2018). Equation (1.1) represents the chemical reaction for liberating ethylene and equation (1.2) shows how it accelerates the ripening process.

$$\mathrm{CaC_2(s)} + 2H_2\mathrm{O(l)} \rightarrow \mathrm{Ca(OH)_2(s)} + \mathrm{C}_2H_2\mathrm{(g)} \tag{1.1}$$

$$\text{Unripe (green)banana} + \mathrm{C}_2H_2 \rightarrow \text{Ripe (yellow) banana} \tag{1.2}$$

Notably, the form of CaC_2 that is usually readily available for purposes of artificial ripening of fruits is the impure form. Impurities of calcium phosphide and calcium arsenide have been discovered in industrial grade calcium carbide (Nowshad *et al*., 2018). Phosphine is liberated when calcium phosphide reacts with water and arsine liberated when calcium arsenide reacts with water. These hydrides are fat soluble and can dissolve through the wax surface of fruits and diffuse from the peel to pulp of fruits exposed to them (Haturusihghe *et al*., 2004). These impurities are the ones which largely contribute to making carbide ripened fruits having adverse health effects to humans.

Workers who are directly involved in applying CaC_2 to the fruits bear the highest risk burden of the negative health effects associated with it. They may suffer from conditions such as vomiting and diarrhea, fluid buildup in the lungs, peptic ulcers and colonic cancer caused by exposure to high levels of arsenic and phosphorus. Further, direct exposure to acetylene gas is known to affect the neurological system as it reduces the brain's oxygen supply causing dizziness, seizures, memory loss and cerebral edema (Fattah *et al*., 2010). Dhembare. (2013) reports that the health risks resulting from consumption of CaC_2 ripened fruits may even be passed down genetically, if consumed by pregnant women, resulting to children born with abnormalities.

1.1.2 Challenges Associated with Conventional Methods for Assessing Artificial Ripeners

There are various analyses methods, devices and procedures that are conventionally used for assessing artificial ripeners in fruits that are premised on chemical analysis. These include HPLC-MS, GC and ELISA. These processes are time-consuming, inconvenient in terms of sample preparation and not environmentally friendly (Liu *et al*., 2011). The chromatographic methods have indeed been successful in such kinds of tests but the destructive nature and analytical cost has hindered their widespread and regular use.

Presently, more emphasis and research is geared towards the development of non-destructive techniques which are rapid. Consequently, vibrational spectroscopic techniques such as near-infrared (NIR) spectroscopy, mid-infrared (MIR) spectroscopy and LRS have shown great potential in the fruit industry to check for ARF (Kangas *et al*., 2007). Nonetheless, the applicability of these spectroscopic techniques in studies involving fruits faces challenges arising from the complex nature of fruits such as sample inhomogeneity. In this case, it becomes difficult to resolve spectral intensity profiles of inhomogeneous fruit samples. It is for these reasons that spectroscopic techniques are coupled with ML techniques to aid in overcoming such challenges. Considering LRS, the technique facilitates quick analysis as the time taken for each cycle of measurement is less than one minute. However, this advantage is eroded in practical applications owing to the low reliability of data processing and hence, the need to validate the same measurements by multiple techniques. Nevertheless, with a high degree of accuracy and reliability, ML assisted LRS has the capacity to solve a wide range of complex issues such as the one in this research.

1.2 Statement of the Problem

The challenge with conventional techniques for assessment of artificial ripeners in fruits is that they are costly, time-consuming, require specialized sample preparation and more often involve destruction of the test sample. The laser Raman spectroscopy (LRS) technique overcomes most of these challenges. However, analysis of trace analyte concentrations in complex matrices such as fruit samples can be challenging. This is because, the traditional data analysis approach assigns known individual peaks to specific vibrational groups but the composition of the entire sample affects each individual peak due to matrix effects, thus this approach cannot adequately represent the resulting peak intensity shifts. In addition, the fluorescence effect tends to be more intense than the Raman effect. This implies that the laser Raman spectroscopy technique may not be sufficient independently. Multivariate machine learning techniques offer alternatives to this traditional approach by using data from the entire wave range collected to solve issues connected with univariate Raman data analysis. Therefore, the machine learning assisted laser Raman spectroscopy approach has potential for rapid, non-destructive and cost-effective assessment of artificial ripeners in fruits.

1.3 Research Objectives

1.3.1 Main Objective

The primary goal of this research was to develop a machine learning-assisted laser Raman spectroscopy technique for the direct and rapid assessment of calcium carbide ripened bananas.

1.3.2 Specific Objectives

i. To design and optimize a protocol for the assessment of calcium carbide ripened bananas for rapid laser Raman spectroscopic measurements.

ii. To pre-process the LRS measurements obtained from specific objective (i) for spectral noise reduction and perform exploratory analysis of the pre-processed data using PCA to assign molecular vibrations and for dimensionality reduction.

iii. To develop calibration models for quantitative analysis of the data obtained from specific objective (ii) above using selected machine learning techniques, namely ANN, SVM/R and RF.

iv. To test the applicability of machine learning-assisted laser Raman spectroscopy technique in assessing the presence of calcium carbide in market samples.

1.4 Justification and Significance

Fruits are a popular source of food as they are a vital source of nutrients for the well-being of humans. However, continued consumption of artificially ripened fruits has raised growing concern due the associated health risks. For instance, consumption of CaC_2 ripened fruits has been reported to cause peptic ulcers and colon cancer. To address this health concern, a rapid method for assessing ARF is needed. Hardly is there a rapid and non-invasive method for assessing ARF as the standard wet laboratory techniques are destructive and time-consuming and therefore inappropriate for such practical applications.

Applied vibrational spectroscopy techniques such as LRS are non-invasive and rapid and have potential to be applied in assessment of ARF. However, when this technique is applied in studies involving fruits, it suffers from the influence of broad fluorescence baseline that obscures the requisite Raman signal. It also becomes difficult to assign Raman bands to different chemicals of interest when the spectra is recorded in the background of interfering molecules. Further, the inhomogeneity of fruit samples results to Raman spectra with variations in intensity making quantification studies difficult. Therefore, the practical application of LRS in studies such as in the current work increases significantly when combined with appropriate data pre-processing techniques as well as ML techniques such as ANN, RF and SVM. The findings in this study highlight optimal conditions for recording Raman spectra of fruits ripened artificially or naturally. In addition, the use of appropriate data mining techniques and optimally tuned ML parameters for the fast and reliable analysis of LRS data are outlined.

CHAPTER 2

LITERATURE REVIEW

2.1 Chapter Overview

There is a discussion in this chapter of the literature on methods of assessing artificial ripeners in fruits. Various techniques such as chromatography, near-infrared spectroscopy and LRS are discussed as well as the role of ML in LRS.

2.2 Common Artificial Fruit Ripeners

The application of CaC_2 to accelerate the ripening process of fruits alongside the negative health effects of using this chemical has been documented over time in several studies (Kesse *et al.*, 2019; Lakade *et al.*, 2018; Chandel *et al.*, 2018; Kathirvelan *et al.*, 2017) as discussed in section 1.1.1.. Other common chemicals used in artificial ripening of fruits include Ethephon (2-chloroethylphosphonic acid) which decomposes into ethylene when hydrolyzed, (Roberts *et al.*, 1998). The liberated ethylene speeds up the ripening process in fruits. Impure ethephon may contain monochloroethyl ester which may degrade to monochloroacetic acid which can cause burn injury (Pirson *et al.*, 2003). Further, health studies report that ethephon gets rapidly absorbed in the gut and has potential to damage liver cells (Bhadoria *et al.*, 2018). The use of kerosene for ripening has also been on the rise. Kerosene is used in kerosene burners to generate smoke for ripening fruits. The generated smoke contains unsaturated compounds that are known to accelerate ripening (Maduwanthi *et al.*, 2019). Kerosene used for this purpose may have impurities like sulfur which is emitted as sulfur dioxide (SO_2) in the process. High exposure to (SO_2) can cause pulmonary injuries (Kan *et al.*, 2010).

2.3 Methods for Assessing Artificial Ripeners in Fruits

Several studies on artificial ripening of fruits have been conducted. However, there is minimal literature that directly highlights spectroscopic methods used in assessing ARF.

One such study was conducted by Nowshad *et al*. (2018) to analyze artificial ripening agents used for ripening bananas. From the elemental analysis of CaC_2 samples obtained from their locality using an EDXRF spectrometer, sulfur, arsenic and phosphorus were found among other metals. The arsenic content found (160 ppm) exceeded the limit set by United States Food and Drug administration (0.5-2 ppm). Phosphorus present in the CaC_2 samples (120 ppm) was much higher than the lethal dose set by Center for Diseases Control and Prevention (CDC) for dogs (10 ppm) and cats (4 ppm) (Francois *et al*., 2015). The toxic dose set by the National Academy of Sciences, USA for sulfur is at 600 ppm. The sulfur found in CaC_2 samples significantly exceeded this value. However the experimental techniques that were used to assess the ripening agents within the bananas were largely wet chemical techniques.

The detection and quantification of CaC_2 in carbide ripened fruits is complex owing to its instability in the presence of moisture. Further, the diffusion of calcium hydroxide and acetylene into the fruits make the detection even more challenging (Ramachandra *et al.*, 2016). In this regard, several studies propose that in order to detect CaC_2 in carbide ripened fruits, the respective tests should be targeted at detecting the common impurities found in CaC_2 like arsenic, sulfur and phosphine. In one such study Lakade *et al.* (2018) demonstrated that carbide-ripened mangoes contained arsenic as evidenced by color changes resulting from reactions between arsenic and lauryl sulphate-capped gold nanoparticles. In another study, Chandel *et al.* (2018) demonstrated that CaC_2 ripened mangoes contained arsenic residues (ranging from 35 to 107 ppb) whereas mangoes which were left to ripen naturally did not show any arsenic residues. Thus, the presence of the afore mentioned impurities in fruits can be used as indicators of ripening using CaC_2.

2.3.1 Wet Chemistry-based Techniques

Various forms of chromatography have been used in assessing artificial ripeners in fruits. These include ICP-AES, GC, HPLC, and HPLC coupled with mass spectrometry (HPLC-MS). These methods are more accurate and sensitive to trace concentrations but are not rapid (Naushad et al., 2014). In particular, chromatographic methods are more costly, destructive and involve complex sample preparation techniques that require high level of expertise. Further, toxic waste are produced during measurement due to the solvents used; this has the implication that the number of samples to be tested must be limited for environmental reasons (Naushad *et al*., 2014). For this reasons, chromatography does not provide a rapid assessment method for artificial ripeners in fruits.

2.3.2 Nuclear Magnetic Resonance (NMR) Spectroscopy

This technique has gained traction in the fruit industry as it has shown to be a radiation-free, non-invasive and fast approach of obtaining chemical structure information of biological samples such as fruits and vegetables. Wyzgoski *et al.* (2010) highlighted that fruits have complex chemical structures and their analysis using the classical 1-dimensional NMR spectra was constrained in a study to ascertain bioactive components of raspberry fruits. The use of partial least squares, a chemometric tool, made it possible to establish relationships between bioactive components of raspberry fruits to biological responses. The biomarkers identified for different fruits can be used as fingerprints to point out differences between chemically and naturally ripened fruits in the present study. Further, the use of chemometrics to deal with highly multidimensional data resulting from the heterogonous fruit samples was employed in the current study.

Lee *et al.* (2014) compared the performance of NMR spectroscopy with HPLC in a study to quantify Cordycepin, a medicinal compound, in mushrooms. The quantitative results showed that NMR was better than HPLC in terms of repeatability and sensitivity. However, preparation of the samples in this study involved preparing mushroom extracts using solvents like methanol. The present work envisages the use of a chemical free method to assess carbide ripened bananas.

In another study conducted by Wu *et al.* (2020), it was demonstrated that NMR spectroscopy could obtain high resolution metabolic profiles of several fruits. The resulting spectra could aid in pinpointing significant metabolites that are unique to particular fruits and this could aid in fruit quality control. It was reported that these studies could be carried on site using portable NMR equipment. This presents an advantage in the present work whereby bananas can be assessed, in a fast way, for artificial ripeners on site. Nonetheless, the method faces challenges in analysis where the fruit sample has non-uniform peel and pulp such that it becomes the metabolites cannot be easily discerned. Multivariate analysis approaches were thus recommended.

2.3.3 Near Infrared Spectroscopy

The challenges faced by chromatographic methods in assessing artificial ripeners in fruits can be addressed using vibrational spectroscopy techniques. These include NIR, MIR and laser Raman spectroscopy (LRS). They are relatively fast, non-destructive and require minimal sample preparation. Rapid assessment of artificial ripeners has been reported in tomato, apple, apricot, citrus, peach and olives using NIR spectroscopy (Gracia *et al*., 2011). As a relatively quick and low-cost approach, NIR spectroscopy is a robust alternative. However, the NIR spectra is generally characterized by broad overlapping peaks. The mixed spectra from layered samples are often computationally difficult to resolve and allocate to the individual components. Further, the extent of light penetration in the NIR reflectance mode is often limited by the relatively low radiation energy of the illumination sources (Qin *et al*., 2017). Thus NIR spectroscopy has not been used extensively for assessment of ARF.

2.3.4 Laser Raman Spectroscopy

LRS is a wide class of spectroscopic techniques based on inelastic scattering of light of a single wavelength. The technique is capable of analyzing molecular composition of a sample with high specificity owing to the unique nature of molecular vibrations (Gomez-Lazaro *et al*., 2017). However, LRS has been used to a lesser extent in assessment of ARF in comparison with infrared absorption primarily because of problems associated with sample degradation and fluorescence. Further, Raman scattering is a weaker process compared to NIR and MIR spectroscopy and it has low signal-to-noise ratio (SNR). In addition, interpretation of Raman spectral data of target samples within different matrices is complex (Morris, 2008). Spectral data that was obtained in this work had a large number of variables with no properly defined peaks. Consequently, the manual correlation of experimental spectra with the standard reference spectra for peak allocation was difficult (Kiyohara *et al*., 2018). Therefore, LRS used independently could not be used to assess artificial ripeners in fruits.

Standard data analysis of LRS data involves preprocessing by cosmic ray removal followed by smoothing and then baseline correction. Studies show that proper application of these steps is able to suppress the fluorescent signal which is normally intense than the Raman signal, especially for biological samples (Wei *et al.*, 2015). Additionally, in classification studies, principal component analysis (PCA) is normally used for reduction of data dimensionality before classification. However, proper choice of appropriate preprocessing steps without loss of vital information remains a challenge (Liu *et al.*, 2017). In this study, appropriate preprocessing techniques were used and the experiment optimized to obtain most useful information with minimal loss of data.

Calcium carbide, when used for ripening fruits chemically, must be hydrolyzed to produce acetylene gas and other by-products. Therefore, LRS could not be used to directly test for CaC_2 in carbide ripened bananas as it was absorbed in different form by the fruits. Thus, in order to detect and quantify how much CaC_2 was used to ripen sample bananas, this study relied on the following Raman peaks of molecules resulting from impurities found in CaC_2 (see Table 2.1).

Table 2.1: Common impurities found in CaC_2 and their Raman peaks.

Target molecule	Raman vibrational modes (cm^{-1})	Reference
Phosphine	2306, 1115, 979 (P-H bending)	(Ceppatelli *et al.*, 2020)
Sulfur S8	160, 225, 480 (S-S bond stretching)	(Nims *et al.*, 2019)
Calcium hydroxide	780 (O-H out-of-plane bending)	(Chiriu *et al.*, 2014)
Acetylene	612 (C-H asymmetric bending)	(Edwards, 1990)

2.4 Applications of Machine Learning in Analytical Spectroscopy

Section 2.3 above highlights some of the limitations encountered in analyzing spectral data. These limitations can be partially solved by applying multivariate analytical technique for spectral data analysis using machine learning. In the case of LRS, hardly are there references pointing out machine learning techniques employed in LRS data for the assessment of artificial ripeners in fruits. Liu *et al*. (2017) demonstrated a unified solution for identification of chemicals whereby a convolution neural network was trained to identify substances according to their Raman spectrum without data preprocessing. The network performed well with high classification accuracies (95%). However, this network was only tested for mineral datasets. Madden *et al.* (2003) investigated the use of ML techniques for predicting the concentration of cocaine in solid mixtures by examining the Raman spectra of the samples. Several multivariate calibration models were tested in this study. The results showed that the neural network model aided with feature selection produced greater prediction accuracy than chemometrics models such as partial least squares. In the present study, feature selection was vital for data reduction so as to reduce the dimensionality of the LRS data.

The NIR spectra for heterogeneous samples, as described in section 2.3 above, is normally characterized by overlapped broad peaks without distinct signature of individual components. Multivariate calibration techniques using ML have since been proposed to solve this limitation. In one such study, RF was evaluated as an alternative multivariate technique for analyzing NIR spectral data of gasoline (Lee *et al.,* 2013). In comparison to partial least squares, the quantitative analysis of varied gasoline samples using the RF model was more accurate without overfitting. ML models that are not properly tuned to the specific spectroscopic problem suffer the problem of poor generalization ability. In the present work, parameters in the RF model were selected and adjusted such that the developed model did not overfit or underfit.

Support vector machine (SVM) as non-linear calibration technique has also been proposed for solving complex problems in spectroscopy. Thissen *et al*. (2004) demonstrated the use of SVM to solve a well-known chemical problem in which NIR spectra was measured at different temperatures leading to non-linear spectral variations. In comparison to the previously applied modelling techniques employed to find a solution to this problem, SVM performed best. In the present study, highly dimensional Raman spectral data was recorded and it was expected that this would present computational complexity for some ML algorithms. SVM calibration models for pattern recognition and regression were therefore among the ML algorithms used for this work as the quality of the SVM is not directly dependent on the dimension of the input data (Liu *et al.*, 2017).

Machine learning was employed in this work to deal with the challenges associated with univariate analysis of spectral data. Among these include data from heterogeneous samples, which normally have varying intensity values for the same sample and highly dimensional data with many correlated variables. The application of machine learning models to build calibration models for concentration of molecules of interest amidst a field of interfering molecules was therefore crucial in this work.

CHAPTER 3

THEORETICAL FRAMEWORK

3.1 Chapter Overview

In this chapter, the principle underlying Raman Scattering is explained in detail. The basics of both classical and quantum mechanical formulations are presented, but not in a rigorous manner. Also included in this chapter, is the theory behind spectral data analysis, exploratory data analysis and multivariate calibration of spectral data using ML algorithms.

3.2 Raman Scattering

The Raman effect was first experimentally observed by C. V. Raman, an Indian physicist around 1928. For this discovery, he was awarded the 1930 Nobel Prize for Physics (Müller *et al.*, 2003). As a precursor to this discovery, a postulation had been put across by Smekal; that radiation scattered from molecules, comprises photons with the incident photon frequency as well as photons with a shift in frequency. In principle, the Raman effect is a manifestation of this.

It is the inelastic scattering of light by matter that is responsible for the Raman effect. There are three unique ways that a photon of light, that does not have enough energy to cause an electronic transition from one state to another, might be scattered when it collides with a molecule. Photons hitting a molecule can can be elastically scattered, meaning their energy is identical to their scattered counterpart's energy. Rayleigh scattering is another name for this phenomenon. As an alternative, the incoming photon can be inelastically scattered by either gaining energy from the molecule or losing to it. In the case where the photon loses energy to the molecule, the phenomenon is referred to as Stokes scattering. The molecule ends up having excess energy that shows up as vibrational energy in the form of a phonon (Schrader, 2008). Figure 3.1 below is a representation of this process.

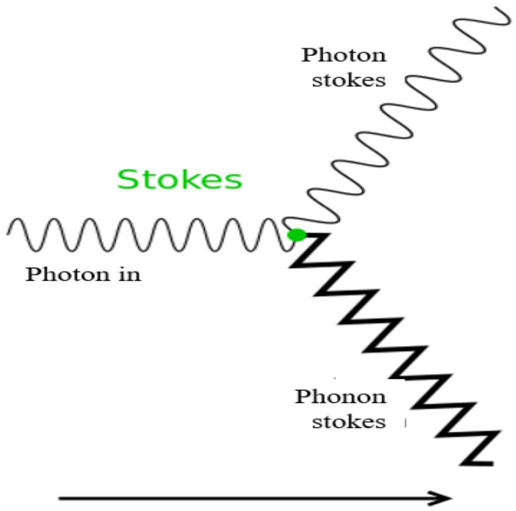

Figure 3.1: Feynman diagram of Stokes scattering. (Source: Feynman *et al.*, 2011). The process is characterized by energy loss in the photon due to phonon creation.

On the other hand, in the case where the photon gains energy from the molecule, the phenomenon is known as anti-Stokes scattering. Consequently, the molecule ends up losing vibrational energy in the form of a phonon to the photon. As indicated in Figure 3.2 below, the anti-Stokes process significantly relies on the phonon population. A strong correlation exists between this process and temperature since an increase in temperature results to increased phonon population.

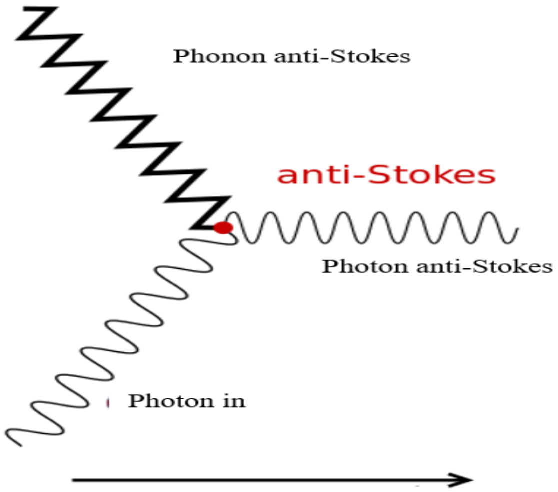

Figure 3.2: Feynman diagram of anti-Stokes scattering. (Source: Feynman *et al.*, 2011). The process is characterized by energy gain in the photon due to the absorption of the phonon.

3.2.1 Classical Theory of Raman Scattering

Raman spectroscopy, in principle, traces foundations to an elementary classical theory (Larkin, 2017). This theory brings forth the idea that molecules are simply atoms that undergo simple harmonic vibrations without quantization of vibrational energy. It is based on polarizability theory of molecules as proposed by G. Placzek in 1934. When a molecule is put into a static electric field, it suffers some amount of distortion whereby the positive charges on the nuclei are attracted toward the negative pole of the electric field and electrons are attracted towards the positive pole of the electric field. This separation of charge centers induces a dipole moment which is set up on the molecules. Because of this, the classical theory presumes that molecules are dipoles. For small fields the induced dipole moment μ is directly proportional to the applied electric field strength *E.* Thus:

$$\mu = \alpha E \tag{3.1}$$

Where α is the proportionality constant called the polarizability of the molecules, which is a measure of the ease with which the electron cloud may be distorted by the presence of an external electric field. For diatomic molecules aligned parallel to the direction of the electric field, polarizability is higher than when aligned vertically.

Where the electric field varies, there will be a varying dipole moment, of the same frequency, in response to that changing field. When a sample molecule is subjected to a radiation of frequencyV_o, the electric field experienced by the molecules varies according to the equation below (Larkin, 2017):

$$E = E_o \sin 2\pi v_o t \tag{3.2}$$

Where $E_\circ$ is the amplitude (intensity) of the vibrating electric field, $V_\circ$ is the frequency and t is the time. Substituting equation (3.2) into equation (3.1), we obtain:

$$\mu = \alpha E_o \sin 2\pi v_o t \tag{3.3}$$

Which is the expression for oscillating dipole which emits its own frequency known as Rayleigh scattering.

When placing any molecule in the electric field there will be a change in the dipole moment and because of the induced dipole moments, the molecule will vibrate slightly. The molecular vibrations causes a change in the polarizability given by:

$$\alpha = \alpha_o + \left(\frac{\partial \alpha}{\partial q}\right) q_o \sin 2\pi v_m t \tag{3.4}$$

Here, $\left(\frac{\partial \alpha}{\partial q}\right)$ is the rate of change of polarizability with vibration, q_o is the period of molecular vibration and v_m is the vibrational frequency. Substituting equation (3.4) into equation (3.3), we obtain:

$$\mu = \alpha_o E_o \sin 2\pi v_o t + \left(\frac{\partial \alpha}{\partial q}\right) E_o q_o \sin 2\pi v_o t \cdot \sin 2\pi v_m t \tag{3.5}$$

From the relation:

$$\text{sinAsinB} = \frac{1}{2}\left[\cos\left(A - B\right) - \cos\left(A + B\right)\right] \tag{3.6}$$

Equation (3.5) can be written as:

$$\mu = \alpha_o E_o \sin 2\pi v_o t + \frac{1}{2}\left(\frac{\partial \alpha}{\partial q}\right) E_o q_o \cos 2\pi (v_o - v_m) t + \frac{1}{2}\left(\frac{\partial \alpha}{\partial q}\right) E_o q_o \cos 2\pi (v_o + v_m) t \tag{3.7}$$

If we consider $\left(\frac{\partial \alpha}{\partial q}\right) = 0$, then the last two terms of equation (3.7) disappear and we remain with Rayleigh scattering. But in actual case, $\left(\frac{\partial \alpha}{\partial q}\right) \neq 0$ for a Raman active molecule; there should be a change in polarizability with respect to the vibration. Thus, the vibrating molecule can be a source of scattered radiation of three different frequencies. First, it can be of frequency $\nu_\circ$ which remains without any change in relation to the incident radiation. This type of scattering is also known as Rayleigh scattering. Secondly the frequency can be $(v_o - v_m)$ which is equivalent to the difference of the frequency of the incident radiation and that of the vibrations of the molecule, known as Stokes scattering. Lastly, it can be of frequency $(v_o - v_m)$ which represents the sum of the frequencies of the incident radiation and the vibration of the molecule, known as anti-Stokes scattering.

3.2.2 Quantum Theory of Raman Scattering

Electromagnetic radiation is dual in nature i.e. it is both particulate and wave in nature. The Raman effect as previously interpreted was on the basis of wave theory, which was a classical approach to electromagnetic radiation. Quantum mechanics, on the other hand, acknowledges that a molecule's vibrational energy is quantized, revealing its particulate form (Larkin, 2017). In the case of electromagnetic radiation they are also known as photons. The relationship between the energy (E) of a photon and its frequency (ν) is described by the Planck formula:

$$E = hv \tag{3.8}$$

In the case of a photon's collision with a molecule, three events can occur: absorption, emission and scattering. In the case of LRS, scattering is the phenomenon of interest. Scattering takes place within a relatively short time (10^{-14}s) of when a photon, of energy not equal to the energy difference between any two stationary levels of the molecule, interacts with that molecule. In particular, the typical Raman effect occurs whenever a photon interacts with a molecule at a far lower energy level than the energy difference between the ground state and the first excited state.

When a photon of frequency $\nu_\circ$ is incident onto a molecule, two types of collision are possible: elastic and inelastic. If the collision is elastic, the frequency of the incident photon will be equal to the frequency of the scattering. On the other hand if the collision is inelastic, the resulting scattered frequency will either be higher or lower than that of the incident photon. The assumption here is that total kinetic energy of the photon and of the molecule remains unchanged before and after the collision (Meier, 2003). From the law of conservation of energy,

$$hv_o + E_o = hv + E \tag{3.9}$$

Where $h\nu_\circ$ and $E_\circ$ are energy of the photon and molecule respectively before collision whereas $h\nu$ and E are energy of the photon and molecule respectively after collision. On rearranging equation (3.9) above, we obtain:

$$\frac{E - E_o}{h} = v_o - v \tag{3.10}$$

From equation (3.10), three cases are possible when identical molecules are illuminated with monochromatic light as illustrated in the Figure 3.3 below. The first case is when the photon with an initial energy of $h\nu_\circ$ proceeds without a change in energy, this is termed Rayleigh scattering. Secondly, the photon can experience a decrease in energy, a phenomenon known as Stokes scattering. Lastly, the photon can experience an increase in energy, this is referred to as anti-Stokes scattering. Raman shift occurs when there is a change between the incident and the scattered frequency (Morris, 2008).

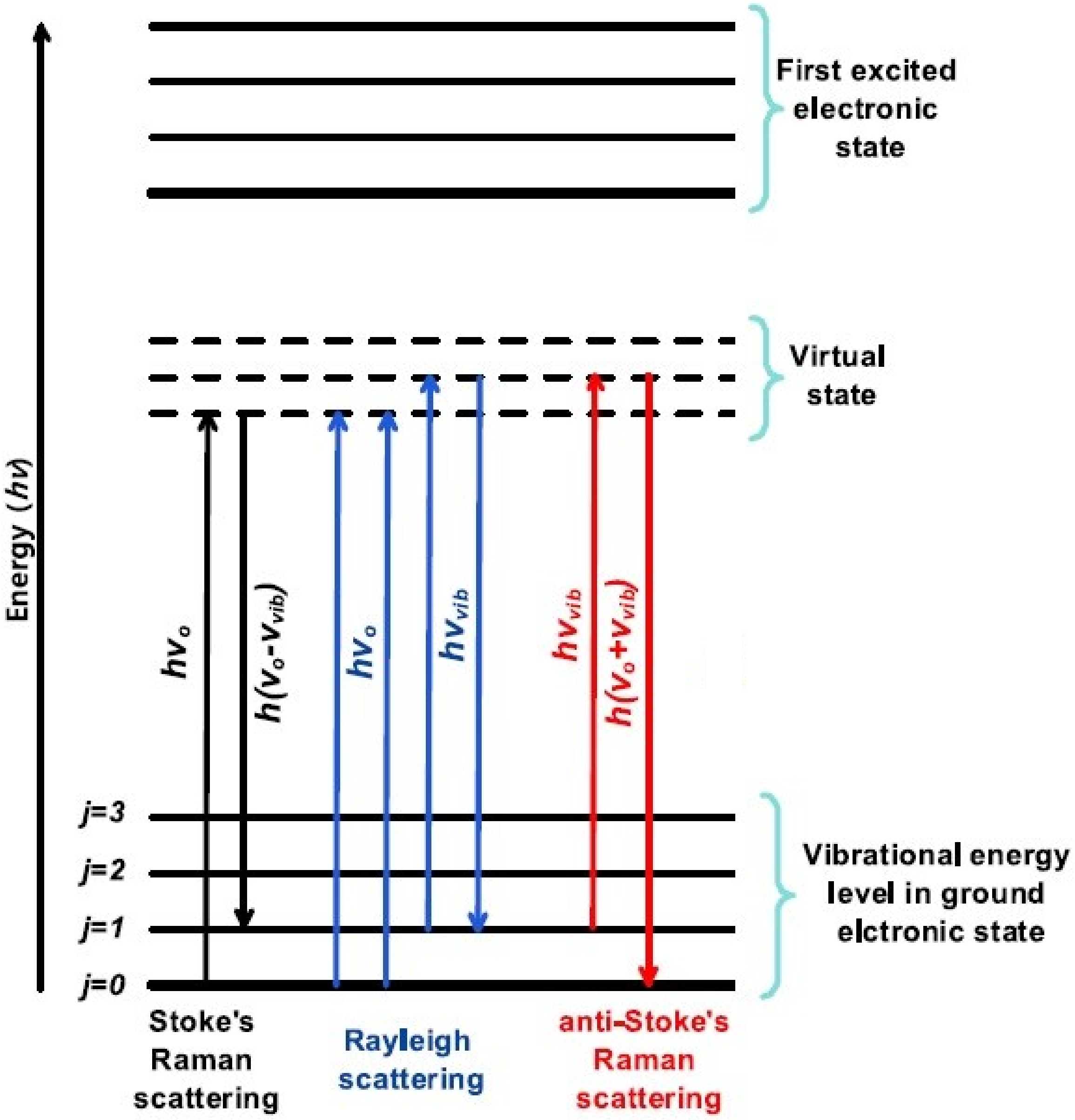

Figure 3.3 Diagram of scattering during illumination with monochromatic light. (Source: Demtröder, 2008)

In Figure 3.3 above, two electronic levels are shown. Ground state level and the first excited state. At the ground electronic state, the vibrational levels are also illustrated. The dashed lines depict virtual levels of the molecule, separated by $h\nu_v$ where ν_v is one of the possible vibrations of the molecule. This follows from the fact that most molecules generally have more than one Raman active vibrational modes. In accordance with

Boltzmann's rule of distribution, the vast majority of molecules will be in the vibrational ground state at room temperature. For this reason, Stokes transitions are more likely to occur than anti-Stokes transitions (Müller *et al.*, 2003). In most of the practical applications of LRS, the Raman scattering is presented only as the Stokes spectrum and is given as a shift in energy form the energy of the incident laser beam. In this work, the Raman system that was employed uses the Stokes configuration.

In the case of Rayleigh scattering, the scattered frequency has a value equivalent to that of the incoming light and is by far stronger than that of the Raman frequencies (Stokes and anti-Stokes). This scattering process is the most intense and the most probable amongst the scattering processes. Consequently, it is necessary to filter away the Rayleigh signal efficiently in order to prevent the Raman signal from being swamped. Also luminescence (fluorescence) signals can easily swamp the Raman signal (Afseth *et al.*, 2005). Figure 3.4 below shows a schematic of the interaction between the incident beam (drawn in thick yellow line) and the sample.

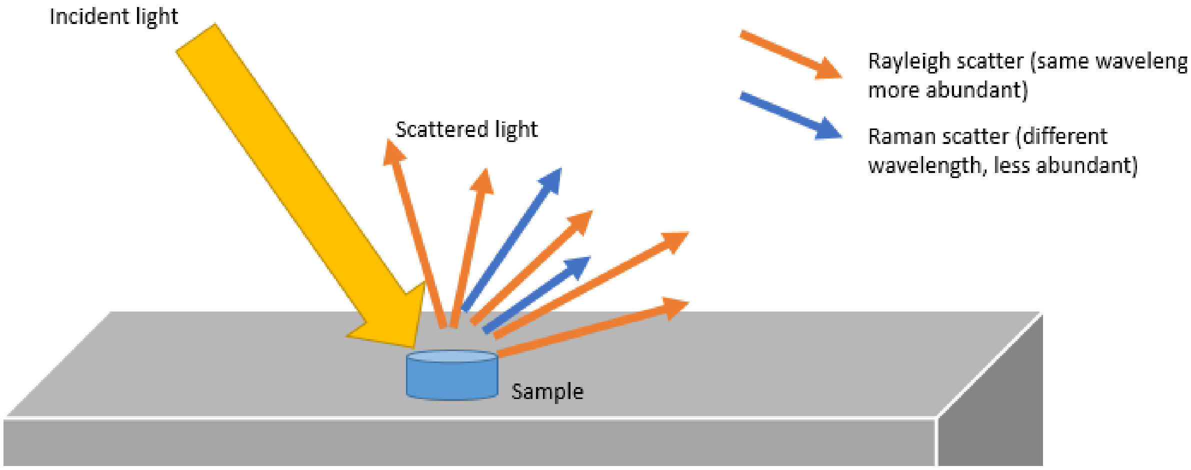

Figure 3.4: A schematic showing Rayleigh and Raman scattering. (Adapted from: Demtröder, 2008). Orange colored rays (majority) represent the intense Rayleigh scattering whereas the weaker Raman scattering is represented by the few blue rays.

Rayleigh scattering, being the dominant process, is represented by multiple rays (drawn in orange lines) whereas Raman scattering is displayed as the weakest process (few rays drawn in blue lines). This is because most photons scatter elastically without energy changes.

The absolute differences between the frequencies of the incident photon and both scattered photons are the same as the molecular vibration frequency.

$$hv_o - hv_{(stokes)} = hv_v = hv_{(anti-stokes)} - hv_o \tag{3.11}$$

From equation (3.11) above, we conclude that the difference in frequency between the incident photon and the scattered photon is characteristic of a molecule and independent of the frequency of the incident radiation. Thus, even at different excitation wavelengths for the same molecule, the same Raman spectra are expected.

It is imperative to highlight the fact that Raman scattering is governed by selection rules which are in turn determined by the symmetry and electronic structure of the molecular system under study. Consequently, such considerations are an important factor in Raman studies of materials as will be discussed in subsequent sections.

3.3 Intensity of Raman scattered light

Several factors affect the intensity of Raman scattered light as shown by equation (3.12).

$$I = K(v)A(v)v^4 I_o J(v)C \tag{3.12}$$

Where *I* is Raman scattered intensity, *K* is the spectrometer's spectral response, *A* is the absorption of the medium, *v* is frequency of the exciting laser, I_o is the excitation intensity, *J* is the molar scattering coefficient and *C* is the concentration of a given sample (Nakamoto, 2006). The intensity of the Raman scattered light is directly proportional to these factors.

A Raman spectrum is a depiction of intensity as a function of the wave shift i.e. the difference between the excitation frequency and the Raman scattered radiation frequency. In the classical sense of Raman scattering as discussed in sections 3.2.1, Raman scattered intensity depends on the polarizability of the molecules, the concentration of these molecules in the sample and the excitation source (Schrader, 2008). Thus, it was expected that the Raman intensity profiles for various analyte concentrations were to vary linearly in the current study.

3.4 Molecular vibrations and the Raman spectra

As discussed in section 3.2, a group of atoms interconnected by elastic bonds make up molecules which can perform periodic motions. Consequently, the energy of a molecule can be divided into a number of different parts of freedom on condition that there is no change in electronic energy. Three of these are translational whereas the other three are rotational. Therefore, polyatomic molecules having n atoms possess 3n-6 normal vibrations with the exception of linear molecules which possess 3n-5 vibrations (Demtröder, 2008). It is from these molecular vibrations that the vibrational spectra of molecules, such as Raman spectra, are defined. These vibrational spectra are highly dependent on their atomic mass, their geometrical orientation and nature of their chemical bonds among other factors.

Raman active vibrations mainly result from symmetric vibrations in molecules, in particular, vibrations that can cause a shift in the polarizability of the electron cloud around molecules (Dieing *et al.*, 2011). These vibrations are unique and thus can be used as fingerprints of molecules. In other words, the spectra shows specific vibration bands that can only be associated with a particular set of molecules. These spectra are more often than not defined by definite ranges of frequency (energy), intensity (polarizability) and shape of the bands (bond environment) (Demtröder, 2008).

Molecular structure can be derived from vibrational spectra using two approaches. Group theory coupled with mathematical model calculations is one of the approaches. The second approach is by use of empirical characteristic frequencies for chemical functional groups (Larkin, 2017). These two approaches form the basis for interpretation of vibrational spectra. Notably, a vast number of the empirical functional groups approach have been confirmed and refined by using the group theory approach. Nonetheless, many identification problems, such as the one presented in the current study, employ the use of the empirical approach to solve them. Functional groups exhibit specific vibrations which only the atoms present in that particular group are dislocated. These characteristic group vibrations frequencies remain relatively unchanged regardless of whatever molecule the group is in (Schrader, 2008). This is to say that molecules of interest can largely be identified in a Raman spectra regardless of the matrix they are immersed in. Further, intensities of the bands in the spectra of a mixture are customarily proportional to the concentration of the individual components (Dieing *et al.*, 2011). These characteristics of vibrational spectra, make LSR an invaluable technique in the current study.

In this work, the assignment and interpretation of Raman peaks of target molecules was based on the type of bonds involved and the different positions each bond occupied on the spectral line. The band regions spanning the functional groups of the target molecules were then evaluated and where possible, Raman shifts of a particular peak was able to give appropriate information regarding the exact molecule in the respective functional group region.

3.5 Spectral Data Preprocessing

The Raman spectra is capable of quantitatively reflecting the composition of a sample. However, noise resulting from instrumental components, surrounding light and software computations among other sources are normally added in varying proportions to the Raman spectra. Further, the physical and chemical inhomogeneity of biological samples makes them to be complex samples as the desirable quantitative information from their Raman signal becomes obscured. Therefore, there is need to apply data correction methods to the Raman spectra to eliminate the unwanted signal (noise) while at the same time enhancing the requisite signal such as the subtle differences between samples.

Spectral data preprocessing normally employs the use of mathematical data correction methods prior to univariate or multivariate analysis. Thus, it is a crucial step in LRS studies where accurate, verifiable and robust quantitative information is required. Several approaches are normally used on LRS data depending on the type of study (Morris, 2008). Some of the most common preprocessing techniques applied to LRS data are discussed in this section. Also included, are the anticipated challenges in the current work which can be resolved by preprocessing.

The LRS system intended to be used in this study has a CCD incorporated for recording the scattered spectra. As the experiment is expected to be carried out across a number of days, the data point spacing recorded by the CCD may vary from day to day due to the different calibration settings. These data points may even drift over the course of a day owing to changes in humidity and temperature. The use of different gratings can also contribute to this challenge. These conditions lead to recordings with differences in the X-axes, a challenge commonly referred to as spectral axis drift (Byrne *et al.*, 2016). Spectral axis alignment is therefore a necessary procedure prior to further analysis as most analysis techniques require common spectral axes to perform meaningful analyses.

Fluorescence, as has been discussed earlier, is orders of magnitude stronger than the Raman signal. The fluorescent signal is characterized by a broad band signal and induces uneven amplitude shifts across the Raman signal for biological samples (Afseth *et al.*, 2006). Some hardware changes like the use of longer excitation wavelength lasers can tackle this limitation. In the current study, a 785 nm laser is expected to be used. Further, baseline correction methods, particularly modified polynomial fitting, will be employed. This involves choosing base points in the spectrum to fit a polynomial to the spectrum baseline and finally subtracting the polynomial fit from the original Raman spectrum.

The Raman signal in certain instances consists of high frequency components which customarily have much lower FWHM compared with genuine Raman bands which need to be removed through a process called smoothing. The use of Savitzky-Golay algorithm is one of the most common smoothing techniques applied to LRS data. It involves the use of a moving window based local polynomial fitting procedure to get rid of the noise in the Raman spectra (Byrne *et al.*, 2016).

Normalization is a crucial preprocessing step that takes care of disparities in intensity levels. This is done by ensuring that for the same sample under same experimental environment, the intensity of a given Raman band is similar as possible across all spectra. The variations in intensity normally arise due to fluctuations in laser power as well changes in sample opacity as the experiment is carried out (Gautam *et al.*, 2015). Correction of these variations is done by normalizing the spectra using normalization approaches like standard normal variate (SNV) and vector normalization.

Spectra which significantly differ from the group are considered as outliers and need to be omitted before further analysis is carried out. The use of thresholding methods in the compressed domain based on the variance of the data in these domains can be used to eliminate such outliers. For instance, when using PCA, the axes which explain most of the variance are used to judge the outliers (Shaver, 2001). Spectra that are also captured outside the region of interest are also considered as outliers.

The use of the data preprocessing techniques must be applied in correct sequence and proper consideration of adjustable parameters must be made where necessary. An optimal combination of steps to be followed is necessary for each specific study so as to realize an increase in the SNR.

3.6 Machine Learning Techniques in Raman Spectroscopy

Machine learning is capable of handling complex data sets such as the laser Raman spectral data that was obtained in this work. ML techniques were used to overcome problems associated with classical/univariate analysis of Raman data. Supervised and unsupervised methods were used for performing exploratory and multivariate calibration of acquired laser Raman spectra.

The LRS measurements basically consist of two parts, i.e. the requisite signal and the rest of the signal which is considered to be noise. The requisite signal represents the underlying chemical information which greatly influences the property of interest. One of the principal roles of multivariate analysis is to filter out the noise from the requisite signal by using statistical measures such as covariance of variables within a data set. It is important to establish which variables have a great impact on the requisite signal. Broadly, the main objectives of multivariate data analysis are data exploration, classification and regression (Varmuza and Filzmoser, 2016).

Adoption of ML techniques in a multivariate manner in LRS serves to overcome some of the limitations encountered during classical/univariate analysis of samples. For instance, the heterogeneous nature of biological samples makes their spectra complex due to the overlapping of several intense Raman bands among other issues. Thus the qualitative and quantitative analysis of spectral profiles associated with these kind of spectra requires the use of multivariate analytical techniques. These methods allow for complicated and large data sets to be analyzed by reducing the dimensionality such that the requisite information can be extracted (Byrne *et al.*, 2016). In the current study, multivariate methods were used to analyze multiple spectra simultaneously and make comparisons between groups of spectra to identify trends of spectral markers in control and non-control samples.

3.6.1 Exploratory Analysis of Raman Spectra Using the Principal Component Analysis Technique

When it comes to exploratory data analysis, an unsupervised method (one that does not require training) is always preferred. An example of a commonly used unsupervised technique is the PCA. It does analysis of data patterns in such a manner that highlights their similarities and contrasts.

Multivariate analysis often starts out with data involving a significant number of correlated variables. Consequently, analysis of such kind of data becomes a challenge owing to the large number of variables. For this reason, data reduction techniques are proposed for such cases. PCA is a dimension reduction tool that can be used to reduce a large set of correlated variables to a smaller set of linearly independent variables that contain as much information as the original data set. The exploratory properties of PCA will be used in this work mainly for visualizing the data and transforming the highly multi-dimensional variables to a smaller set of latent variables that will be used in the supervised ML models.

PCA works by decomposition of a correlation or covariance matrix into eigenvalue and corresponding eigenvectors which are orthogonal to each other (Shaver, 2001). In the process, the axes of original variables are rotated to a new coordinate system having principal axes (components). The axis or direction with maximum variation of the projected values of the original data points defines the first principal component (PC). The corresponding projected values are referred to as the scores whereas the coefficients of the PCs are known as the loadings. Each of the successive PC will have maximum variation of the projected points and will be orthogonal to its predecessor. Mathematically, PCA decomposes a data matrix X into an outer product of scores matrix T and loading matrix P' plus a residual matrix E, as expressed in equation (3.13) below (Varmuza and Filzmoser, 2016):

$$X = T.P' + E \tag{3.13}$$

The graphical interface of PCA provides valuable information that helps in visualizing the data so as to get a general understanding of the data. For instance, the score plots show the covariance between samples. A different approach is to look at the loadings plots to see how the original factors affect each of the PCs. As a result, they demonstrate how a certain PC responds to changes in a variable over a period of time. The loadings in conjunction with the scores can therefore be used to give information leading to the process of identifying molecular bands that are unique to the constituents of the sample being studied. This feature will aid in identifying molecular signatures of analytes of interest from Raman spectra of fruits ripened naturally and artificially.

3.6.2 Modeling Approaches for Multivariate Data Sets Using Supervised Machine Learning Methods

The major aim of multivariate calibration of data is to define the relation that exists between a response (Y) variable and several input variables (X). ML models utilize both linear and nonlinear functions in a multivariate manner to establish the relationship between X and Y (Varmuza and Filzmoser, 2016). In this work, the X variable, representing the spectral data, was a matrix *X (n* x *p)* having *n* as the measured spectra and *p* as intensities and other spectral features. On the other hand, the *Y* variable was a matrix *Y (n* x *m)* with *m* being the concentration of the analyte and *n* being the samples.

The following are three ML techniques that were used to develop models for classification and regression of the Raman spectral data in this work.

3.6.3 The Artificial Neural Network Model

Artificial neural network (ANN) is a type of machine learning technique that emulates the working of the human brain and the nervous system to process information and solve problems. ANNs have three basic parts each having neurons i.e. the input layer, hidden layer and the output layer. The artificial neuron is the foundation of every ANN and it can be viewed as mathematical model which performs the functions of multiplication, summation and activation (Ciaburro *et al*., 2017). Several neurons in an ANN work in parallel to solve problems. Figure 3.5 below shows a side by side comparison of the biological and the artificial neuron.

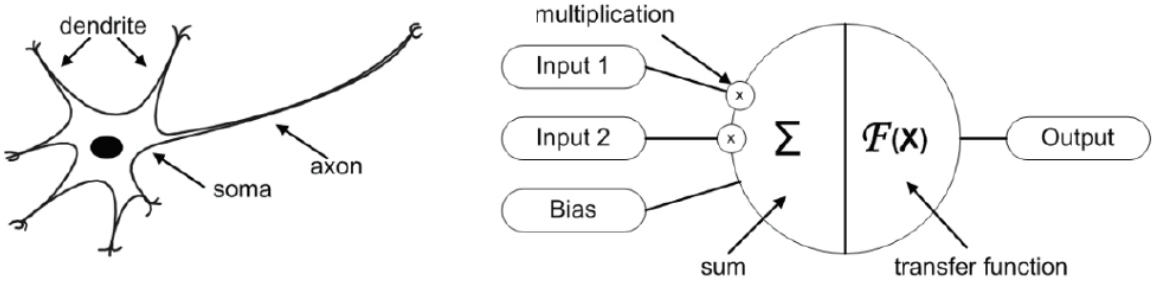

Figure 3.5: Biological and artificial neural design. (Source: Andrej Krenker *et al.*, 2011). The figure to the left shows the main parts of the biological neuron whereas to the right, an artificial neuron is shown.

In a biological neuron, inputs are first received into the neuron through the dendrites then processed in the soma before being finally relayed via the axon. The artificial neuron is fed with weighted inputs at its entrance. These weighted inputs as well as the bias terms are then summed up by a summation function in the middle section before being passed through an activation (transfer) function at the exit of the neuron. Given an output y_i, input x_i, weight w_i, bias b and a transfer function, the artificial neuron model can be summarized by the mathematical equation below (Gershenson, 2003):

$$y_i = f\left(\sum_i w_i \cdot x_i + b\right) \tag{3.14}$$

The only unknown variable in this artificial neuron model equation is the transfer function which is usually chosen on the basis of the problem at hand. In the current study, this was advantageous as the model could be tuned to handle data which was either linear or non-linear in nature.

When several individual artificial neurons are interconnected, they form artificial neural networks (ANNs). There are two ways in which the ANNs topology are built based on the direction of flow of information: feedforward and feedbackward. The network topology of BP-ANNs is such that the artificial neurons are organized layers and transmit their signals forward whereas the network errors are relayed backwards. Figure 3.6 below is a general depiction of the BP-ANNs.

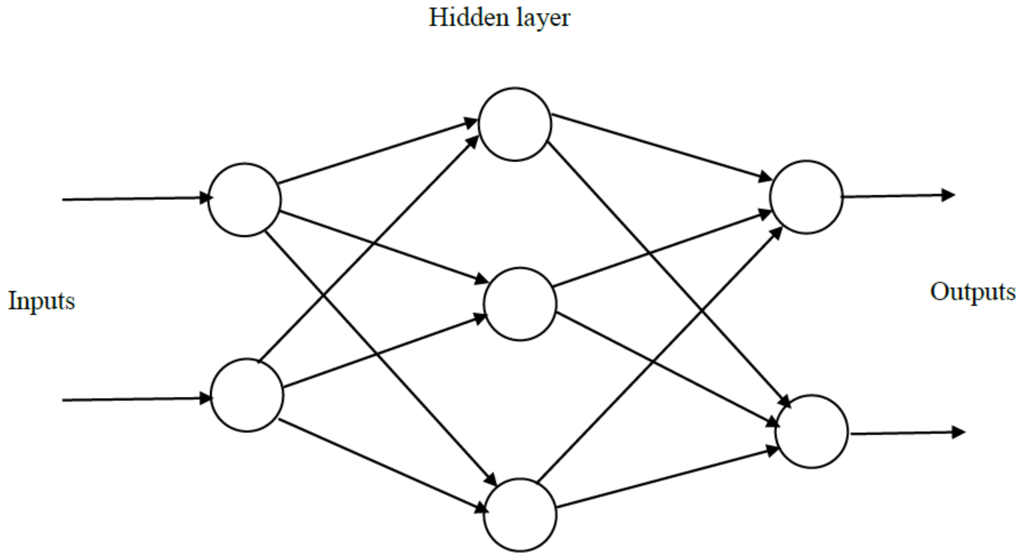

Figure 3.6: A generalized schematic of the BP-ANN. (Source: Ilonen *et al*., 2003)

The input layer is the independent variables used to predict the output layer (response variable). For regression problems, only one neuron in the output layer is normally expected whereas for classification problems, the output layer will have as many neurons as the distinct classes which are expected. The hidden layer transforms the input variables into higher order functions and by having more than one hidden layer results in achieving non-linearity (Krenker *et al.*, 2011). In order for ANNs to make sense from complex, non-linear data such as Raman spectra from fruits in this study, more than one hidden layer is necessary. Consequently, this explains why in deep learning neural networks, the number of hidden layers is quite substantial.

Just as the human brain uses experiences to give responses to new environmental inputs, the BP-ANNs have to learn by supervised learning approaches. The training process entails providing the network with inputs and expected outputs that the network should compute.

Initially, the network is assigned random weights. The cost (loss) function then calculates the total error of the network, which is simply the difference between the network's actual and the expected result. The main goal of BP-ANNs is to reduce this error to an acceptable level that is close to a predetermined threshold by adjusting the weights (Ciaburro *et al.*, 2017). When this happens, the network is said to have learned the training data and can therefore be used to generalize to new data. During training, the total network error is broken down and distributed back to each weight, hence the reason as to why the network is referred to as backpropagation. The network updates the weight for each layer until the lowest network error is achieved. This whole process from the forward pass to the backpropagation makes up an epoch and it is an iterative process.

The performance of the neural network can be assessed by several criteria including the coefficient of correlation (R), root mean square error and mean absolute error among others. A well trained model should result in an R value close to one and have very small values of error terms (Krenker *et al.*, 2011).

3.6.4 The Random Forest Model

Random forest (RF) is a supervised ML algorithm that can simply be described as a bunch of decision trees bundled together. A proper understanding of decision trees is required to implement the RF algorithm. A decision tree is basically a step by step process that uses certain criterion and thresholds to classify or predict the output values of a variable of interest. In classification and regression problems, decision trees have several advantages. Their informative output and visualization makes them easy to interpret. Further, they are less sensitive to outliers as compared with some traditional regression techniques in addition to having the ability to analyze highly dimensional data (Ayyadevara, 2018).

A decision tree, just like the biological tree, has roots, branches and leaves. In training the decision tree, the root decision node in the tree is first determined. The root node represents the whole or part of the input samples. In the case of Raman spectra, the input can be the wavelengths or intensities. The root node splits into two more sub-nodes through a decision process governed by some rules. Further sub-division among the daughter nodes proceeds in an iterative process until the final node in a decision tree is reached; this is called the leaf or terminal node. Therefore, a branch or sub-tree in this case is a sub-section of the entire tree.

In order to use decision trees, we begin from the top and make decisions in series until we reach the bottom where an outcome is realized. At each decision making step, A decision is made between only two viable options. The criterion for splitting at the decision nodes depends on the nature of the variables we are predicting. In particular, we need to establish the variable which will form the basis of the first split in the root node. A variable which has the greatest potential to separate different classes as much as possible is required. For this purpose, two important measures are used to check for the quality of the split; "entropy" and "Gini impurity" (Hartshorn, 2016).

Entropy is used for information gain and is a measure of uncertainty after splitting a node. Given input features i, the mathematical equation for entropy is:

$$entropy = \sum_{i} (-P_i log_2 P_i) \tag{3.15}$$

Where P_i is the probability of picking a data point within class i. Considering that we expect maximum uncertainty at the root node, the choice of a good split should be such that the variable chosen decrease uncertainty the most. Thus, the lower the number of entropy, the better. Once the decision for the first split has been made, the next decision is to decide on which side of the split the distinct variable will go to. The Gini impurity metric, which refers to the extent of inequality within a node, is used to evaluate quantitatively how good a split is. It is given by the formula (Hartshorn, 2016):

$$Gini = 1 - \sum_{i} {P_i}^2 \tag{3.16}$$

Where P_i is the probability of having data points within class i. The best possible value we could have is an impurity of 0 and this occurs when a class has all values belonging to one class or the other. The splitting process then continues until all the leaf nodes of a tree achieve their purest form possible. However, this process becomes disadvantageous as it suffers overfitting of the data and does not generalize well to new data. Nevertheless, this limitation faced by a single decision tree can be solved by having several independent decision trees.

Random forest overcomes the problems of overfitting by fitting multiple classification and regression trees to a data set and averaging the results. The randomness in the forest is achieved by bagging, which is short of bootstrap aggregating. Training the RF begins by taking a subset of the original data and building a decision tree based on this subset. The grown tree is then used to predict the out-of-bag (OOB) data, which is the subset of training samples that had not been selected during the growing of the first tree. OOB estimation is a significant cross validation method in RF regression (Zhang *et al.*, 2014). It ensures that the final result from the model is drawn from the majority vote of the independent trees. Once the first tree is fully grown and its OOB error estimate calculated, these steps are repeated *n* times to have *n* trees. The final prediction of the result is the weighted average from all the *n* trees. As a result of fitting many independent trees by the bagging procedure, the risks of biased decisions and overfitting are greatly reduced.

Random forest cannot be visualized as a single decision tree as it is a combination of several decision trees. Nonetheless, RF has a very important feature that allows to evaluate the variable importance of the input variables (Lee *et al.*, 2013). This is achieved by evaluating the effect of the variable on the Gini impurity and entropy. Consequently, this was important in this study as we were interested to know the distinguishing features from the Raman spectra of naturally and artificially ripened bananas.

3.6.5 The Support Vector Machine and Support Vector Regression Model

Support vector machines (SVMs) are a set of supervised ML methods that were developed mainly for binary classification problems but have been extended to regression problems (Liu *et al.*, 2017). Both SVM and support vector regression (SVR) have the same architecture with slight differences in the inputs for the models; SVMs have categorical variables as inputs whereas SVR has continuous variables as their input. In the years leading up to1980, most of the ML methods were based on linear decision surfaces. Decision trees and neural networks were developed in subsequent years to allow for efficient learning of non-linear decision surfaces. However, these methods suffered from local minimum problems. It is against this background that Vapnik et al., (1995) suggested a way to create non-linear classifiers by applying the kernel trick to maximum-margin hyperplanes.

The basic operating idea of SVM, just like 1 layer or multi-layer NNs is to find an optimal hyperplane for linearly separable patterns. Nonetheless, extensions to patterns that are not linearly separable has been achieved using the kernel function, where the original data is mapped (transformed) into a new dimensional space (Ma and Guo, 2014). A hyperplane is defined depending on the dimension of data where each data point is viewed as a p-dimensional vector in (p-1) dimensional space. The goal is to find an optimal hyperplane in (p-1) subspace which can separate these data points with the largest margin or separation possible. Consequently, the larger the margin the lower the generalization error of the classifier. A sample hyperplane for a 2 dimensional space is presented in Figure 3.7 below.

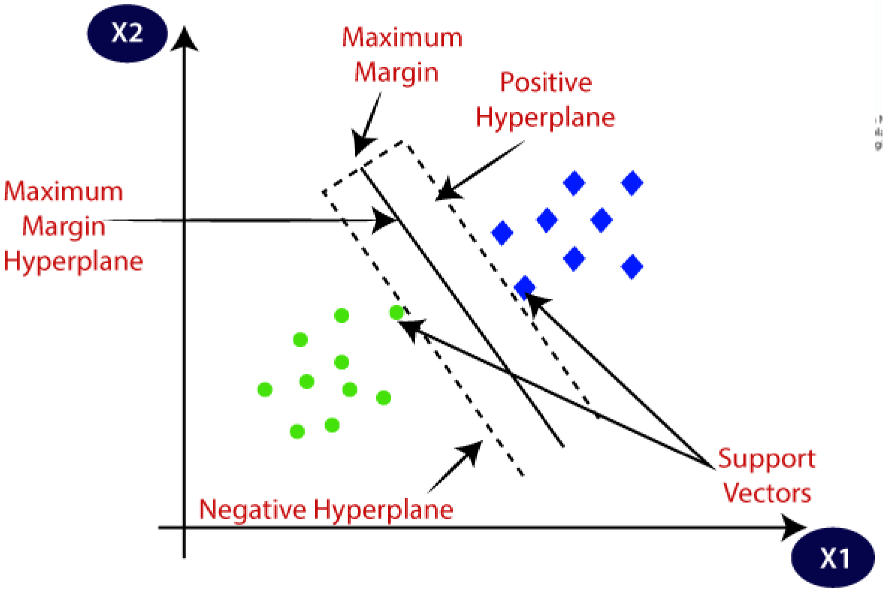

Figure 3.7: A sample hyperplane in a 2 dimensional space. (Source: Ma and Guo, 2014). The best hyperplane is the one that maximizes the margin between the support vectors.

The hyperplane is found using support vectors and margins. The support vectors are the data points lying closest to the decision surface (hyperplane) and usually form the set of data points most difficult to classify. They directly influence the location and of the optimal decision surface. During the training phase, a SVM or SVR algorithm builds a model of support vectors in space such that support vectors of different categories are separated by maximal margin. In the predictive phase, new data points introduced to the algorithm are mapped into the same space and predicted to a class based on which side of the hyperplane they fall on. In other words, SVM maximizes the margin around the separating hyperplane.

SVM is based on the use of the linear discriminant function:

$$f(x) = w^T x + b \tag{3.17}$$

This function represents the hyperplane in feature space where $x = \{(x_i,\ y_i)\}$ is a set of training pair (input, output) sample with features $x_1, x_2, x_3, \dots . x_n$ and the output result y. w (or w_i) is a set of weights, one for each feature, whose linear combination predicts the value of y. the last term b is the bias. The number of hyperplanes that can separate patterns such that the patterns form two classes lying on opposite sides of the decision tree are infinite. Thus, the linear discriminant function with maximum margin is the optimal solution to equation (3.17). The maximization of the margin around the separating hyperplane is a constrained optimization problem that can be solved using standard methods like the Lagrangian multiplier method (Amarappa *et al.*, 2014).

In a two class classification problem, the equation of the separating hyperplane or line is given by:

$$w^T x + b = 0 \tag{3.18}$$

Therefore, points in the negative class will satisfy the equation:

$$w^T x + b \leq -1, when\ y_i = -1 \tag{3.19}$$

Conversely, points in the positive class will satisfy the equation:

$$w^T x + b \geq -1, when\ y_i = +1 \tag{3.20}$$

The hyperplane parameters w and b are optimized in the SVM/SVR algorithm during training such that the support vectors are equidistant to the hyperplane on either sides. The optimization algorithm to generate the weights proceeds in such a manner that only support vectors determine the weights and thus, the boundary (Balabin and Lomakina, 2011). This was important in this study since there was an expectation that the Raman spectral features for the two classes of samples would have subtle spectral differences. The samples would therefore be classified based on these subtle differences arising from the different molecular composition in them.

In the case where support vectors are not linearly separable, kernel functions are employed to transform (map) the data to a higher dimensional space. A common example of non-linear kernel function is the radial basis function (RBF) of the form:

$$K(x, x_i) = exp\left\{-\frac{\|x - x_i\|^2}{2\sigma^2}\right\} \tag{3.21}$$

Where x_i is the input vector and σ is the radial width.

CHAPTER 4

MATERIALS AND METHODS

4.1 Chapter Overview

The application of LSR for classification and quantification of additives arising from artificial ripening of bananas using calcium carbide has been investigated using multivariate classification and regression models developed from the molecules vibrational bands. The acquisition of spectra was done over an exposure time of 10 S per sample and Raman shift centered at 1050 cm^{-1} to cover the fingerprint spectral region of sulfur, acetylene, calcium hydroxide and phosphine compounds resulting from hydrolyzed calcium carbide which is between 200 and 1200 cm^{-1}. The procedures used in preparation of control and treated samples and the methodologies for developing, customizing and fine-tuning the ML models are presented and discussed.

4.2 Instrumentation for the Laser Raman Spectroscopy Set-up

This study used a confocal laser Raman Spectrometer (STR Raman Spectrum, Seki Technotron Corp, Japan) equipped with an imaging spectrograph and a 785 nm exciting laser. A backscattered-illuminated CCD camera was also included in the system for acquiring spectra within optimal time frames.

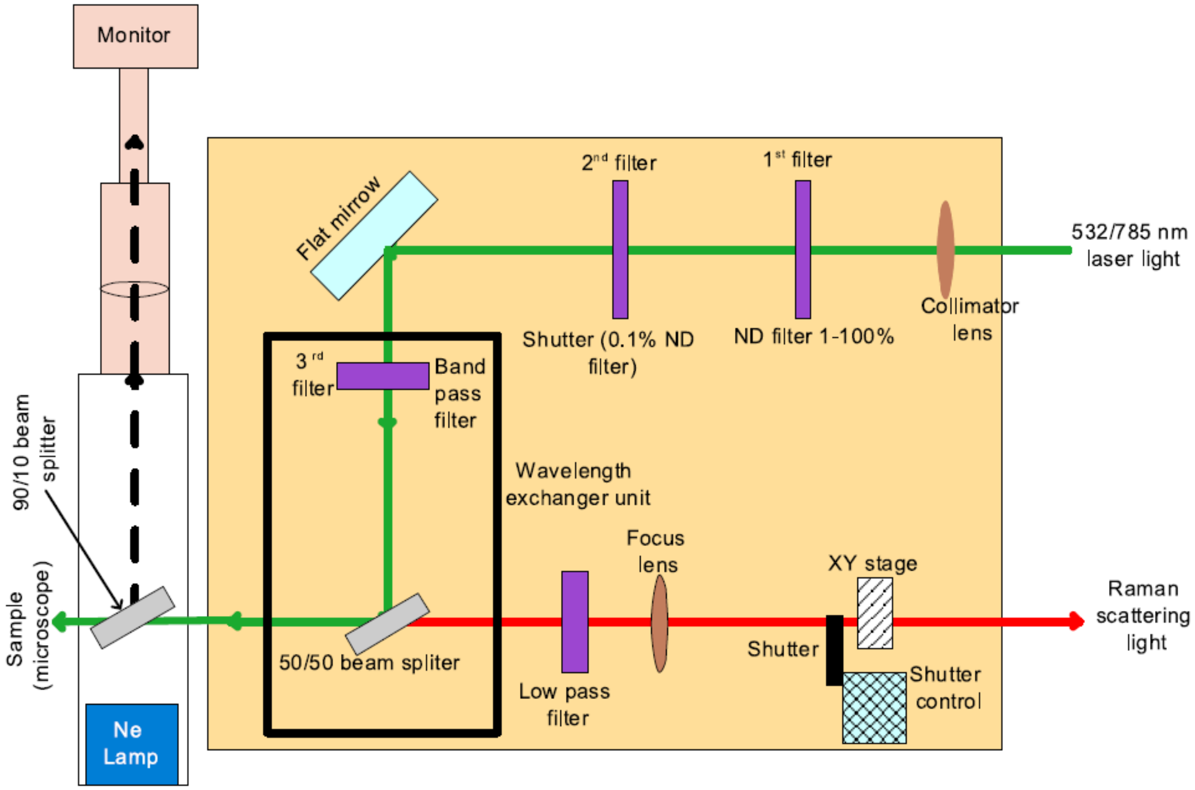

Figure 4.1: Layout of confocal laser Raman spectrometer. (Source: Cornes, 2012)

During measurements, a beam of red (785 nm) laser light was delivered to the Raman optics via a system of optical fiber. Once at the Raman optics, the beam is passed through the neutral density (ND) filter where it is 1-100% filtered to the shutter. Thereafter, the shutter then delivers the beam to the band pass filter. The band pass filter then conveys the beam of light to the beam splitter that splits it into two equal parts such that fifty percent of the beam is reflected while the rest passes through the beam splitter onto the sample where scattering occurs (see Figure 4.1).

The set up was equipped with a microscope that uses a lever to regulate the motorized stage for controlling the movement of the focused laser spot on the sample. Therefore, it was possible to focus the laser and record the Raman spectra at different points of a sample mounted on the stage. Once the sample has been excited, the scattered beam is then passed through the objective to the 785 nm low pass filter which blocks the Rayleigh scattered beam and only allows the Raman beam to pass through. Thereafter the beam is passed through fiber optic cables to imaging spectrometer and the CCD camera. Finally, the signal from the CCD camera is relayed to the computer equipped with STR software. This software provided an interface for controlling the components of the whole system in addition to providing decoding and visualization of the spectra recorded.

4.3 Preparation of Fruit Samples

The presence of CaC_2 was assessed on the peels of banana fruit samples. In order to carry out these studies, three groups of banana samples were required. For the first and second group of samples, mature but unripe bananas were harvested from Muga farm in Kisii County. The farm practices organic farming and this was vital as pure banana samples free from pesticides or insecticides were required for this study. All the bananas were cleaned first by washing them under running water.

A portion of the banana fruits were placed in a container and left to ripen naturally. These samples formed the control group (1st group). The other samples (2nd group) were ripened using commercial grade CaC_2 in the form of solution using the following treatment schedule. Different concentrations of CaC_2 solution were prepared ranging from 0.024% (0.24 g/ L water) to 0.2% (2 g/L water) in the same manner as prepared by Chandel *et al.* (2018). The masses were measured carefully using an electronic measuring balance whose sensitivity was 0.001 g. Starting with the smallest measured mass of CaC_2, the chemical was dissolved into a container having one litre of deionized water. Deionized water was used here to try and eliminate common impurities found in tap water. Banana fingers were then immersed into the solution and left for about half an hour. After this step, the fruits were then taken away from the solution and left to dry under air to remove adhering droplets. The treated fruits were then placed in containers and allowed to ripen for 48 hours. The process was repeated in steps for the subsequent concentrations up to the largest measured mass of CaC_2. Table 4.1 below shows the concentrations and the assigned labels for analysis purposes.

Table 4.1: Concentration levels of CaC_2 used in ripening samples

Sample labels	**CaC_2 ratio (g/L of water)**	**Concentration of solution (%)**
A1	0.24	0.024
A2	0.26	0.026
A3	0.28	0.028
A4	0.3	0.03
A5	0.34	0.034
A6	0.36	0.036
A7	0.4	0.04
A8	0.6	0.06
A9	0.8	0.08
A10	1	0.1
A11	1.2	0.12
A12	1.4	0.14
A13	1.6	0.16
A14	1.8	0.18
A15	2	0.2
A16	4	0.4
A17	6	0.6
A18	8	0.8
A19	10	1
A20	12	1.2
A21	14	1.4
A22	16	1.6

The last group of samples (3rd group) to be used in these studies was bananas which were already ripe. These bananas were purchased from local markets including Rongai, Kiserian, Marikiti and Gikomba markets, as well as from banana vendors around Chiromo campus.

4.4 Procedure for Laser Raman Spectra Acquisition

The confocal Raman set up in this study could not allow the spectra of banana to be taken as a whole. This is due to the limiting focusing distance between the objective lens and the sample placed on the stage. Thus, the different samples were therefore sliced to a profile depth of about 5 mm (spreading through the peel and flesh) and placed on glass slides. At this depth, the laser power was largely absorbed and scattered within the samples with minimum chances of reaching the glass slides upon which the slices were mounted (Gierlinger *et al.*, 2012). Figure 4.2 below shows how the sliced banana samples were mounted on the glass slides.

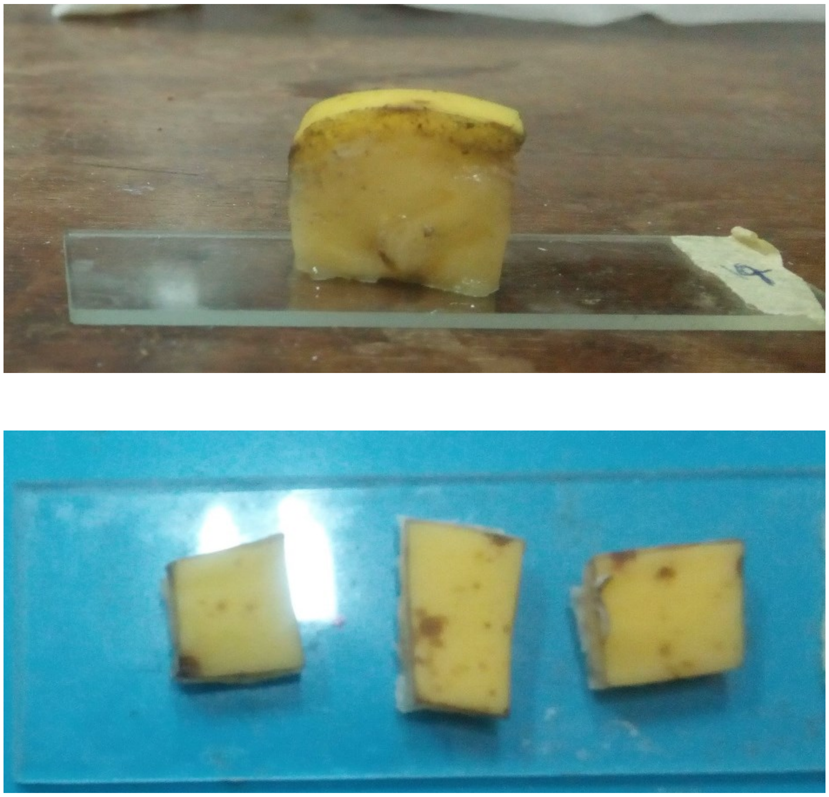

Figure 4.2: Banana sample slices for mounting under the microscope. Top image is a side view of the mounted slice whereas the image at the bottom is a top view same samples. Confocal LRS requires the samples to be as flat as possible.

The effect of substrates, upon which the sliced samples was placed, on the Raman spectra was evaluated. The Raman spectra was taken on samples placed on glass slides and then on samples placed on glass slides coated with conductive silver paste. Conductive silver paste is normally used to enhance the Raman signal. Notably, the net effect on the spectral response was negligible. Therefore, Raman spectra for all the sliced banana samples were taken when the samples were placed on glass slides. Nonetheless, as a precautionary measure, the glass slides were cleaned in alcohol and rinsed with distilled water before placing the sliced banana samples on them.

In order to have reliable data for analysis, further precautions were taken to optimize the Raman measurements. For instance, before any measurements were taken, the instrument was calibrated continuously over the course of the experiment using a standard silicon wafer which has peak at 520.5 cm^{-1} (McCreery, 2001). This ensured the recorded spectra was free from the problem of spectral axis drift discussed in section 3.5. This problem could have arose due to spectra which were recorded in different days. Further, to eliminate the effects of fluorescent light bulbs in the laboratory room, all measurements were done awhen the room was darkened. The temperature in the experiment room was also maintained at room temperature (24^{o} C) as variations in temperature could affect the measurements obtained. In a study conducted by Ghita *et al.* (2018) to investigate the change in intensity of Raman signals versus temperature, it was established that an elevation of temperature from 20 to 40^{o} C, lead to an increase in the signal (up to 2 fold) for biological samples. It was therefore imperative in this work to perform all measurements at the said constant temperature.

For every banana sample, five slices were obtained as described earlier to represent the whole sample. For each of these sliced samples, 6 spectra were obtained translating to 30 spectra per single banana sample. Thus, inhomogeneity due to unequal distribution of chemicals on the samples was compensated for in this manner.

During acquisition of the Raman spectra, 10 seconds exposure time was used over 5 accumulations as per the optimized conditions described earlier. We also used X50 objective lens with a laser power of 6.28 mW for the 785 nm laser as the best optimized combination for all Raman measurements (Gierlinger *et al.*, 2012).

Over the entire process of preparing the CaC_2 solutions and the banana samples as well as during spectra acquisition, safety precautions were observed. As discussed in section 1.1.1, continued exposure to CaC_2 presents neurological and carcinogenic health risks. Further, in the presence of moisture, CaC_2 liberates acetylene gas which is highly flammable. For this reason, the experiment was conducted in a well-ventilated room. Additionally, gloves and face masks were put on at all times when handling this chemical. The confocal LRS system used in this study is designed to ensure that the laser beam is maintained within the system's optics at all times. Nevertheless, laser safety goggles were worn to protect the eyes from stray laser beams, should they have been propagated.

4.5 Elemental Analysis of Calcium Carbide

Energy dispersive X-ray fluorescence spectroscopy (EDXRF) was used for determining the elemental composition of CaC_2 that was used in ripening bananas in this study. Industrial grade CaC_2 (75 % pure) was purchased from a local store and it comprised of chunks (<2cm wide) and granules. The CaC_2 granules were ground to loose powder to ensure uniform grain size. Thereafter, the loose powder was weighed in masses of 3.5 g before being pressed into pellets to ensure that the CaC_2 samples had uniform density across the analysis area. The pellets were then irradiated for 120 S and analyzed using a tube-excited EDXRF spectrometer. The results for elemental analysis are attached in Appendix 5.

4.6 Preprocessing of laser Raman spectra

The confocal LRS system used in this study was embedded with mechanism for background noise removal. The system had inbuilt parameters for removing cosmic-ray signals by use of interpolation based algorithm in the STR software. During the measurements, the spectrometer was calibrated continuously after a number of measurements. Thus, the recorded spectra were devoid of spectral axis drift discussed in section 3.5. Once the data was acquired, it was transferred to spectragryph software (Menges, 2017) for pre-processing and graphing before further analysis.

To begin with, the fluorescence background from the spectra had to be removed. This was done to obtain accurate laser Raman spectra as the fluorescence background tends to suppress the Raman signal in such measurements. Baseline offsetting was done by applying a modified polynomial fit with a coarseness value of 3 points to the raw data. The baseline was then subtracted from the spectra to have spectra free from the broad fluorescence baseline. Smoothing was then done using Savitzky-Golay algorithm with an interval of 9 points and polynomial of order 2 to remove noisy artefacts in the spectra as discussed in section 3.5. In order to take care of disparities in the intensity levels, normalization was done by SNV. This was to ensure that the intensity of a given Raman band of the same sample group was as similar as possible across the spectra recorded. Finally, the spectra were cut to a region of 200 – 1300 cm^{-1} as this was our region of interest.

The goal of doing data pre-treatment was to have spectrally cleaned data without loss of vital information. Therefore, the choice of adjustable parameters in the pre-processing steps was chosen on how best the preview fitted the data.

4.7 Software for Data Analysis

Two software were used for data analysis in this study; Spectragryph (Menges, 2017) was mainly used in the previous step of data pre-treatment. Once the data was cleaned, it was then transferred to R (version 3.5.3), an open source software, which was used as the main software for data visualization and analysis. R has inbuilt packages dedicated for spectral data analysis and machine learning applications. The software consists of functions for plotting and inspecting spectra, peak alignment, principal components analysis and model-based clustering, regression and prediction. Thus, the software was well suited for this study. The specific packages which were used include chemospec, caret and neuralnet.

4.8 Utility of Machine Learning Techniques in the analysis of laser Raman spectra

Machine learning is capable of handling complex systems such as the laser Raman spectral data that was obtained in this work. Multivariate ML techniques were used to overcome problems associated with classical/univariate analysis of Raman data. Supervised and unsupervised methods were used for performing exploratory and multivariate calibration of acquired laser Raman spectra.

4.8.1 Exploratory analysis of laser Raman spectra utilizing PCA

In this work, the spectral data was represented by a data matrix consisting of 42 rows that corresponded to samples for use in qualitative studies. These included 20 spectra for naturally ripened samples while the rest were spectra for artificially ripened samples. On the other hand, for qualitative studies, the spectral data was represented by a data matrix consisting of 224 rows correspond to artificially ripened samples which were treated at different concentrations. The original number of columns for both matrices was 1024 but this was reduced to 636 after performing a spectral cut for the region of interest as discussed in data-pretreatment.

PCA models were built using the samples with the main aim of detecting the composite features (PCs) that would be later used as inputs for the ML models. Some PCA models were developed using subsets of wavelengths corresponding to certain Raman vibrational bands of the compounds in the artificial ripener. The graphical visualization of scores and loading plots provided insights on the vibrational modes of the molecules present in the samples.

4.8.2 Multivariate Calibration of Laser Raman Spectra Utilizing ANN

In this work, ANN models were developed for classification and regression. In both cases, PCs were used as inputs to the ANNs. Further, the wavelengths and their corresponding intensities were used as inputs in developing the models and a comparison made with the former choice of inputs. Regardless of the choice input, the input data was scaled using the min-max approach. Feedforward ANNs trained by backpropagation were used in this study. In a broad view, the process of model development consisted of assembling the data, creating the network object, training it and predicting the response to new inputs.

Initial values of weights were generated randomly by the command in R once the BP-ANNs were fed with input data (training set). Nonetheless, some parameters had to be manually selected and tuned for the models to converge in the shortest time possible. The number of neurons and size of the hidden layer was selected based on the number of input variables (Huang *et al.*, 2007). A network's complexity is determined by the number of neurons and layers, and maximizing these parameters is crucial; too many neurons and/or layers may result in overfitting to the training data, while too few neurons and/or layers may not have enough complexity to discriminate the data spectrally (Madden and Ryder, 2003). In practice, this range of parameters can be chosen using optimization functions such as random search (used for this work) or genetic algorithms.

The output layer size was determined from the size of the targets and whether the model was regression or for classification. The transfer function used for hidden layers (non-linear) was logistic while purelin (linear) was employed as transfer function for the output layer in the case of regression models. For classification, both the hidden and the output layers had non-linear transfer functions. The error function that was used by the BP-ANNs for adjusting the network weights was the sum of square errors (SSE). The network was trained until the SSE was minimized. However, since the initial weights were randomized by the network, the output varied each time. Therefore, the training process was repeated several times and the trained network that provided the best performance was retained.

Once the neural network analytical model was adopted, the model was first put to test by simulating the output of the neural network with the measured input data in the case or regression. The results herein were then compared with the measured outputs. The models were then validated using an independent (test) data set that the BP-ANNs had not been exposed to. The model's performance was evaluated by root mean square error of prediction (RMSEP) and coefficient of determination (R^2). RMSEP was calculated using the formula:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n}} \tag{4.1}$$

where y_i is the predicted concentration by ANN, $\hat{y_i}$ is the actual concentration, and n is the number of test samples. R^2 was calculated using the formula:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y_i})^2}{\sum(y_i - \bar{y})^2} \tag{4.2}$$

where the symbols have the same meaning as in equation (4.1) and $\bar{y}$ represents the actual mean concentration. In the case of classification, confusion matrices were drawn and misclassification error rate calculated. Appendices 2 and 3 contain a detailed description of the commands and algorithms used for the neural network models.

4.8.3 Multivariate Calibration of Laser Raman Spectra Utilizing RF

Random forest (RF) is a highly predictive machine learning technique that is able to model and extract data from target systems that are complex and non-linear (Cutler *et al.*, 2007). In this work, RF was used to develop a classification scheme for samples ripened naturally versus samples ripened artificially using the PCs data of their Raman spectra. Additionally, multivariate calibration schemes were developed from RF models for predicting the concentration of artificial ripeners in samples ripened using CaC_2.

Decision trees were built in the following manner in R: The PCs data set was first divided into a training and test data set. Initially, a number of samples were randomly selected from the training set with replacement (bootstrap resampling technique) and used to construct an equivalent number of classification or regression trees (Zhang *et al.*, 2014). Some samples were repeated whereas others were left out to form the OOB data, which was later used to calibrate the performance of each tree line. The splitting (decision making) criteria as discussed earlier in section 3.6.4 was applied from the root to the terminal nodes. The node building procedure was applied repeatedly until full classification of the selected samples for a given tree was achieved and each classified sample assigned to the corresponding reference concentration or class. By repeating each tree building step, k independent trees were built to form the RF model(s). The average results from the *k* trees was therefore the output.

Two important parameters were optimized in this model using the OOB data: number of trees (ntree) and the number of variables randomly tried at each split (mtry). This was done by comparing the effect of the number of trees on the OOB error and by evaluating the variables that had much more significance than others and adjusting the RF model accordingly. The trained model which gave the least errors (MSE) of OOB data was adopted.

The performance of the adopted RF classification model was evaluated by drawing confusion matrices and calculating correct classification accuracy of the test data. On the other hand, the performance of the random forest regression model was evaluated using the RMSEP and R^2 metrics on the test data. Appendices 2 and 3 contain a detailed description of the commands and algorithms used for the RF models.

4.8.4 Multivariate Calibration of Laser Raman Spectra Utilizing SVR

In this work, support vector machines was used for discrimination of samples based on the presence and the concentration of compounds resulting from hydrolyzed CaC_2. Quantitative features (sample concentration owing to intensity values) and qualitative features (spectral peak bands) of each of the samples can infer whether the samples were ripened artificially or not. Considering the Raman spectra of naturally ripened bananas in relation to artificially ripened ones, SVR hyperplanes can be used to classify a new sample as to whether it contains the artificial ripeners or not.

Using R, PCA for Raman spectral data of banana samples ripened naturally versus artificially was done and the principal components plus their score values were saved as a matrix. The saved data was then split into a training and a test set before the data was fed into the model. The PCs were used as the predictors whereas the class labels and the concentrations were the response values in the case of classification and regression respectively.

The model was run several times by varying the number of support vectors (PCs) in a bid to establish the PCs which would best classify the data. Additionally, the model was tuned by adjusting the gamma and cost parameter so as to obtain the best parameter combination that produced optimal hyperplanes. Notably, both linear and radial kernels were utilized. Training automatically stopped when the model's calibration error were reduced to a minimum. Samples that had not been shown to the model (test set) were then fed to the model to determine the clustering ability of the model. Confusion matrices were drawn and misclassification error calculated. The classification of the samples into their correct groups is a useful component of this work as this can be utilized in a rapid identification of fruit samples ripened artificially. For regression, the model's ability to predict correctly the concentration of the measured values for the test set was evaluated by RMSEP and R^2. Appendix 2 and 3 contain a detailed description of the commands and algorithms used for the SVM and SVR models.

In order to validate the predictive ability of multi-molecular calibration models in banana samples, we compared the result of support vector regression with random forest regression and artificial neural network regression, by means of prediction accuracy, R^2 and RMSEP. The schematic of the ML modelling approaches proposed in this work is given in figure 4.3 below.

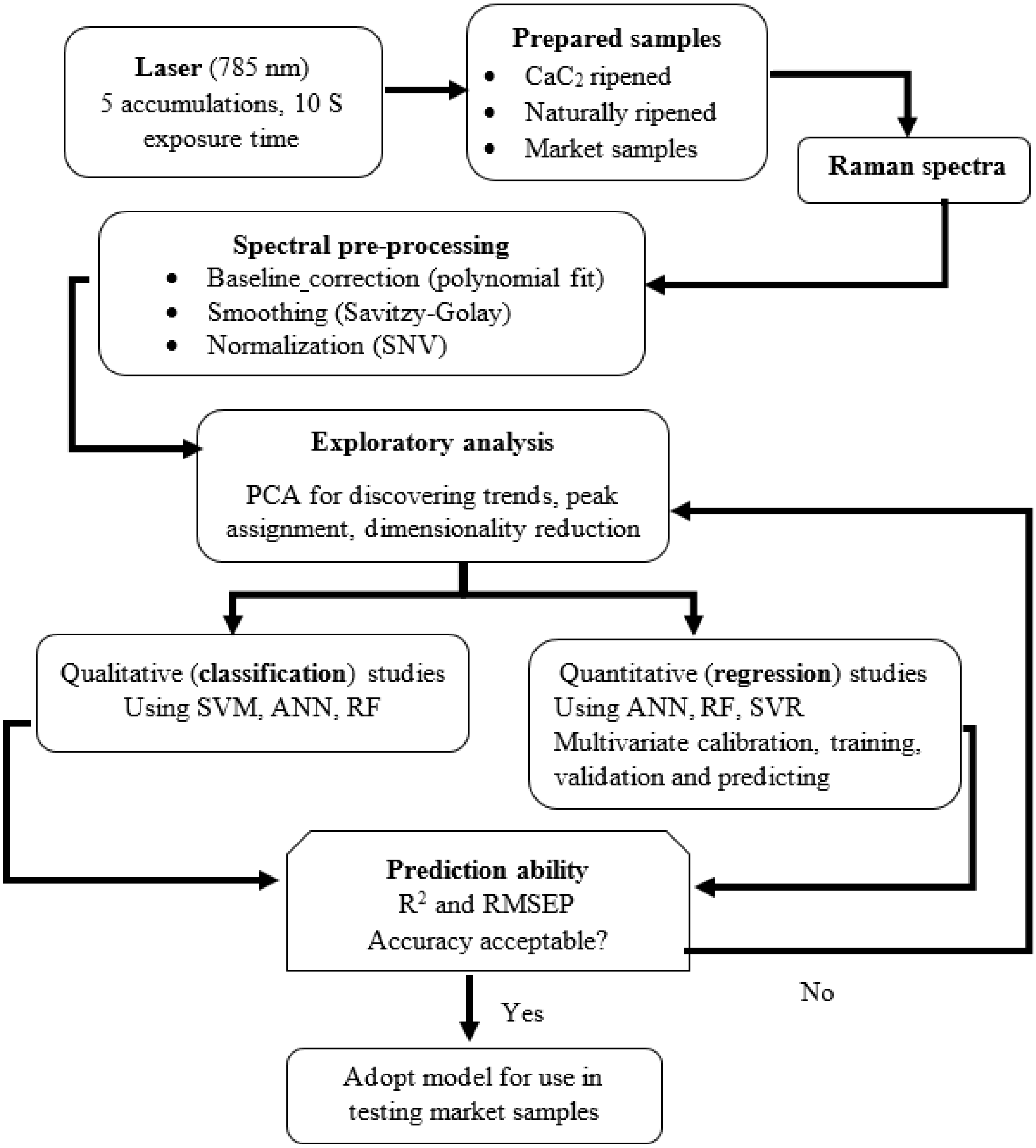
Laser (785 nm)
5 accumulations, 10 S exposure time
Prepared samples
CaC$_2$ ripened
Naturally ripened
Market samples
Raman spectra
Spectral pre-processing
Baseline_correction (polynomial fit)
Smoothing (Savitzy-Golay)
Normalization (SNV)
Exploratory analysis
PCA for discovering trends, peak assignment, dimensionality reduction
Qualitative (classification) studies
Using SVM, ANN, RF
Quantitative (regression) studies
Using ANN, RF, SVR
Multivariate calibration, training, validation and predicting
Prediction ability
R^2 and RMSEP
Accuracy acceptable?
No
Yes
Adopt model for use in testing market samples

Figure 4.3: Conceptual framework for machine learning methodologies employed towards calibration and prediction of Raman spectral data.

4.9 Evaluation of Limits of Detection (LOD) and Quantification (LOQ)

In order to determine the least CaC_2 concentration that could be detected and quantified by the ML-assisted LRS system, the LOD and the LOQ had to be evaluated. For any given analytical technique, the LOD refers to the minimum analyte concentration that is detectable and can be proven to be present with a certain degree of confidence (Uhrovčík, 2014). On the other hand, LOQ refers to the minimum analyte concentration that can be quantified with reasonable reliability. These two limits are normally derived from univariate calibration approaches whereby single instrumental measurements are done per sample. The formulae for calculating univariate LOD and LOQ are well prescribed by the International Union of Pure and Applied Chemistry (IUPAC) (Chandran and Singh, 2007).

In the present work, the Raman spectra for the analytes of interest were recorded amidst a background of unknown number of interfering species. As such, there were multiple instrumental data for a single sample. The universal univariate calibration approach in this case fails to adequately cater for the multiple spectral responses in this data as discussed in section 3.6. Consequently the well-defined IUPAC formulae for evaluating LOD and LOQ could not be correctly applied in this case. Thus, the multivariate approach for determining LOD and LOQ values was adopted in this work.

Currently, there is no well-defined procedure for obtaining LOD and LOQ in multivariate calibration as this is still an active research area (Uhrovčík, 2014). However, the normative approach involves plotting the model-predicted analyte concentration against their measured concentration (Shrivastava and Gupta, 2011). In this work, the analyte concentration were obtained from the multivariate calibration curves of the ML models and plotted against actual concentrations as recorded in Table 4.1. The graphs were then processed as univariate graphs. In this pseudounivariate approach, the LOD and LOQ were computed using the following reduced formulae derived from the standard LOD and LOQ equations (Shrivastava and Gupta, 2011):

$$LOD = \frac{3\sigma}{S} \tag{4.3}$$

$$LOQ = \frac{10\sigma}{S} \tag{4.4}$$

Where S is sensitivity (slope of the calibration curve) and σ is the standard deviation of response obtained by several approaches. The method of using standard deviation of the y-intercept of the regression line was used herein as it is more accurate than using the mean blank signal approach. It is important to note that this approach provided averages for LOD and LOQ of the whole methodology and not the LRS instrument independently.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Chapter Overview

In this chapter, the results of the analysis of LRS of banana samples based on the spectra obtained from the control and treated samples as well as market samples are presented. In addition, optimal instrumental and analysis conditions for conducting this study are discussed in this chapter. In addition, results of exploratory qualitative and quantitative analysis of LRS data using multivariate ML methods are presented.

5.2 Optimized Conditions for Rapid Laser Raman Spectroscopic Measurements

The STR system used in this study was equipped with two lasers for excitation; 532 nm and 785 nm. Owing to the nature of this study, whereby Raman spectra of bananas are to be recorded, the 785 nm was preferred. Such kinds of sample tend to suffer from a strong fluorescence background, particularly, when short excitation wavelength lasers are used. A laser in the NIR region, say 1025 nm, will likely overcome this challenge and so the justification for the choice of 785 nm laser in this study (Zhang et al., 2006a) .

Optimization of the spectral window to be used depends on the region where most vibrational bands of the sample occur. The Raman spectrometer used in this study was therefore set to cover the shift range 37 – 1800 cm^{-1} which encompasses the fingerprint region. Further, most of the vibrational bands for the compounds of interest were expected to be found in this region. The spectrograph's CCD camera was set at 256 X 1024 pixels.

The excitation laser light intensity was regulated by use of ND filters with different attenuating capabilities. For this study, the laser power was set to maximum (100%) as better quality spectra were realized for this laser power setting as compared to lower power settings. In addition, no burning of the sample was observed. This was ascertained by inspecting the microscope objective for presence of residues and inspecting the spectra for flat-line effect which normally arises due to CCD saturation brought about by high laser intensity (Butler *et al.*, 2016). In both cases, these common pointers of photo-damaging did not feature. In order to measure the laser excitation field powers at different objective lenses, an Orion Laser Power meter from Ophir photonics was used. The intensities are as shown in Table 5.1 below.

Table 5.1: Power of different lasers at different objectives

Objective	**Numerical Aperture (NA)**	**Power to sample** (mW)
X10	0.30	10.45
X20	0.45	8.87
X50	0.80	6.72
X50	0.50	6.28
X100	0.90	5.57

In this work, the objective was set to X50 with NA = 0.5. At this resolution, it was possible to see the features of the banana samples up to the cellular level.

Figure 5.1 below shows the microscopic images of naturally and carbide ripened banana samples at different concentrations obtained from the confocal LRS system at X50 objective lens magnification.

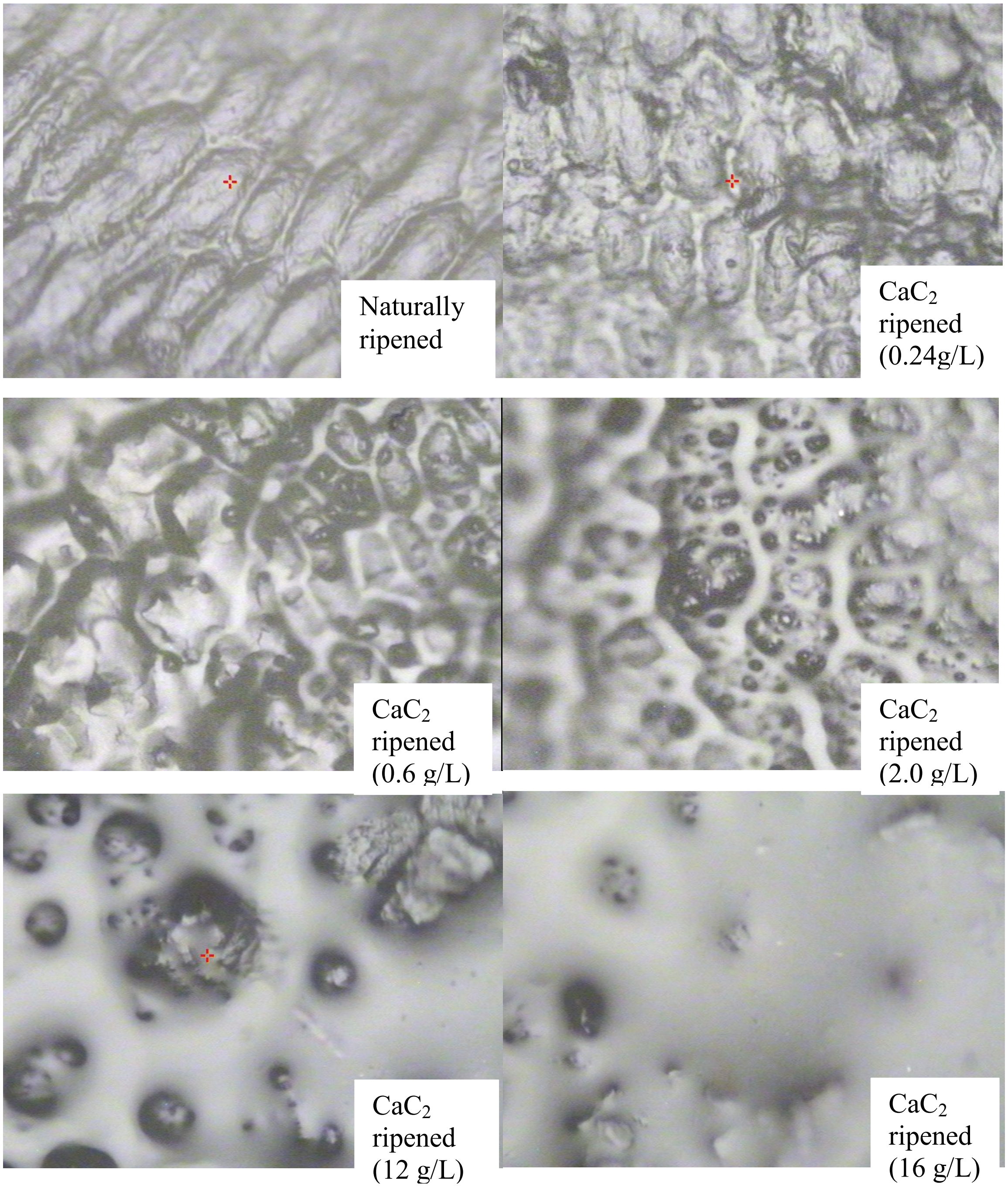
Naturally ripened
CaC_2 ripened (0.24g/L)
CaC_2 ripened (0.6 g/L)
CaC_2 ripened (2.0 g/L)
CaC_2 ripened (12 g/L)
CaC_2 ripened (16 g/L)

Figure 5.1: Images of banana samples as seen under the microscope (X 50 lens) Deposits are seen on the artificially ripened samples and the amount of deposits seen increases with the increase in CaC_2 used in ripening.

From the images in figure 5.1, it is clear that when CaC_2 is used for ripening banana samples, deposits of its impurities remain embedded in these samples. As discussed in section 1.2, these impurities are fat soluble and they can diffuse through the cell wax of the peel into the flesh. Further, it is also evident from the images in

Figure 5.1 that when higher concentrations of the ripening agent are used, the residue concentration also increases accordingly.

One of the advantages of using confocal LRS in this study was its ability to focus the laser to particular points on the mounted samples which showed residues of CaC_2. Setting the objective lens at X50 made it possible to focus the laser on the points of interest by utilizing the laser spot diameter.

5.3 Raman Spectra of Samples before and after Data Preprocessing

Under the same conditions described in section 4.4, the Raman spectra of banana samples ripened naturally and artificially were recorded at several surface points. The typical spectra obtained are shown in Figure 5.2 and Figure 5.3 below. Figure 5.2 shows the spectra for the samples ripened naturally whereas Figure 5.3 shows the spectra for samples ripened using CaC_2 at different concentrations. The fluorescence background was strong for the two groups of samples, and the spectra demonstrated no obvious characteristic peaks among various samples. Notably, intensity differences were readily noticeable in the two groups of spectra. The carbide ripened spectra portrayed relatively higher intensity values as compared to the naturally ripened spectra. This agrees with discussions in section 3.3 that Raman scattering (intensity) varies directly as the concentration of the analyte present amongst other factors. However, this relationship did not hold for all spectra recorded for different concentrations as will be discussed.

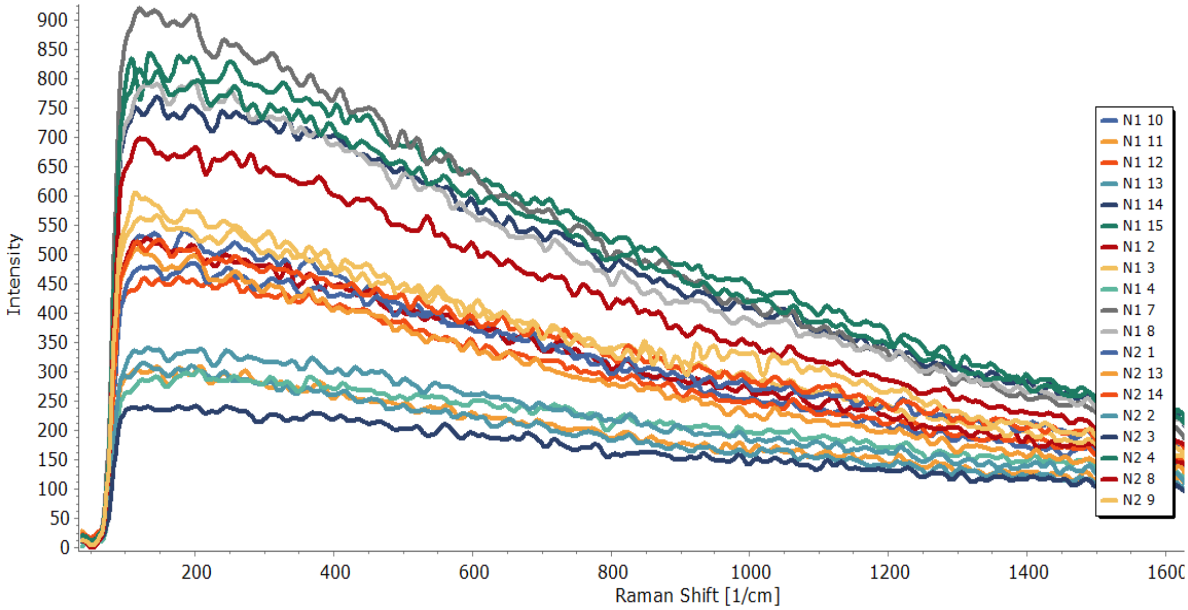

Figure 5.2: Raw Raman spectra for naturally ripened banana samples. The N numbers represent different samples ripened naturally.

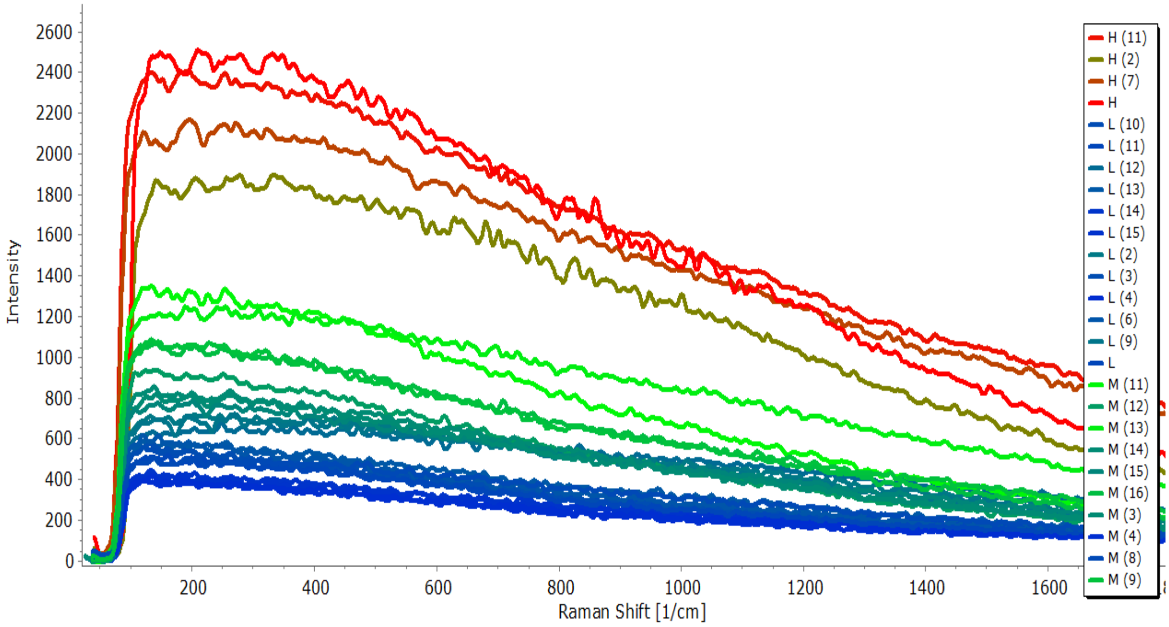

Figure 5.3: Raw Raman spectra of CaC_2 ripened banana samples. H represents samples ripened with high concentrations, M represents medium whereas L represents low concentrations.

As observed from the Raman spectra above, the broad fluorescence background (Raman shift region <400cm^{-1}) obscures the Raman signal of the molecules of interest. The strong autofluorescence was produced by molecules of carotenes (alpha and beta), tyrosine and folate contained in the bananas (Yakubovskaya *et al.*, 2019). Thus the spectra had to be cleaned to reduce the noise and enhance the desired spectral features. Since multivariate ML methods rely on the shape of the peak/ Raman band to a small extent, band overlaps and other spectral matrix effects were resolved through a series of spectral pre-processing steps to remove spectral noise.

The first step was to subtract the broad fluorescence baseline. Modified polynomial fitting was applied to correct the drifting baseline. This was then followed by smoothing using Savitsky-Golay algorithm with a window of 7 points and polynomial of order 2. The choice of these parameters was based on the best smoothing capability for the spectra specific to this study.

To correct for intensity variations in the spectra, normalization was done using SNV. The non-linearity in the spectral response that was observed (in terms of the signal intensity relative to the concentration) can be attributed to several factors. To begin with, the banana samples were non-homogenous. This had the implication that the spectra obtained had overlapping signals owing to the numerous Raman active molecules contained therein. Imperfections in the LRS system optics and CCD detector giving non-linear responses in the presence of stray light could also have contributed to the non-linearity in the recorded spectra. Lastly, chemical factors like intermolecular reactions would have affected the bonding structures resulting to a shift or broadening of the Raman bands. For these reasons, deviations in the spectra in terms of peak regions and non-linear intensity response was observed and that explains why normalizing the spectra was imperative.

The classical approach of assigning peaks to a Raman band of interest was not sufficient under these circumstances. A multivariate approach would thus form a better basis as will be described later in section 5.6. Notably, the proposed multivariate ML methods that were used in this work were capable of overcoming the problem of non-linear data in developing regression models as will be discussed in subsequent sections.

Outlier detection was an important aspect before a calibration strategy was employed. At this point, spectra which manifested deviation to a greater extent relative to the others were removed. As a final step, data compression was done by selecting only appropriate region of interest (200 cm^{-1} – 1200 cm^{-1}). This was important to reduce the computational burden during development of the ML models as there would be fewer inputs relating directly to the analyte of interest and that our models will not be modeling noise.

The preprocessed Raman spectra for the two groups of samples are shown below:

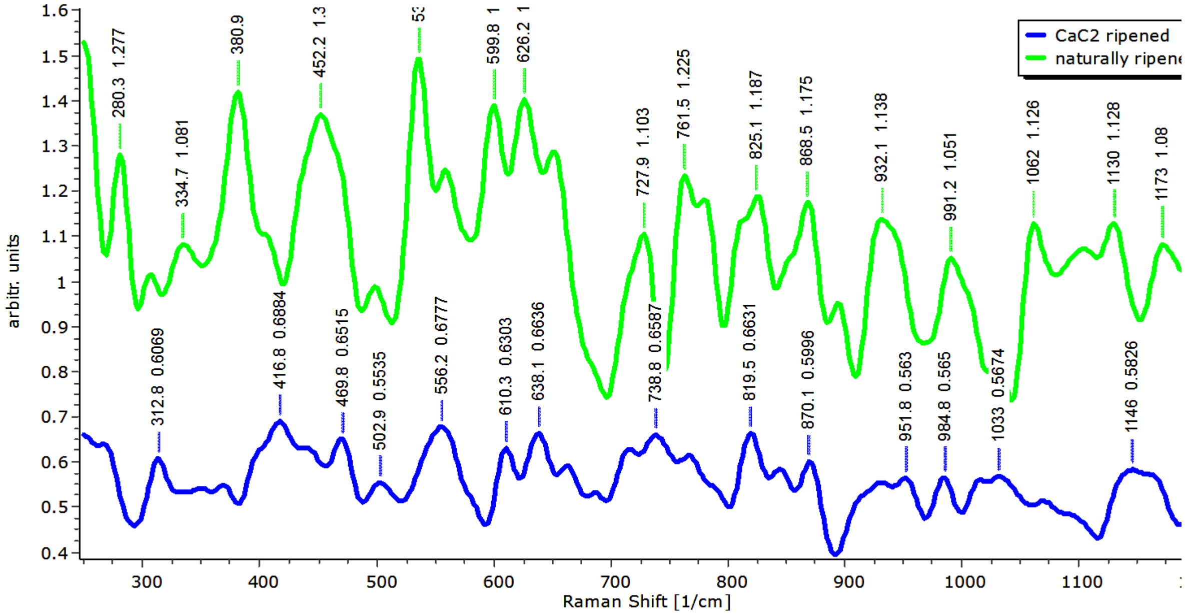

Figure 5.4: Pre-processed Raman spectra of naturally and carbide ripened banana samples

After pre-processing, it can be seen that the obscured Raman signal becomes enhanced and the random instrumental noise is reduced significantly. This shows that appropriate mathematical software computations can be carefully employed to clean noisy Raman spectral data. The cleaned data can thereafter be used appropriately in the subsequent analysis steps.

5.4 Exploratory Multivariate Analysis of Raman Spectra by PCA and Assignment of Peaks.

The chemical compositions of fruits are complex. The cleaned Raman spectra in Figure 5.4 above exhibit some characteristic features of the various chemical components therein. However, it is difficult to visually distinguish between the two groups of samples from their spectra as several peaks overlap and some of the peaks are not clearly defined. It is also evident that the spectra contain noisy artifacts (unwanted spectral information) which make it difficult to identify the spectra by visual inspection. PCA was therefore used for exploratory multivariate analysis of the whole spectral region of the cleaned Raman spectra. The scores and loadings plot were used for identifying and assigning peaks responsible for the group differences. Principal components are the basis behind groupings in score plots since for some spectra to be clustered in one group they must have factors that are similar (Shaver, 2001). In the case of quantitative studies, PCA was used for dimensionality reduction such that the PCs were used as the inputs as will be discussed later. PCA for the whole spectral ROI gives a clear distinction between the samples ripened artificially and samples ripened naturally as shown in Figure 5.5 below:

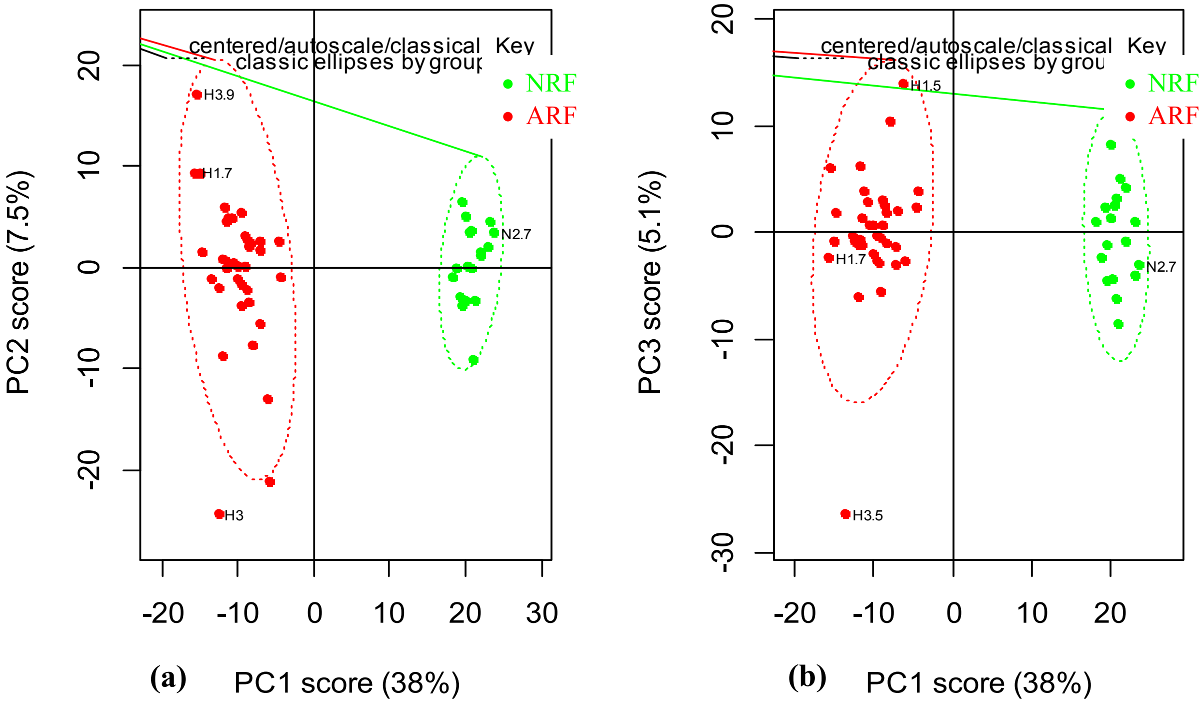

Figure 5.5: PCA scores plots of naturally (NRF) and artificially (ARF) ripened samples for waveshift region 200 -1200 cm^{-1}.

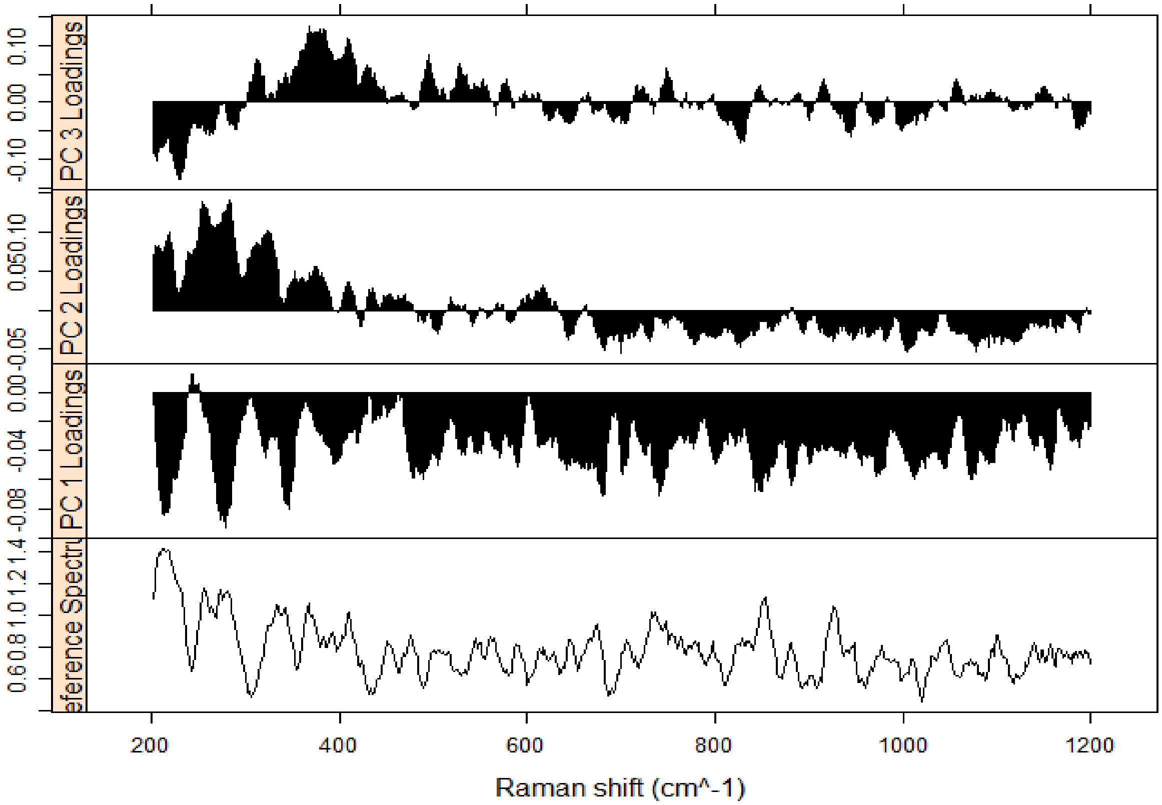

Figure 5.6: PCA loadings plots of naturally and artificially ripened samples for waveshift region 200 -1200 cm^{-1}

However, the PC scores recorded for this explained low percentages of the variability. About 10 PCs were needed to explain 95 % 0f the variability in this dataset which contained the spectral region after pre-processing. This could be attributed to the fact that PCA was done including regions which did not have molecules of interest and therefore constituted noise in the PCA model. Therefore, PCA was further done in the Raman shift regions where the target molecules resulting from artificial ripening by calcium carbide are found (as highlighted earlier in Table 2.1).

5.4.1 PCA for Raman shift Region 450 cm^{-1} - 500 cm^{-1}

Owing to the production process of CaC_2, sulfur is usually contained in larger proportions as compared to other impurities in CaC_2. The focus for PCA for this band region was the sulfur peak centered at 480 cm^{-1}. Figure 5.7 and Figure 5.8 below show the scores and loading plots for this ROI. The sulfur molecule bonding is also shown alongside.

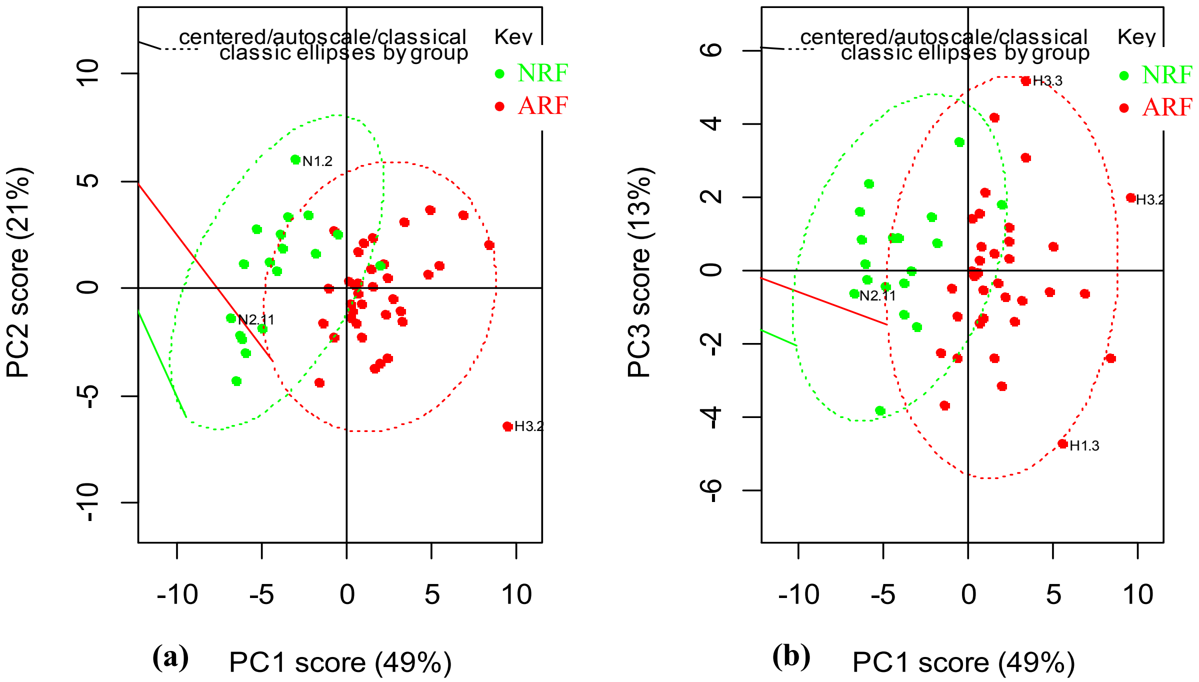

Figure 5.7: PCA scores plots for sulfur molecules ROI (450 -500 cm^{-1})

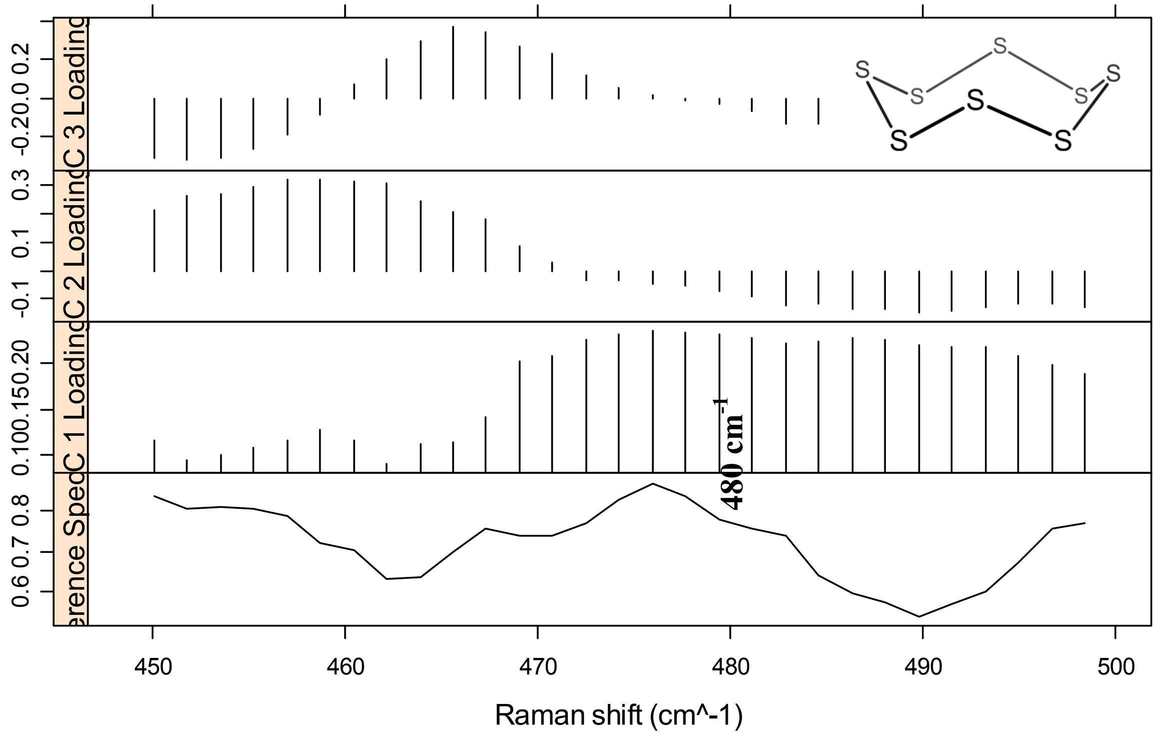

Figure 5.8: PCA loadings plot for sulfur molecules ROI (450 -500 cm^{-1}). The puckered ring bonding of the sulfur molecule is shown at the top right corner.

The scores plot of the first two PCs depict a distinction between the carbide ripened and the naturally ripened samples. The positive of PC1 receives influence almost exclusively from the ARF samples. Correlating this to the loadings plot, this influence peaks at around 480 cm^{-1}. This peak can be assigned to ν(S-S) cross-ring stretching vibration mode of sulfur molecules that exist as puckered rings. Thus, from the loadings and scores plots, we can conclude the clustering of carbide ripened samples in this band region was influenced by the presence of sulfur molecules in the ARF samples.

5.4.2 PCA for Raman shift Region 600 cm^{-1} - 650 cm^{-1}.

This wavelength region was evaluated to investigate the presence of Acetylene molecules in the ARF samples. As discussed earlier, acetylene liberated from hydrolyzed CaC_2 is the one which speeds up the ripening process. The scores and loadings plots are as shown in Figure 5.9 and Figure 5.10 below:

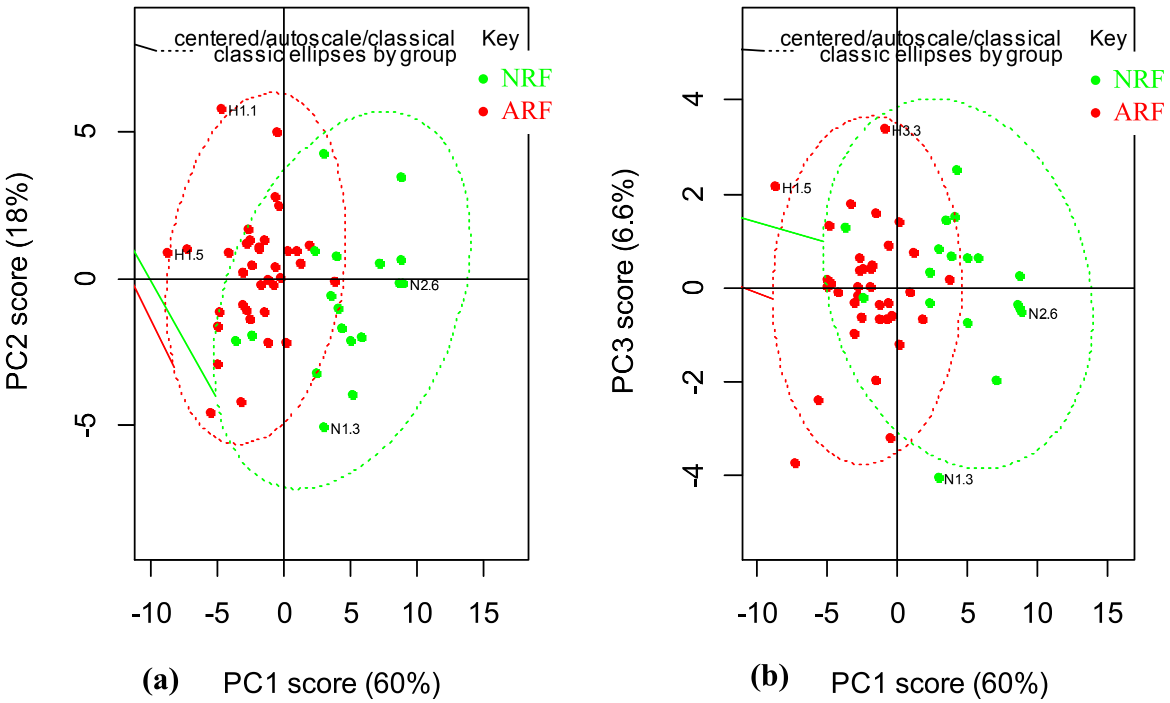

Figure 5.9: PCA scores plots for acetylene molecules ROI (600 -650 cm^{-1})

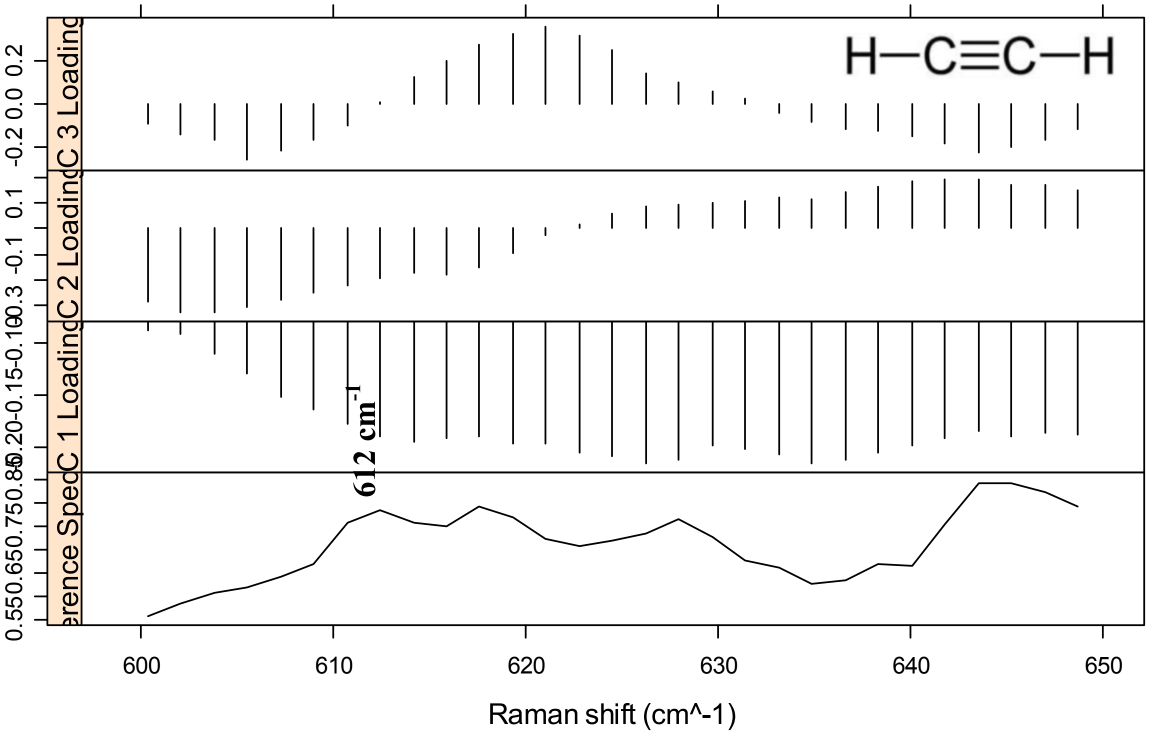

Figure 5.10: Loadings plot for acetylene molecules ROI (600 -650 cm^{-1}). The bonding structure of the acetylene molecule is shown at the top right corner.

The negative of PC1, which explains about 60 % of the variability in this wavelength region, is dominantly influenced by the ARF samples. The negative of PC3 is also largely influenced by the ARF samples. A comparison of these influences with the loadings plot can be attributed to the (C-H) out-of-plane bending mode of acetyl molecules with a peak at 612 cm^{-1}. This mode presents a very weak vibrational mode of the acetyl group.

5.4.3 PCA for Raman shift Region 750 cm^{-1} - 800 cm^{-1}.

In the chemical reaction of CaC_2 with water in equation (1.1), we saw that calcium hydroxide is formed alongside the liberated acetylene gas. PCA was done for this region to investigate whether the two groups of samples could be distinguished based on their Raman spectra for this band region. Figure 5.11 and Figure 5.12 below shows the scores and loadings plot for this ROI.

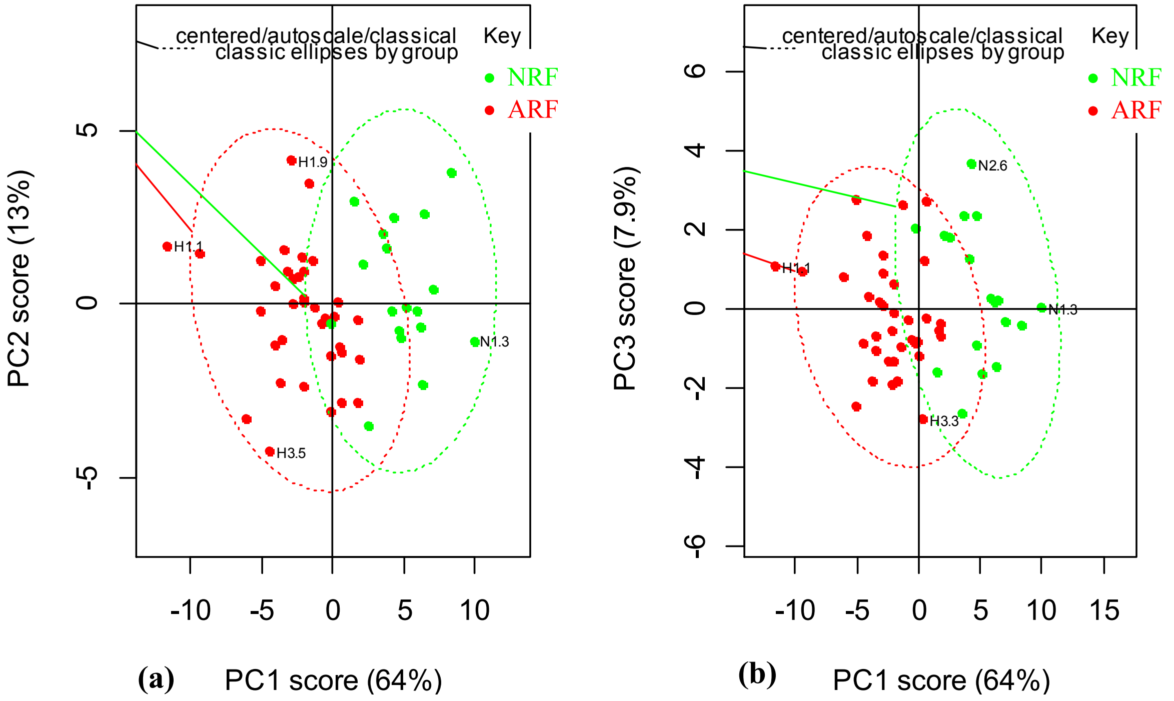

Figure 5.11: PCA scores plots for hydroxyl molecules ROI (750 -800 cm^{-1})

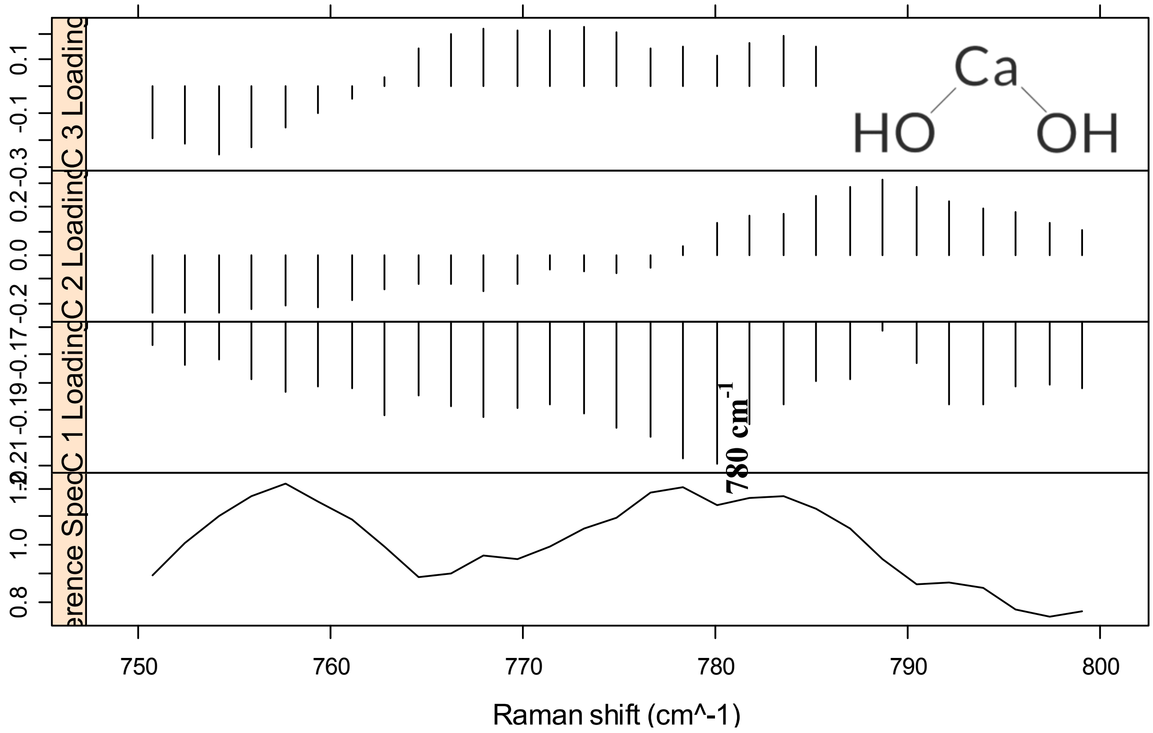

Figure 5.12: Loadings plot for hydroxyl molecules ROI (750 -800 cm^{-1}). The bonding structure of calcium hydroxide molecule is shown at the top right corner.

From the scores plots, it is clear that the ARF samples give major influence to the negative of PC1, which explains 64 % of the variability of the data in this band region. When compared to the loadings plot, this negative influence of PC1 is more pronounced at the peak centered at 780 cm^{-1}. This peak can be assigned to the (O-H) asymmetric bending vibration of the hydroxyl molecule. Therefore, the clustering in this band region is as a result of the calcium hydroxide molecules present in the ARF samples.

5.4.4 PCA for Raman shift Region 950 cm^{-1} - 1000 cm^{-1}.

In order to assess whether fruits have been ripened artificially using CaC_2, the test for the presence of phosphine molecules in those fruits is one of the common pointers of this practice. The scores and loadings plots for this ROI are shown in Figure 5.13 and Figure 5.14 below:

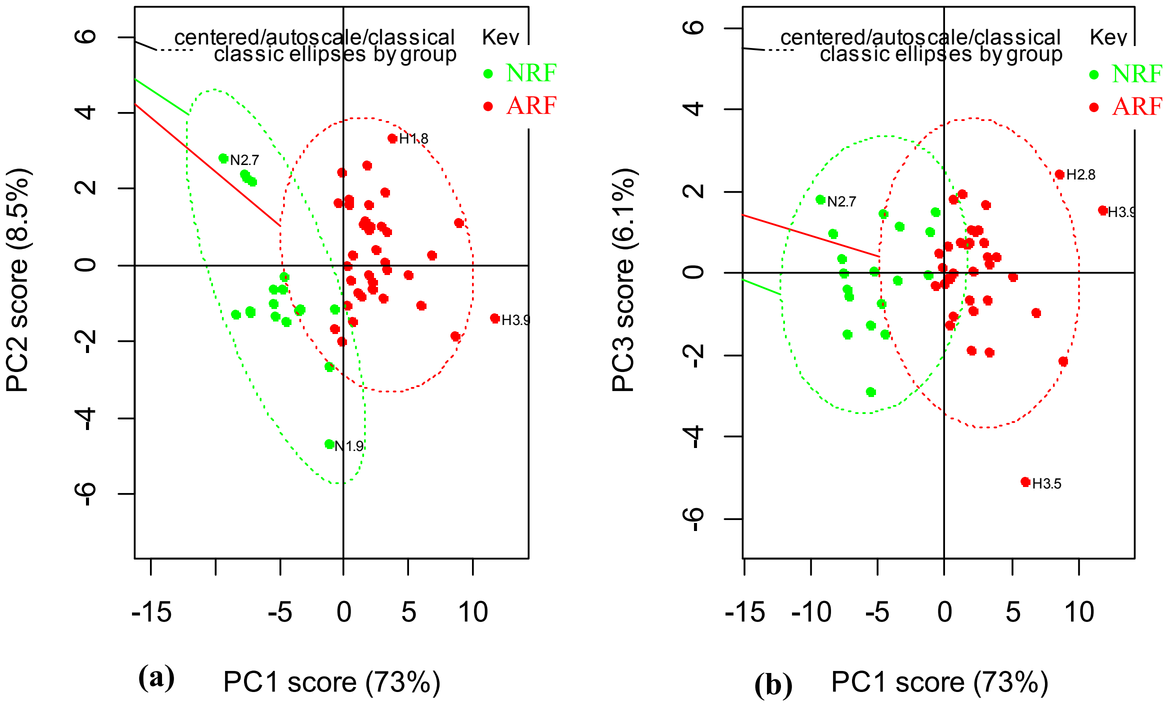

Figure 5.13: PCA scores plots for phosphine molecules ROI (950 -1000 cm^{-1})

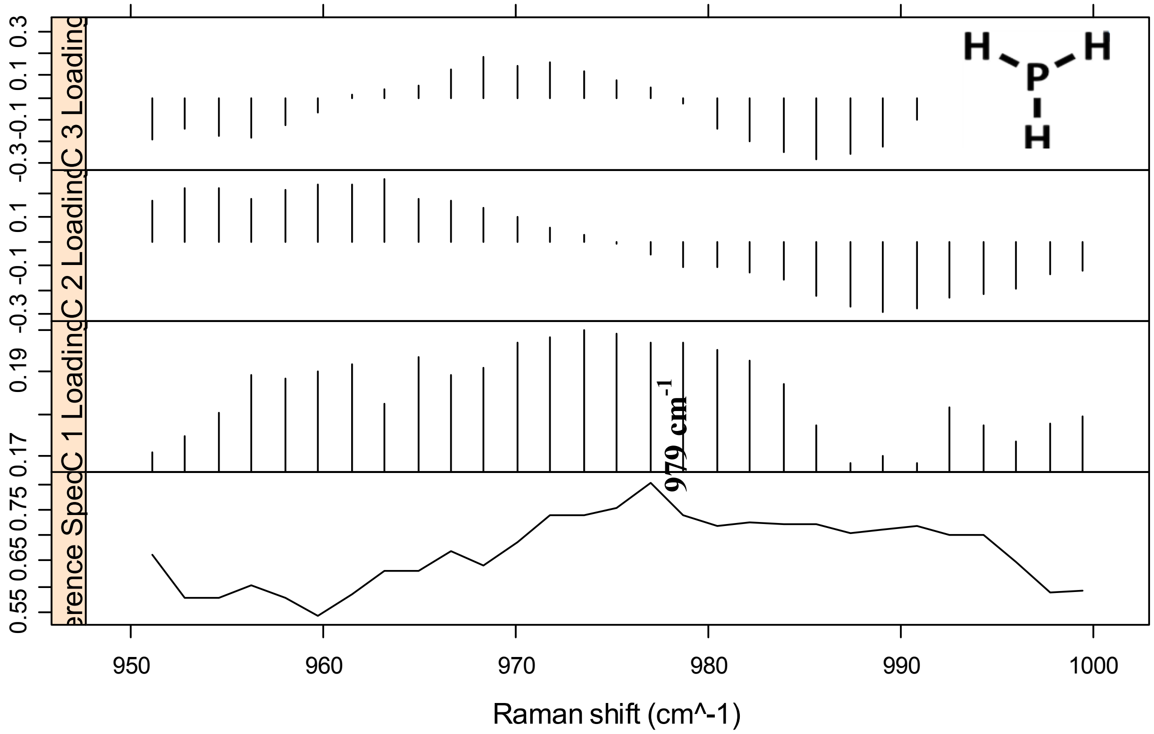

Figure 5.14: Loadings plot for phosphine molecules ROI (950 -1000 cm^{-1}) The bonding structure of phosphine molecule is shown at the top right corner.

Positive PC1, explaining 73 % variability of data in this ROI, is strongly and exclusively influenced by the ARF. Correlating this with the peak at 979 cm^{-1} in the loadings plot, where PC1 is most positive, we can conclude the clustering in this band region is due to presence of phosphine molecules in the ARF samples. This Raman vibrational peak centered at 979 cm^{-1} can be assigned to the out-of-plane bending mode of the non-polar covalent bonds of phosphine molecules.

From these results, we arrive at the initial finding that PCA can be used, on an exploratory basis, for qualitative identification of fruits ripened naturally versus fruits ripened artificially from their Raman spectra. Nonetheless, we note that the chemical composition of bananas is complex as there are usually different amount of water, sugar, carotene, protein, fat, vitamin, as well as other components and elements such as calcium, iron and phosphorus (Bari *et al.*, 2018). The assignment of peaks to molecules in this complex domain of mixtures using univariate way of analysis was therefore not sufficient as there were peak shifts owing to the complexity of the mixture background. Further, at 785 nm excitation, the fluorescence background was relatively strong. The incident excitation energy at 785 nm was consumed, to a great extent, by the resonance absorption and fluorescence, hence the interaction with the components of interest was weak. Therefore, the Raman peaks of such components did not show up distinctively in the recorded spectra.

Owing to the reasons stated above, it was therefore difficult to characterize and analyze the constituent molecules in fruits by Raman technique using the classical/univariate approach. Several research papers report on how to tackle Raman spectra of samples with such characteristics as highlighted in this work (Wei *et al.*, 2015, Liu and Liu, 2011). Some studies propose hardware changes, whereas others propose mathematical software computations to address these challenges. For instance, fluorescence can be overcome by using lasers with high excitation wavelengths such as 1064 nm (Zhang *et al.*, 2006b). However, in this study we were limited to a 785 nm laser. Thus the mathematical software approach was adopted for this work in the pre-processing steps.

Whereas the classical approach for analysis of LRS data could not perform exploratory analysis of LRS data in this study in a satisfactory manner, it has been shown that the use of multivariate chemometric techniques can offer solutions to this limitation. PCA was applied successfully for the exploratory analysis of the complex, multidimensional LRS data in this work. Apart from the information deduced about the peaks that were responsible for the different clusters, new set of fewer variables (PCs) were obtained to be used in the subsequent ML steps.

The results obtained from exploratory analysis were in agreement with the results collected from the elemental analysis of CaC_2 by the EDXRF spectrometer (Appendix 4). Among others, elements of phosphorous, calcium and sulfur were found to be present in the industrial carbide. As discussed in section 1.1.1, reactions between calcium and phosphorus produces calcium phosphide which in turn liberates phosphine in the presence of water. Using LRS (a molecular method) in conjunction with PCA, the presence of phosphine in the carbide ripened bananas was verified.

5.5 Qualitative Analysis (Classification) of Raman Spectra Using Machine Learning.

One of the main goals of this work was to establish whether LRS coupled with machine learning can be able to be used for assessing artificial ripeners in fruits qualitatively. To this end, selected ML classification models were developed form the pre-processed Raman spectra and the PCs obtained in section 5.4 above were used as inputs. Supervised ML techniques have the potential to reveal any hidden properties and they can learn selectively even in the presence of noise. This feature is particularly important as the classification model envisaged should be able to classify distinctly samples ripened naturally versus artificially using slight differences in their Raman spectra. The samples were split into a training set (27 samples) and a test set (14 samples).

5.5.1 Classification of Naturally and Carbide Ripened Samples Utilizing Support Vector Machine Classifier

In this sub-section, SVM was first used to visualize the decision boundary between banana samples which were ripened naturally and samples ripened artificially using the latent variables obtained in section 5.4 above. Some of the critical parameters that affect the output of an SVM classifier include the kernel type, the cost and the gamma functions. The cost factor controls the degree of violation of the margin such that a small cost factor results to widening of the margin to accommodate more support vectors and vice versa (Amarappa et al., 2014) The gamma function gives a measure of similarity between two given points. The SVM classifier was tuned using these parameters through 10-fold cross validation and the best parameter combination with the lowest error retained. The radial basis kernel function provided the most flexible separating hyperplanes and was found to be most appropriate for Raman data such as in this work. The SVM classification plot for the test data is shown in Figure 5.15 below.

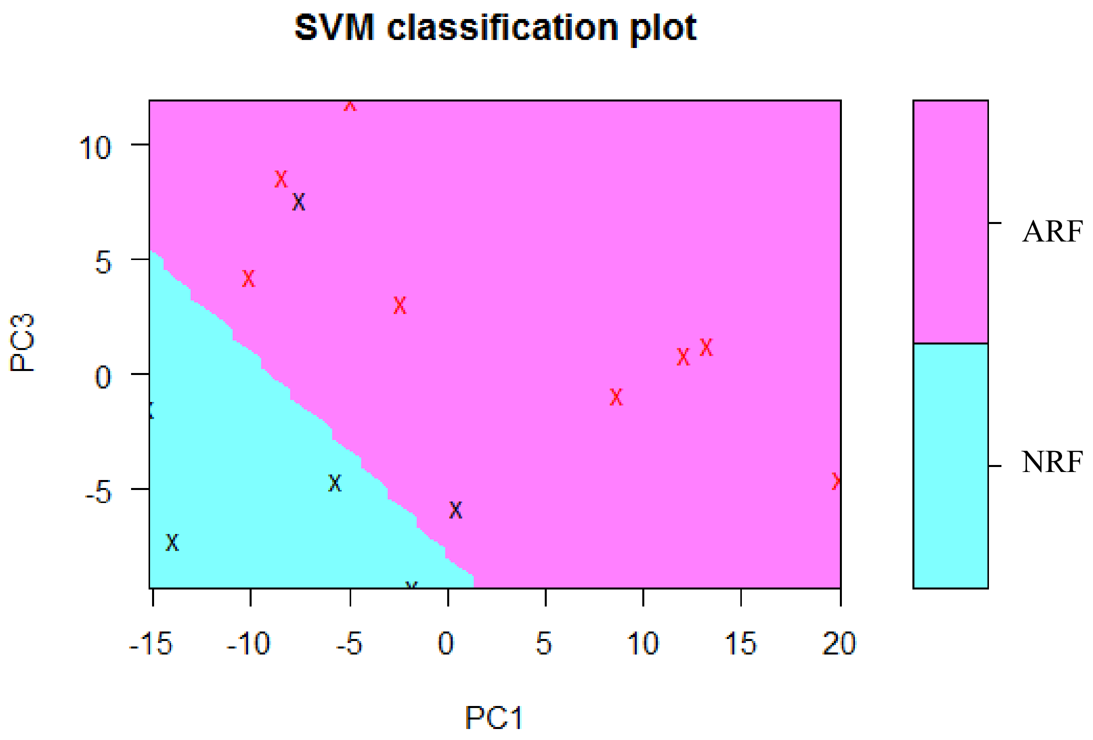

Figure 5.15: SVM classification plot utilizing PCs as inputs with RBK function

For test samples data hidden from the model, the confusion matrix in Table 5.2 shows that all samples ripened either artificially or naturally were correctly classified as belonging to their respective class. There was no misclassification and therefore correct grouping accuracy was 100 %.

Table 5.2: Confusion matrix of test data set

Classification ability		Predicted class	
		NRF	ARF
Actual class	NRF	6	0
	ARF	0	8
Classification accuracy (%)	100		

5.5.2 Classification of Naturally and Carbide Ripened Samples Utilizing Random Forest Classifier

Multiple classification decision trees were grown as discussed in section 4.8.3 and the average result from the prediction results were used to classify the input samples as either naturally or artificially ripened. Two key tunable parameters which were optimized included the number of trees (ntree) and number of variables tried at each split (mtry). These range of parameters were tuned during the training phase of model development and the outputs cross validated using the OOB data. The lowest OOB estimate of error rate achieved was 3.33 % and the corresponding model was adopted and used to predict the classes of the test data. Further, variable importance was evaluated and it was found that PC1 followed by PC5 were mostly responsible for the classification of the two groups of samples as shown in Figure 5.16. This is consistent with the discussion in section 3.6.4. A variable that gives a good split at the nodes is one that lowers the value of Gini impurity and entropy to the lowest possible values (Ayyadevara, 2018).

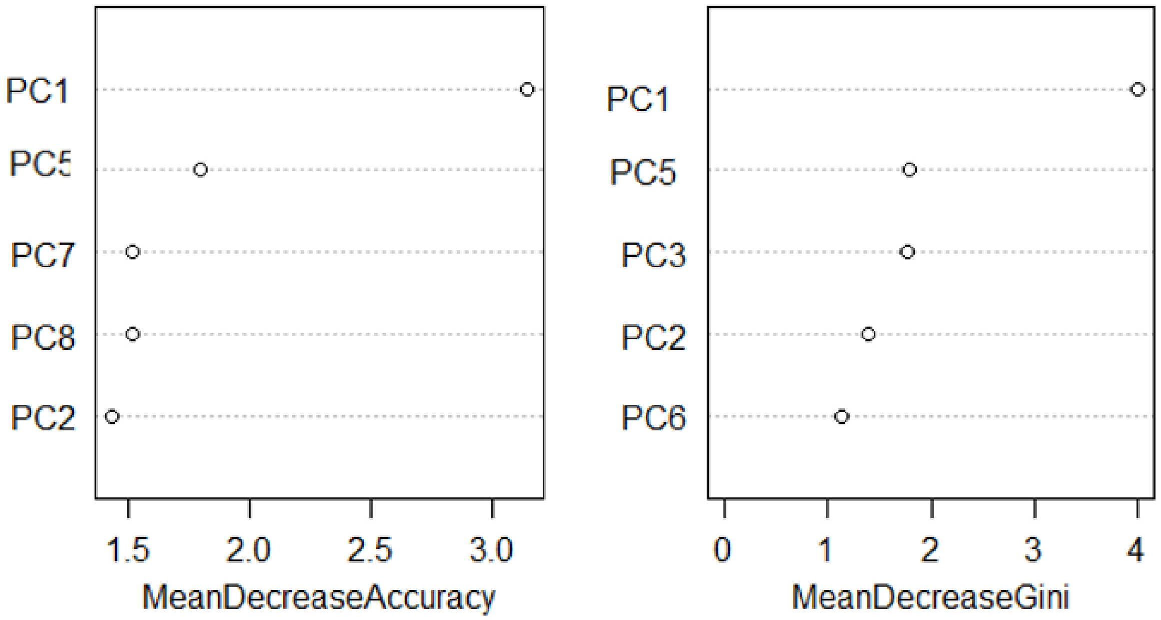

Figure 5.16: RF top 5 variable importance plot. The variables that had the most influence in the decision nodes to classify samples into different classes.

Referring to the loadings and score plots in section 5.4, PC1 and PC5 are linear combination of variables (Raman band regions) centered at 480 cm^{-1}, 612cm^{-1}, 780 cm^{-1} and 979 cm^{-1}. These are Raman vibrational bands associated with compounds resulting from artificial ripening using CaC_2. Thus, the RF classifier model classified the samples as either artificially or naturally based on the presence or lack of those compounds respectively.

For test samples data hidden from the model, the confusion matrix in Table 5.3 shows that one sample of NRF and another one sample of ARF were wrongly classified. An overall correct grouping accuracy of 85.71 % was realized in the RF classifier.

Table 5.3: Confusion matrix of test data set (ntree = 12, mtry=6)

Classification ability		Predicted class	
		NRF	ARF
Actual class	NRF	5	1
	ARF	1	7
Classification accuracy (%)	85.71		

5.5.3 Classification of Naturally and Carbide Ripened Samples Utilizing Artificial Neural Network Classifier

The ability of LRS coupled with the use of ANN for distinguishing fruits ripened naturally versus artificially was assessed using the first 20 PCs from the Raman datasets obtained in this work. The idea behind using PCs instead of the whole wavelength as inputs is to have a smaller subset of inputs, thus reducing significantly the computational burden and making ANN models converge faster.

Several parameters were tuned in the ANN classifier. To begin with, the ANN was created with 18 total neurons in the hidden layers: 10 neurons in the first hidden layer, 5 and 3 neurons in the second and third hidden layers respectively. The network training times and error rates were most acceptable with those settings. This agrees with earlier discussions in sections 3.6.3 and 4.8.2 that more than one layer allows the model to assess different levels of detail in the data. The ANN was able to learn the key spectral features amidst the background noise. The logistic transfer function was used for the hidden layers as well as for the output layer. Resilient backpropagation with weight backtracking algorithm was used to update the network weights. These range of parameters were optimized through random search and training stopped when the network realized low errors.

The trained ANN classifier model was then applied to test samples data hidden from it. The confusion matrix in Table 5.4 shows that only one sample of NRF was wrongly classified as belonging to the ARF group whereas all ARF samples were correctly classified. An overall correct grouping accuracy of 92.86 % was realized in the ANN classifier.

Table 5.4: Confusion matrix of test data set

Classification ability		Predicted class	
		NRF	ARF
Actual class	NRF	5	1
	ARF	0	8
Classification accuracy (%)	92.86		

In all the three classification models, important spectral differences between samples ripened naturally versus artificially are associated with the presence of sulfur (225, 480 cm^{-1}), acetylene (612 cm^{-1}), phosphine (979, 1115 cm^{-1}) and calcium hydroxide (780 cm^{-1}) molecules in the ARF group. The ML models were able to learn these spectral features and use them in predicting the classes of test data not exposed to them with high correct classification accuracies (>85 %). In this regard, SVM classifier provided the best results followed by the ANN classifier and then the RF classifier model.

5.6 Multivariate Calibration Using Machine Learning for Quantitative Analysis

In this section multivariate modeling involving ANN, RF and SVM are discussed. This approach has proven to be reliable with data sets that are known or suspected to be non-linear such as in this work. The univariate approach for calibration of such data is insufficient (Zhang *et al.*, 2014). This is true especially when trace quantities of the molecule of interest are involved. In this case, it becomes difficult to achieve a linear relationship between the emitted intensities of the spectral lines and concentration. Such linear relationships work well for high concentrations and fail for very low concentrations. Moreover, achieving a linear relationship of the analyte of interest with respect to the Raman signal becomes difficult in the presence of interfering molecules.

Multivariate calibration takes into account multiple instrumental data for a single sample. Herein, the use of inverse regression models on latent variables allows for quantification of analytes of interest in a sample without knowing the chemical identity of the interfering molecules. The presence of the latter is adequately compensated for by the calibration model which is built from a training set where the interfering agents have been incorporated (Lee *et al.*, 2013). This is important in food safety analyses such as in the current work where the number of interfering species is unknown. The constituent signals are proportional to their concentration.

ANN, RF and SVM were utilized in multivariate calibration and prediction of the concentration of artificial ripening compounds in banana samples as described in section 4.8. Exploratory analysis was done as a first step whereby the ARF samples were sub-divided into groups of low (0.024-0.06%), medium (0.08-0.2%) and high (0.4-1.6%) concentrations. Figure 5.17 below shows the scores plot.

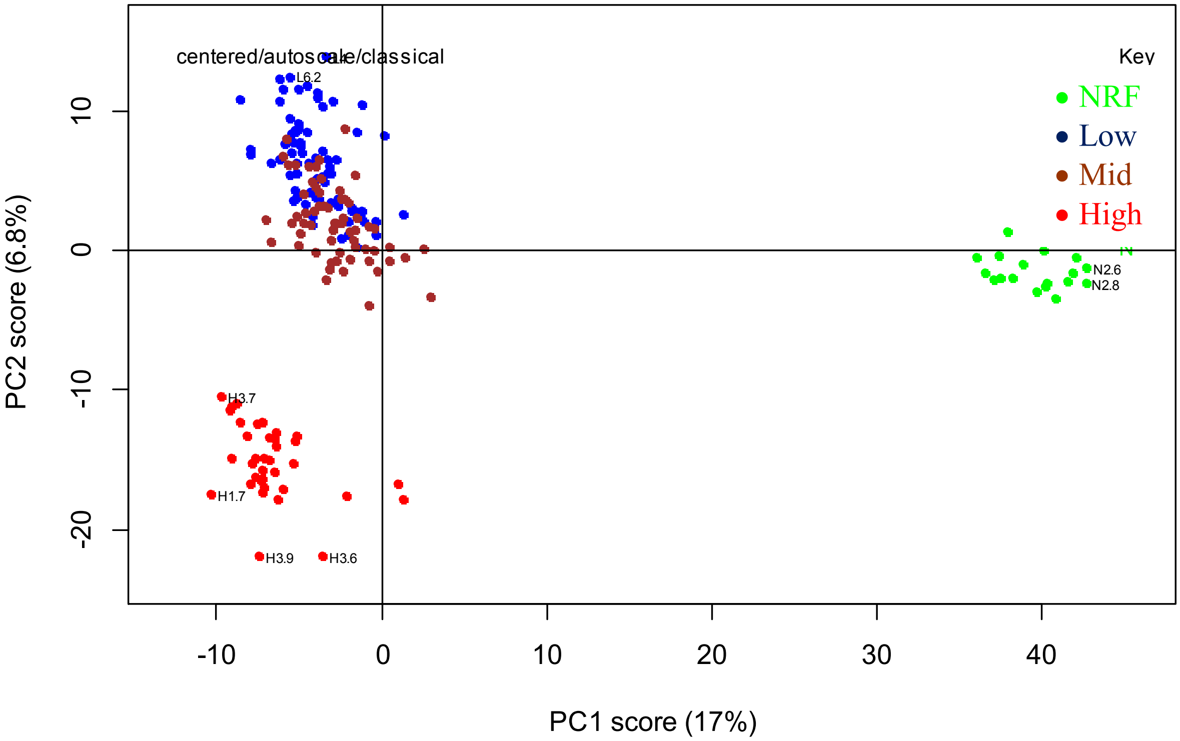

Figure 5.17: PCA scores plot of naturally (NRF) and artificially ripened samples showing clusters based on concentration of CaC_2. The NRF samples cluster on their own on positive side of PC1. The carbide ripened samples cluster on the negative side of PC1 and also according to the level of concentration used in ripening beginning from low to medium and high concentrations.

It can be seen from the scores plot in Figure 5.17 above that the samples can be distinguished as either naturally or carbide ripened by PC1 (horizontal axis). PC2 (vertical axis) and the subsequent PCs show the concentration profile of the ARF samples. Having successfully reduced the dimensionality of the LRS data using PCA, the PCs were used as inputs in developing the ML regression models.

5.6.1 Quantitative Analysis of Carbide in Samples Using Artificial Neural Network

The merits for using ANN for multivariate calibration have been published over time (Fan *et al.*, 2019; Allouche *et al.*, 2015). Of importance in the present work, includes the capability of ANNs to be able to learn from input-output target examples amidst an environment coupled with noise. Additionally, ANNs are capable of modelling non-linear relationships between the input and target response variables. These make up some of the reasons why ANN was preferred for use in calibration.

Principal components were used as the input to the ANN which was implemented in R software. In particular, the first 20 PCs were used meaning the ANN had 20 neurons at the input layer. The corresponding known concentration were administered as targets in the network. Prior to running the network, the input matrix was scaled using the min-max function which transforms the inputs to have mean of zero and a standard deviation of one. This ensured that the input variables participated equally in the modelling process. The data was split into training and test data in the ratio 4:1 respectively. This resulted to a training set of 123 samples whereas the test samples were 31.

Optimization for the best combination of parameters was done. This included a variation in the number of hidden nodes and neurons as well as the transfer function. The number of nodes and hidden layer were determined after trying various network structures as presently, no theory gives an exact number of neurons and hidden layers required to approximate a given function. A trial and error approach was therefore used following recommendations by Ciaburro and Venkateswaran (2017) that the neurons in the hidden layer should be an average of the neurons in the input and the output layer. The number of hidden layers was varied with other parameters held constant such that the number of layers that produced the lowest RMSEP and highest prediction accuracy was adopted. Thereafter, the number of neurons was varied as the number of hidden layers was kept constant and this was also evaluated against RMSEP and R^2 values. The outcomes are as summarized in Table 5.5 and Table 5.6 below:

Table 5.5: Model prediction ability with a variation of hidden layers

Hidden Layers	**Number of neurons**	**Transfer function**	**RMSEP** (g/L)	**R^2**
1	12	logistic	0.656	0.9369
2	12	logistic	0.331	0.9831
3	12	logistic	0.466	0.9686
4	12	logistic	0.273	0.9890
5	12	logistic	0.597	0.9477
6	12	logistic	0.597	0.9477

Table 5.6: Model prediction ability with variation of the number of hidden neurons

Hidden Layers	**Number of neurons**	**Transfer function**	**RMSEP** (g/L)	**R^2**
2	8	logistic	0.363	0.9807
2	10	logistic	0.316	0.9854
2	12	logistic	0.331	0.9831
2	14	logistic	0.416	0.9747
2	16	logistic	0.518	0.9606
3	6	logistic	0.597	0.9477
3	9	logistic	0.483	0.9658
3	12	logistic	0.466	0.9686
3	15	logistic	0.269	0.9893
3	17	logistic	0.311	0.9859
4	8	logistic	0.363	0.9807
4	10	logistic	0.316	0.9853
4	12	logistic	0.273	0.9890
4	16	logistic	0.355	0.9815

It was found that a three hidden layered network with fifteen hidden neurons and logistic transfer function resulted into the rapid convergence of the network. This is consistent with literature discussions that a model with only one or two hidden nodes is strongly biased and therefore limited in the number of functions that it can fit (Ciaburro and Venkateswaran, 2017). On the other hand, a model which is ideally unbiased (having infinite hidden neurons) tends to overfit to the training set and could only work well with noise-free data. Having significantly many hidden neurons was not conducive for the current study as the Raman spectra for the molecules of interest was recorded in an environment of interfering molecules. A balance was therefore achieved by using three hidden layers having fifteen hidden neurons. Other transfer functions like tansig resulted in models with high error rates In addition, highest R^2 and the lowest RMSEP values for the training and test data sets were achieved under these conditions. As such, these parameters were adopted and utilized for subsequent modeling.

The type of backpropagation rule under which these were realized was gradient descent algorithm. The output of the layers was calculated from the net input by means of the logistic transfer function. Consequently, differences between the output and expected values from the ANN's input were interpreted as errors in the network and were back-propagated through the layers until the network converged to the lowest acceptable errors.

The optimized network was then stored before the next step of analyzing the predictive capabilities of the BP-ANN. The basis of this procedure was to demonstrate the prediction performance of the developed models in predicting the concentration of artificial ripeners in banana samples ripened using CaC_2. To this end, the testing set which was originally hidden from the network was exploited to evaluate the confidence in the performance of the trained network.

The obtained results were graphically plotted showing comparison of predictions through ANN analysis method. Figure 5.18 shows predicted concentrations of CaC_2 compounds used in ripening banana samples. The predictions on Figure 5.18 are based on data from the testing set implemented to samples that were not in the training set. The figure clearly show that experimentally measured concentrations of the ripening agent are in strong consistency with the values predicted through ANN for most of the samples.

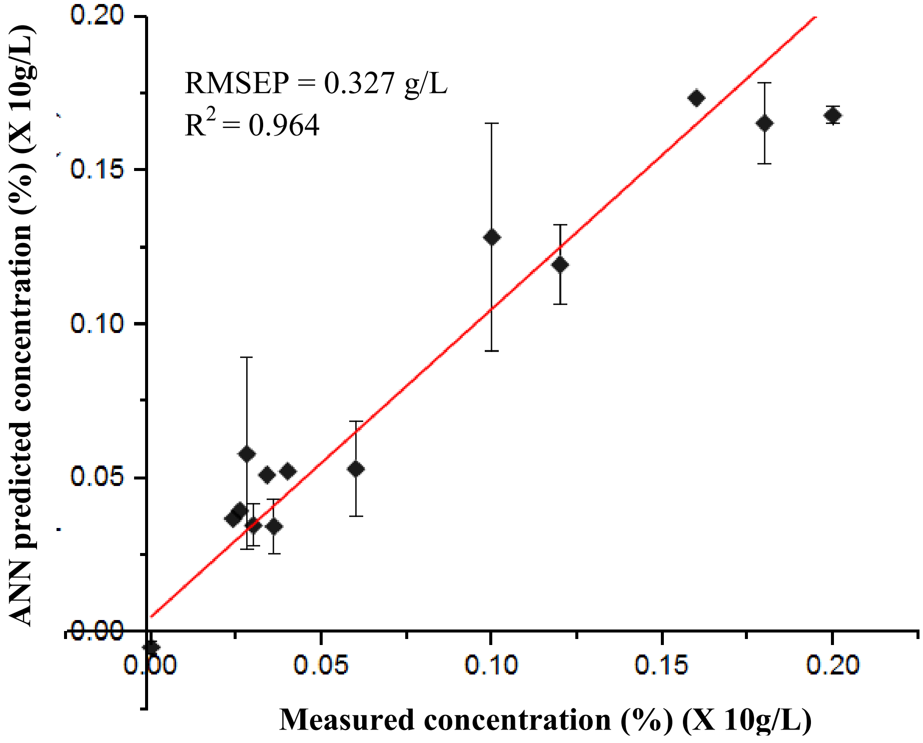

Figure 5.18: ANN regression plot for test data set. The error bars were obtained as standard deviations associated with spectra recorded at different spots on the same sample (concentration).

In evaluating the multivariate LOD and LOQ for the ANN model, a linear fit function was applied to the calibration curve and the values of slope and standard deviation calculated as discussed in section 4.9. Figure 5.19 below shows the linear fit from the calibration curve and the calculated values.

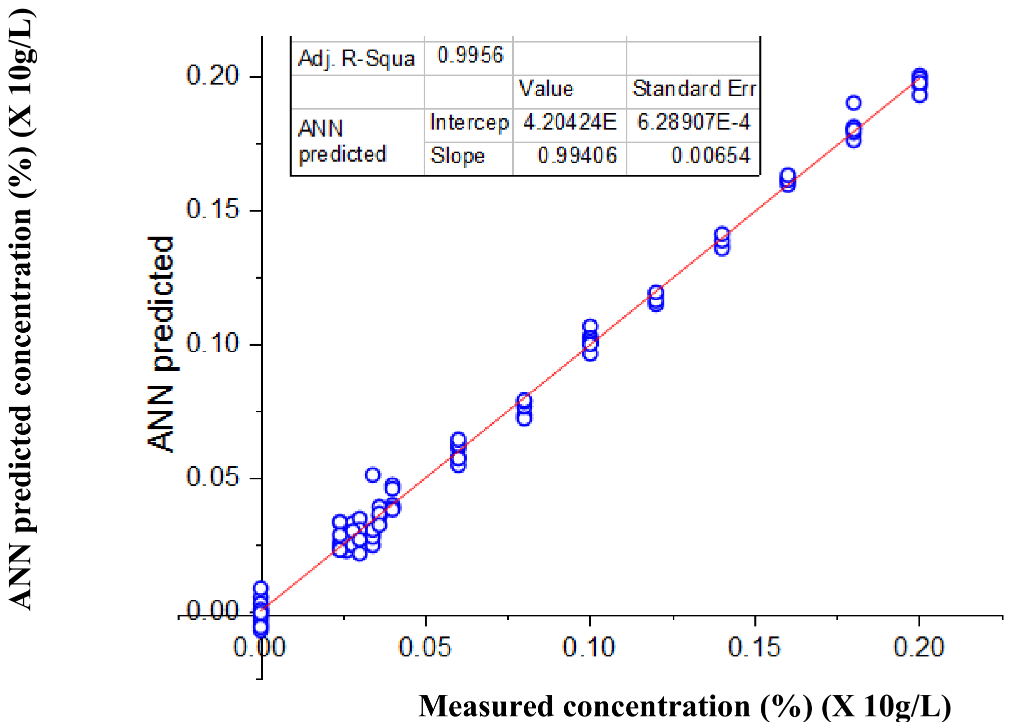

Figure 5.19: ANN calibration plot for calculating LOD and LOQ

The computed values for LOD and LOQ were 0.00189 % (0.0189g/L) and 0.00633 % (0.0633g/L) respectively for this method.

5.6.2 Quantitative Analysis of Carbide in Samples Using Random Forest Regression

The use of RF in multivariate calibration for quantitative analyses has been reported in several studies (Lee *et al.*, 2013; Zhang *et al.*, 2014; Ayyadevara, 2018). Many researchers have proved that the method has a good tolerance to noise, it is not heavily affected by non-linearity in data and avoids the problem of overfitting. For these and other reasons like being able to give valuable information about variable importance, RF was employed for multivariate calibration in this work.

Optimization of the RF regression models involved tuning two key parameters: no of trees (ntree) and number of variables tried at each split (mtry). The optimum number of trees for predicting concentration of the CaC_2 used in ripening banana samples was evaluated from the OOB error estimates. An increase in the number of decision trees was reported to decrease the OOB error up to a certain limit where further increase did not have much effect on the model. Figure 5.20 below shows that beyond 230 trees in the RF model, there was no change in the OOB error. On the other hand, when the value of ntree was too low, the OOB error estimate increased significantly. This supports the idea of having several independent trees to improve the quantitative accuracy of a RF model. Further, the randomness in the forest was achieved by the bagging technique discussed in section 3.6.4. This ensured that for each grown tree, different samples were picked up for training and on the final result, the vote from each of these trees was averaged to obtain the final values. This helps overcome the challenges of bias in predicting values and so avoiding overfitting (Lee *et al.*, 2013).

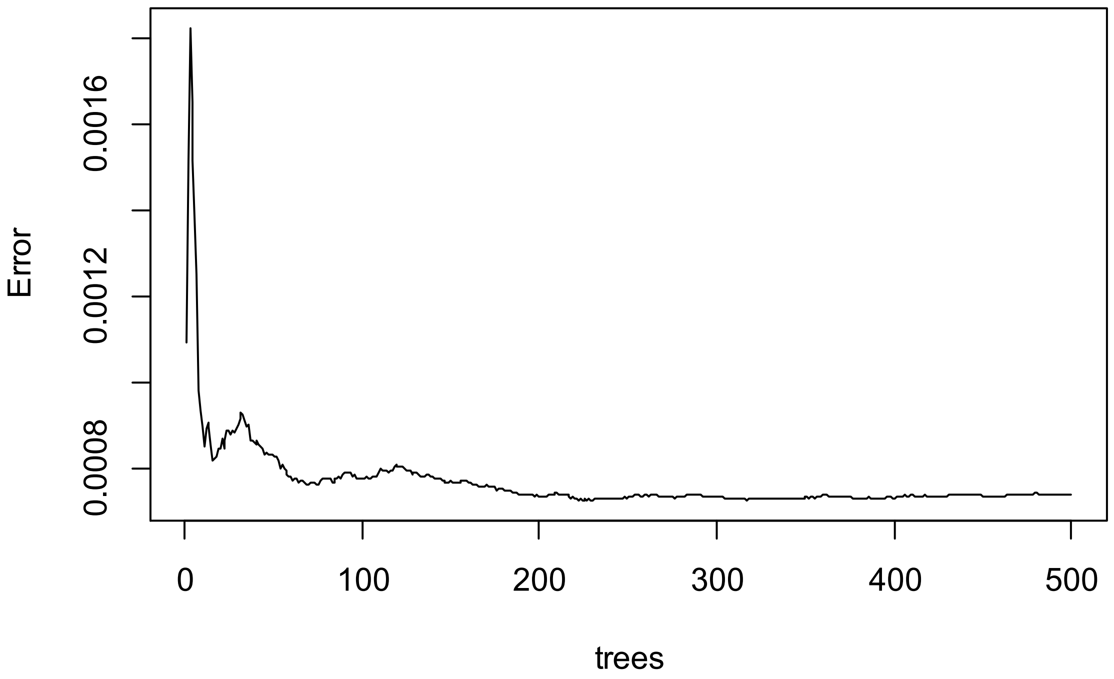

Figure 5.20: RF variation of OOB error against number of trees in the forest. The lowest OOB error rate was achieved at 230 trees. This was the optimum number of trees for the model

Input variables form a critical component of RF regression models. Although the cleaned Raman spectra in section 5.3 brought out rich spectral information concerning the analytes of interest, there was a lot of interfering information from other molecules. The advantages of using PCs as inputs in such a case have been discussed previously. However, there was need to establish the number and the importance of the PCs used as input variables in developing the RF predictive models. An evaluation of the input variables showed that the first 20 PCs were able to explain 86% of the variance in the training data in the RF model. In the decision nodes, it was established that 16 variables were optimum to be tried at each split to give the best predictive accuracy. Figure 5.21 below shows the variation of OOB estimate error with the mtry parameter. When mtry becomes too small, the variables considered at each split are insufficient resulting to diminished predictive accuracy.

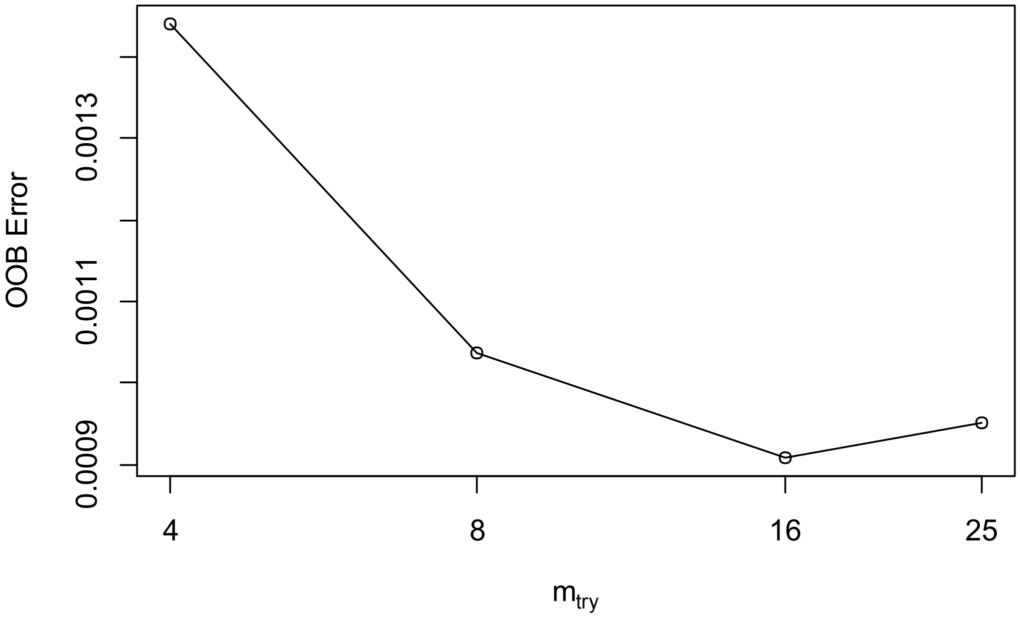

Figure 5.21: RF variation of OOB error against number of variables tried at each split. The lowest OOB error rate was achieved when the variables tried at each decision node was 16.

The variables used in the decision nodes for prediction in this model were PCs which arose from the Raman shift of the molecules present in CaC_2 ripened samples. As discussed in section 5.5.2, variable importance is an important feature which can be evaluated in this RF algorithm. The variables which lowered the Gini impurity the most were PC2 followed by PC1 as shown in Figure 5.22 below. These two variables had the greatest impact on the decision nodes for predicting the concentration values.

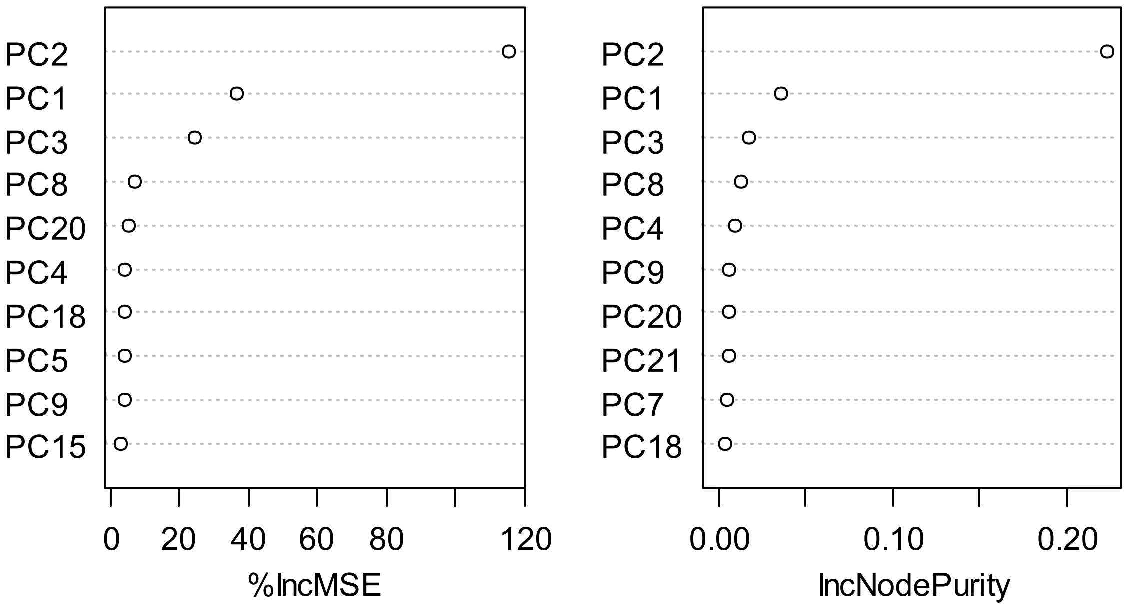

Figure 5.22: RF regression top 10 variable importance plot

A score plot of PC2 against PC1 (see Figure 5.17) depicts the concentration profile and reveals that these PCs give vital information necessary for predicting the concentration of CaC_2 used at different levels. In comparison with the loadings plot in section 5.4, the regions associated with these PCs are Raman bands centered at 480 cm^{-1}, 612 cm^{-1}, 780 $cm^{-1}$and 979 cm^{-1}. Thus, the RF model was able to predict different concentrations based on the present amount of the molecules represented by the listed peaks.

Once the two key parameters were properly tuned as described in the preceding paragraph; such that the lowest OOB error estimates were obtained, the trained RF model was used to predict concentration values for the test set. The calibration and validation plots for the predicted against the actual concentration are shown in Figure 5.23 below. The R^2 and RMSEP values were calculated and show a strong correlation between the actual measured concentrations and the RF model estimates.

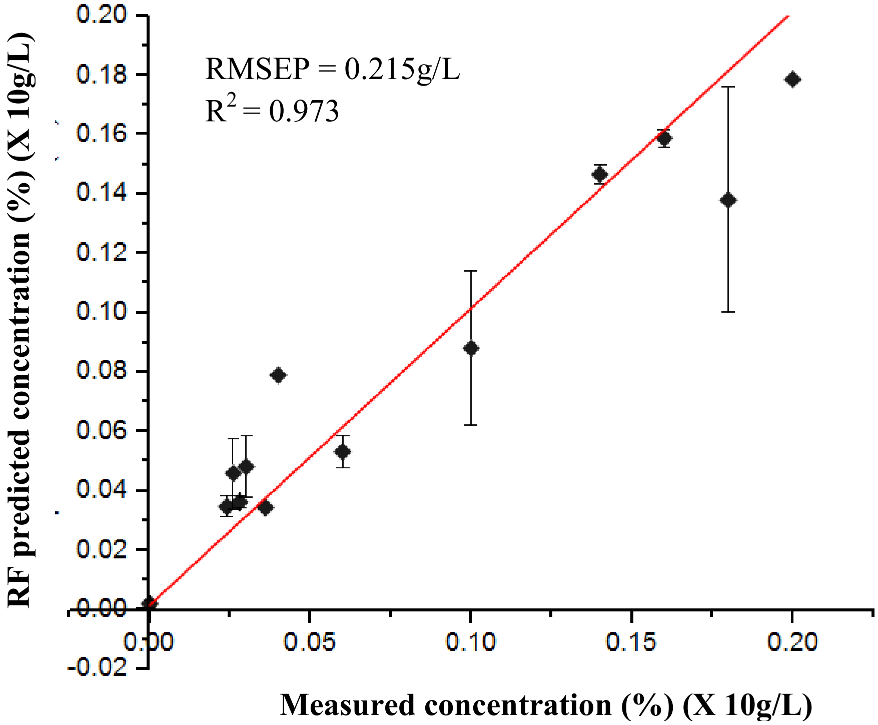

Figure 5.23: RF regression plot for test data set. The error bars were obtained as standard deviations associated with spectra recorded at different spots on the same sample (concentration)

Detection and quantification limits were calculated using the 3σ and 10σ approach discussed in section 4.9 using the pseudounivariate calibration curve. The RF concentrations were plotted against the actual concentrations and from the linear fit, of slope and standard deviation of the y-intercept were calculated. The plot is as shown in

Figure 5.24 below. The calculated values of LOD and LOQ were 0.00539 % (0.0539g/L) and 0.01796 % (0.1796g/L) respectively.

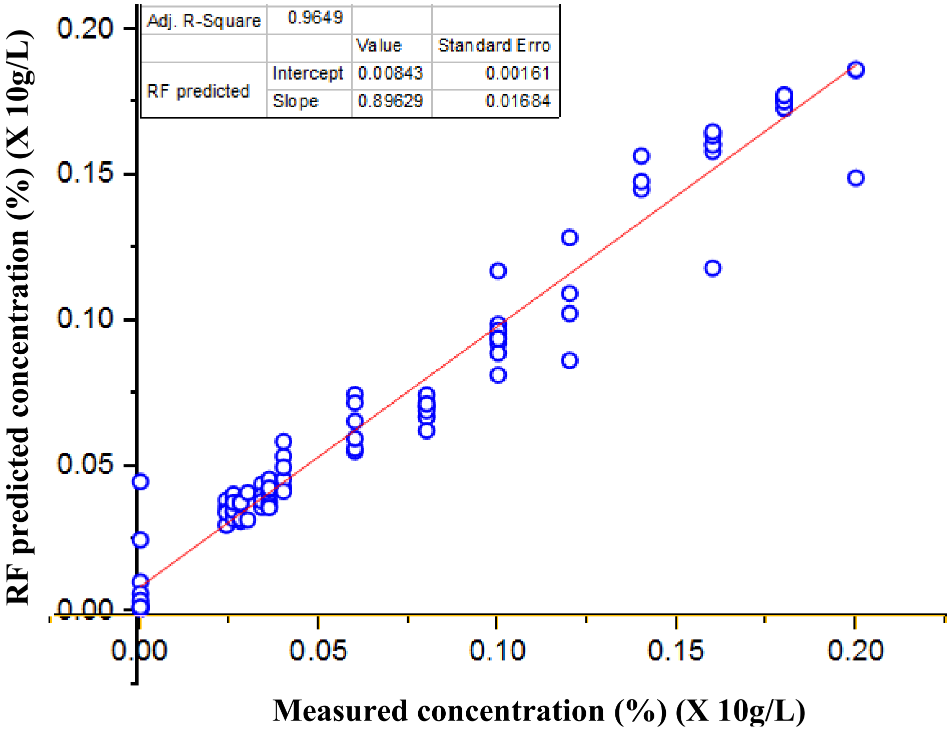

Figure 5.24: RF calibration plot for calculating LOD and LOQ

In order to increase the accuracy of RF model, the inclusion of samples with narrower concentration intervals in a calibration set is necessary (Lee *et al.*, 2013).

5.6.3 Quantitative Analysis of Carbide in Samples Using Support Vector Regression

The use of SVR for multivariate calibration and prediction of concentration of CaC_2 used to ripen banana samples is discussed in this sub-section. Several parameters were optimized to have the final model with low RMSEP values and relatively high R^2 values.

To begin with, the optimal number of input variables (PCs) was evaluated. The PCs were set as the predictors whereas the different concentration used in ripening were set as the target response values. Using all PCs as input variables resulted to calibration curves that overfitted to the training data such that SVM models in this case generalized poorly to the test set. The number of input variables (PCs) was reduced gradually up to the first 6 PCs whereby the SVM model performance improved for both the training and test data. The input PC scores determined the number of support vectors (data points which affect the decision surface). The inclusion of the least dominant PCs resulted to an increased number of support vectors leading to overfitting of the training data. The use of the first 6 PCs (explaining 50 % variability) with 96 support vectors was found to be optimal for this study.

The choice of kernel function is an important consideration in developing SVR regression models. Consequently, the radial basis kernel (RBF) had a better performance compared to the linear kernel function for the data in this work. This is consistent with literature discussions (section 3.6.5); that non-linear data sets work well with non-linear mapping functions in the SVR models (Balabin and Lomakina, 2011). Finally, the cost and the gamma were optimized by running the model through a 10-fold sampling cross validation and evaluating the model error. Figure 5.25 below shows that a cost value of 1 provided the least model error and therefore provided optimum degree of violation of margin to accommodate the support vectors.

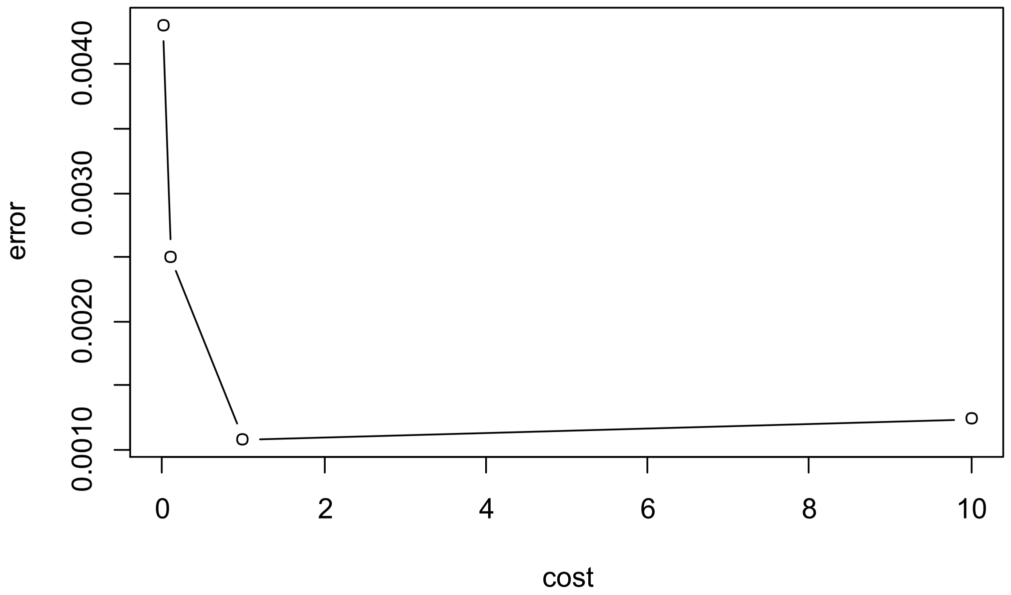

Figure 5.25: variation of the SVR model error with the cost function

The performance of the trained SVR model was evaluated by a test set not previously shown to the model. The SVM predicted estimates were plotted against the actual experimentally measured concentrations as shown in Figure 5.26 below. The R^2 and

RMSEP values were calculated and show a strong correlation between the actual measured concentrations and the SVR model estimates.

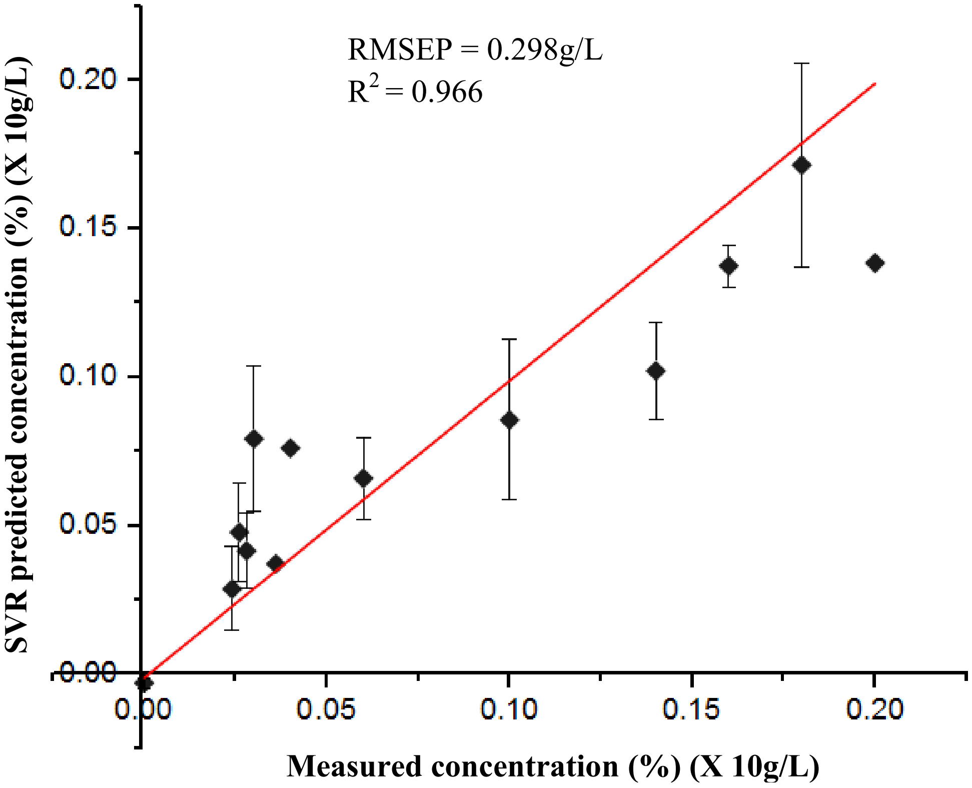

Figure 5.26: SVR plot for test data set. The error bars were obtained as standard deviations associated with spectra recorded at different spots on the same sample (concentration)

In determining the multivariate LOD and LOQ for the SVR model, a pseudounivariate graph was drawn as described in section 4.9. Using the 3σ and 10σ approach in equations (4.3) and (4.4) the LOD was found to be 0.008741% (0.0874g/L) while the LOQ was 0.0291% (0.291 g/L).

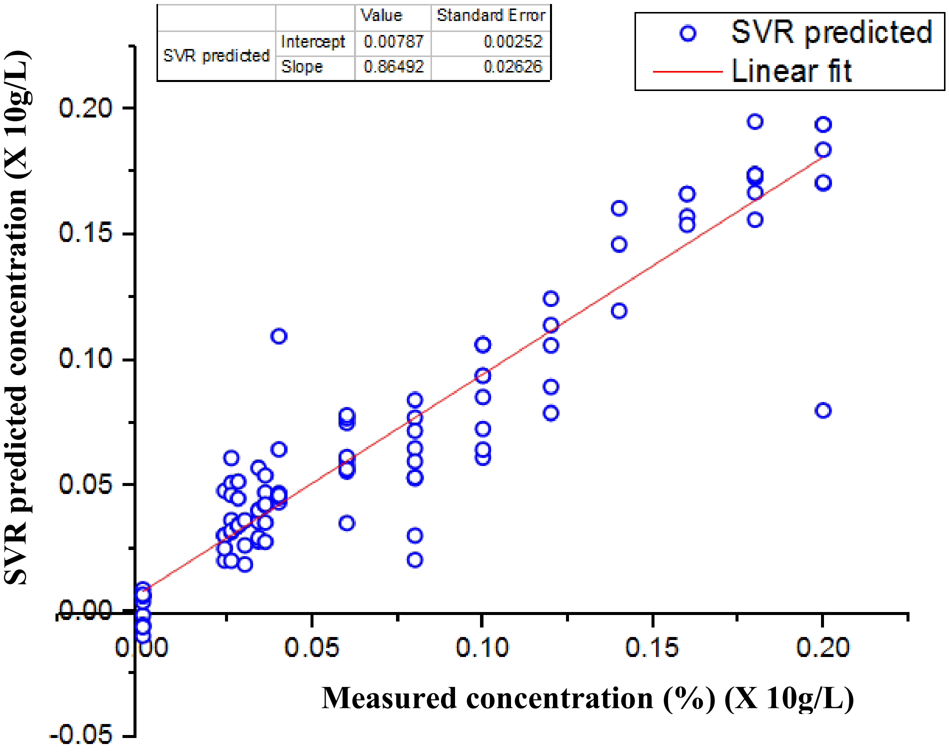

Figure 5.27: SVR calibration plot for calculating LOD and LOQ

In order to validate the predictive ability of the developed multivariate calibration models in test data sets initially hidden to the models, we compared the result of ANN with RF regression and SVR by means of RMSEP and R^2. In regression models, RMSEP gives a robust measure of how accurately the model predicts the target response variable. Table 5.7 shows the comparison of the performance of the three models used in this work based on these parameters. The model predicted values are contained in Appendix 4. Notably, all the three models used in this study provided nearly same indications. This justifies the use of the three models since they all have different architectures but they were able to make similar predictions for the same inputs.

Table 5.7: Model performance based on explained variance (R^2), root mean square error of prediction (RMSEP) and multivariate LOD and LOQ

Model	R^2	RMSEP (g/L)	LOD (g/L)	LOQ (g/L)
ANN	0.964	0.327	0.0189	0.0633
RF	0.973	0.215	0.0539	0.1796
SVR	0.966	0.298	0.0874	0.2913

The method developed was envisaged to be a rapid method. Once the Raman measurements were optimized and the machine learning models trained, tested and validated, a quick evaluation was done to establish how rapid the method was. In this regard, the time required to assess whether a sample had been ripened by calcium carbide and the extent thereof i.e. from sample preparation to data analysis was as summarized in the table below:

Table 5.8: Results turn-around time: from sample preparation to data analysis

Stage		Activity	Time (minutes)
1	Sample preparation	Washing, drying, slicing	<5
2	Raman spectroscopy	Spectra data acquisition (10 S exposure time, 5 accumulations)	<2
3	Data analysis	Preprocessing, classification, quantification	<10
Total time per sample			<17 minutes

In comparison with other techniques such as HPLC and ICP-AES which have average results time of 12-48 hours and 0.5-12 hours respectively for these kind of measurements (Cramer *et al.*, 2017), the method developed in this work is better suited.

5.7 Prediction of Calcium Carbide Concentration used in Ripening Market Samples

After exploration of data using PCA, testing and validating the developed quantitative analytical models, data from market samples was input into the models to predict, if present, the concentration of CaC_2 used in ripening them. Raman spectra obtained from banana samples collected from Nairobi open air markets and its environs were preprocessed as described earlier in section 4.6. The ANN, RF and SVR models developed in sections 5.6.1, 5.6.2 and 5.6.3 were then used to predict CaC_2 concentrations used to ripen the samples. The results of these predictions were averaged and summarized as shown in Table 5.9 which indicates mean concentration per sampling market together with the standard error of the mean (SEM).

Table 5.9: Market samples predictions

Sampling market	Model predicted concentrations of CaC_2 used in ripening (g/L)					
(8 samples per market)	ANN		RF		SVR	
	mean	SEM	mean	SEM	Mean	SEM
CHIROMO	1.204	0.198	1.130	0.207	1.083	0.200
RONGAI	1.066	0.076	1.157	0.084	1.030	0.074
GIKOMBA	0.819	0.069	0.667	0.059	0.660	0.062
MARIKITI	0.795	0.104	0.786	0.107	0.779	0.088
KISERIAN	0.267	0.028	0.304	0.052	0.312	0.039

Table 5.9 shows that banana samples from Chiromo and Rongai area have been ripened using CaC_2 concentrations of up to >1g/L. These samples were suspected to be artificially ripened from their outlook. Previous studies have shown that bananas ripened artificially tend to have an attractive bright yellow, spotless colour with green or yellow stalks. On the other hand, naturally ripened bananas present an unattractive, light yellow colour with black spots as well as blackish yellow stalks (Akter *et al.*, 2020). A side by side comparison of a banana from Kiserian and Chiromo is as shown in Figure 5.28 below.

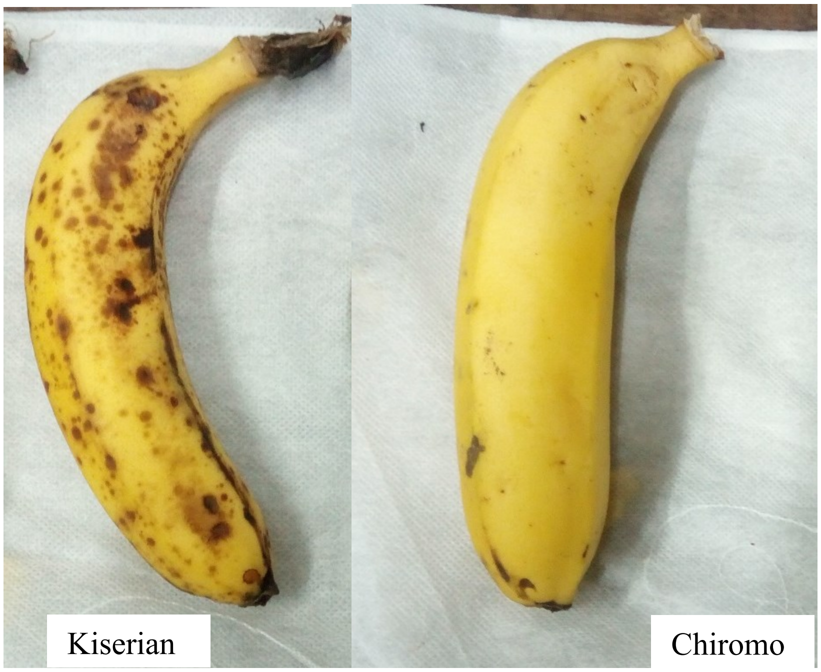

Figure 5.28: A visual comparison of suspected artificially ripened banana and naturally ripened banana from local markets. Carbide ripened bananas usually have a spotless, bright yellow appearance whereas naturally ripened bananas normally present an unattractive yellow color with black spots.

The samples from Kiserian market (left) in Figure 5.28 are indicative that they might have ripened naturally. Indeed this is as evidenced by the relatively low mean concentrations of CaC_2 as predicted by the ML models (0.267 g/L – 0.312 g/L). The ML models predictions for most samples in this market had concentrations <0.150 g/L which was close to the quantification limits of these ML models. This leads us to the conclusion that these samples were ripened naturally as most of the predictions of concentration from the ML models were below the LOQ; suggesting very low concentrations.

In general, the predicted concentrations from the ML models were characterized with high standard deviations as evidenced by the SEM values in Table 5.9. In some cases, the mean and the median concentration varied significantly. This can be attributed to the fact that different samples within a sampling region may have been exposed to different levels of the artificial ripener. Thus, the mean values in Table 5.9 are population mean for all samples within the specified market. The SEM metric is therefore used as a dispersion measure in this table as it gives an estimate of how the sample mean varies with respect to the population mean.

A comparison of the CaC_2 concentrations from the ML models predictions show similar trends for the sampled regions. This shows that LRS coupled with ML can be used to detect the presence and quantify artificial ripeners in fruits. The performance of the models to predict market samples was compared using the ANOVA technique.

Table 5.10: Anova statistics showing comparison between different model performance

Model comparison	t-test	P-test
RF - ANN	-0.099	0.995
SVR - ANN	-0.266	0.962
SVR - RF	-0.167	0.985

The results summarized in Table 5.10 indicated that all the three models were significant with regards to the market sample predictions and that none was better than the other. This was evidenced by the small variations in the mean values from the different models as shown in Table 5.9. The predictions from all models were comparable. Nonetheless, RF model had the least errors in its predicted values followed by the SVR model and then ANN model. In this regard, RF's performance was marginally better than ANN and SVR for this study.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusion

This study was undertaken to explore the possibility of the laser Raman spectroscopy method as an alternative for assessment of carbide ripened bananas owing to its advantages of being a fast and non-invasive technique.

The presence and quantity of CaC_2 in carbide ripened bananas was evaluated based on impurities that are always present in the industrial CaC_2 that was used in this study. Sulfur, acetylene, calcium hydroxide and phosphine, having Raman peaks centered at Raman shift regions 480 cm^{-1} (S-S bond stretching), 612 cm^{-1} (C-H asymmetric bending), 780 cm^{-1} (O-H bending) and 979 cm^{-1} (P-H stretching) respectively, are some of the impurities in CaC_2 that have been reported to have been found in carbide ripened fruits. Detection and quantification of these molecules from the Raman signal require quality and repeatable spectra. Consequently, optimization of LRS equipment and parameters for acquisition of spectral data need to be done appropriately to suit the study at hand. In this work, using a 785 nm laser delivering 6.28mW to the sample through a microscope objective of X50 over an exposure time of 10 seconds and 5 accumulations provided optimized conditions for spectra acquisition.

Herein, Raman data obtained from naturally and carbide ripened banana samples were analyzed using ML models; ANN, RF and SVM. PCA was utilized for exploratory data analysis and dimensionality reduction prior to qualitative and quantitative modeling. It has been shown that the multivariate approach for analysis of data such in this work is necessary as opposed to the univariate approach. The Raman spectra of the analyte of interest were buried in the background of interfering molecules. However, the application of several pre-processing steps and use of PCA (a multivariate ML approach) enhanced the spectral features that were otherwise not easily visible in the raw Raman spectra.

Moreover, it has been shown that ML aided LRS model is a superior alternative to the quantitative methods used in classical Raman quantification models which solely rely on the intensity of target peaks. It has been shown that the ML models are capable of modeling calibration schemes accurately from Raman spectra without properly defined peaks or even in the presence of noise as well as in cases where the spectral response is non-linear. Accurately tuned ML models were developed with the ability to classify and quantify banana samples based on how much CaC_2 was used to ripen them. Banana samples were classified as either artificially ripened or naturally ripened with high correct classification accuracies (> 85 %) using the ML classification models. Further, the quantitative models recorded low RMSEP values for the validation data sets as well as high R^2 values (> 90%). This showed that there was a high correlation between the known concentrations and the model predicted concentrations. The ML assisted LRS models also recorded low detection limits ($\leq 0.088\ g/L$) and quantification limits ($\leq 0.291\ g/L$).

In conclusion, coupling machine learning techniques (PCA, ANN, RF and SVM) with Raman spectroscopy eliminates the challenges commonly encountered when using classical Raman analysis techniques. Properly trained ML models can learn Raman spectral features of samples exposed to them and use them to make predictions on new data as was the case with the test and market samples data. This work shows that LRS coupled with ML models can be used to assess levels of artificial ripeners in fruits. Furthermore, the method can give test results in about 15 minutes, which makes it a rapid alternative to the conventional wet chemistry methods.

6.2 Recommendations and Future Prospects

Apart from artificial ripening of fruits, there are harmful practices of preserving already ripened fruits. The methodology presented in this work, using bananas and calcium carbide can be extended to several fruits that are prone to these practices that pose risks to food safety and human health. A comprehensive LRS spectral library consisting of fruits in their pure form as well as fruits that have artificial additives can be developed and used as a basis for evaluating their safety for consumption.

The major challenge encountered in this work with the use of 785 nm excitation laser was the broad fluorescence signal that was several orders above the requisite Raman signal. In as much as mathematical corrections can be used to correct this limitation, there is always the likelihood of loss of vital information alongside the subtracted baseline. The use of 1025 nm excitation or higher excitation wavelengths in the analysis of artificial ripeners in fruits is therefore recommeded. Zhang *et al.* (2006) demonstrated that at 1064 nm excitation, the fluorescence signal was suppressed and different fruits could be characterized by their Raman spectra at that excitation wavelength.

Lastly, this study was carried out in a lab using a confocal LRS system with controlled parameters. Use of a portable Raman micro-spectrometer system that is field is also recommended. The system should be embedded with programmable software from the developed ML models to enable direct and instantaneous assessment of artificial ripeners in fruit samples at source.

REFERENCES

Adeniji, T. ., Sanni, L. ., Barimalaa, I. ., and Hart, A. . (2010). Nutritional and anti-nutritional composition of flour made from plantain and banana hybrid pulp and peel mixture. *Niger. Food J.* **25**.

Afseth, N. K., Segtnan, V. H., Marquardt, B. J., and Wold, J. P. (2005). Raman and near-infrared spectroscopy for quantification of fat composition in a complex food model system. *Appl. Spectrosc.* **59**, 1324–1332.

Afseth, N. K., Segtnan, V. H., and Wold, J. P. (2006). Raman spectra of biological samples: A study of preprocessing methods. *Appl. Spectrosc.* **60**, 1358–1367.

Akter, B., Talukder, N., Bari, L., and Rabeta, M. S. (2020). Evaluation of ripening period, shelf-life, and physiological properties of Sobri (Musa cavendish) and Sagor (Musa oranta) bananas triggered by ethephon and calcium carbide. *Food Res.* **4**, 407–412.

Allouche, Y., López, E. F., Maza, G. B., and Márquez, A. J. (2015). Near infrared spectroscopy and artificial neural network to characterise olive fruit and oil online for process optimisation. *J. Near Infrared Spectrosc.* **23**, 111–121.

Andrej Krenker, Bešter, J., and Kos, A. (2011). Introduction to the Artificial Neural Networks, Artificial Neural Networks - Methodological Advances and Biomedical Applications. *Prof. Kenji Suzuki (Ed.).*

Ayyadevara, V. K. (2018). “Pro machine learning algorithms : a hands-on approach to implementing algorithms in Python and R.”

Balabin, R. M. and Lomakina, E. I. (2011). Support vector machine regression (SVR/LS-SVM) - An alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. *Analyst* **136**, 1703–1712.

Bari, L., Akter, B., Talukder, N., Islam, A., and Akter, D. (2018). Identification of different physiological properties of Sagor. **19**, 300–306.

Bhadoria, P., Nagar, M., Bharihoke, V., and Bhadoria, A. S. (2018). Ethephon, an organophosphorous, a Fruit and Vegetable Ripener: Has potential hepatotoxic effects? *J. Fam. Med. Prim. care* **7**, 179.

Butler, H. J., Ashton, L., Bird, B., Cinque, G., Curtis, K., Dorney, J., Esmonde-White, K., Fullwood, N. J., Gardner, B., Martin-Hirsch, P. L., Walsh, M. J., McAinsh, M. R., Stone, N., and Martin, F. L. (2016). Using Raman spectroscopy to characterize biological materials. *Nat. Protoc.* **11**, 664–687.

Byrne, H. J., Knief, P., Keating, M. E., and Bonnier, F. (2016). Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. *Chem. Soc. Rev.* **45**, 1865–1878.

Ceppatelli, M., Scelta, D., Serrano-Ruiz, M., Dziubek, K., Garbarino, G., Jacobs, J., Mezouar, M., Bini, R., and Peruzzini, M. (2020). High pressure synthesis of phosphine from the elements and the discovery of the missing (PH3)2H2 tile. *In* "Nature Communications," Vol. 11.

Chandel, R., Sharma, P. C., and Gupta, A. (2018). Method for detection and removal of arsenic residues in calcium carbide ripened mangoes. *J. Food Process. Preserv.* **42**.

Chandran, S. and Singh, R. S. P. (2007). Comparison of various international guidelines for analytical method validation. *Die Pharm. Int. J. Pharm. Sci.* **62**, 4–14.

Chiriu, D., Ricci, P. C., Polcaro, A., Braconi, P., Lanzi, D., and Nadali, D. (2014). Raman study on Pompeii potteries: The role of calcium hydroxide on the surface treatment. *J. Spectrosc.* **2014**.

Ciaburro, G. and Venkateswaran, B. (2017). "Neural network with R." Packt Publishing.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* **20**, 273–297.

Cramer, B., Hübner, F., and Humpf, H. U. (2017). Applications of High-Performance Liquid Chromatography-Mass Spectrometry Techniques for the Analysis of Chemical Contaminants and Residues in Food. *Chem. Contam. Residues Food Second Ed.*, 51–66.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and

Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology* **88**, 2783–2792.

Demtröder, W. (2008). “Laser spectroscopy: vol. 2: experimental techniques.” Springer Science & Business Media.

Dhembare, A. J. (2013). Bitter truth about fruit with reference to artificial ripener. *Sch. Res. Libr. Arch. Appl. Sci. Res.* **5**, 45–54.

Dhembare, A. J. and College, P. V. P. (2013). Bitter truth about fruit with reference to artificial ripener. **5**, 45–54.

Dieing, T., Hollricher, O., and Toporski, J. (2011). “Confocal raman microscopy.” Springer.

Edwards, H. G. M. (1990). Vibration�rotational Raman spectra of acetylene, 12C2H2. *Spectrochim. Acta Part A Mol. Spectrosc.* **46**, 97–106.

Fan, X., Ming, W., Zeng, H., Zhang, Z., and Lu, H. (2019). Deep learning-based component identification for the Raman spectra of mixtures. *Analyst* **144**, 1789–1798.

Fattah, S. A. and Ali, M. Y. (2010). Carbide ripened fruits-A recent health hazard. *Faridpur Med. Coll. J.* **5**, 37.

Feynman, R. P., Leighton, R. B., and Sands, M. (2011). “The Feynman Lectures on Physics, Vol. I: The New Millennium Edition: Mainly Mechanics, Radiation, and Heat.” Basic books.

Francois, M. R. and Stephen, F. (2015). Phosphorus Compounds. *In* “Hamilton & Hardy’s Industrial Toxicology,” pp383–390. John Wiley & Sons, Inc. Hoboken, New Jersey.

Gautam, R., Vanga, S., Ariese, F., and Umapathy, S. (2015). Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Tech. Instrum.* **2**.

Gershenson, C. (2003). Artificial neural networks for beginners. *arXiv Prepr. cs0308031*.

Ghita, A., Matousek, P., and Stone, N. (2018). Sensitivity of Transmission Raman

Spectroscopy Signals to Temperature of Biological Tissues. *Sci. Rep.* **8**, 1–7.

Gierlinger, N., Keplinger, T., and Harrington, M. (2012). Imaging of plant cell walls by confocal Raman microscopy. *Nat. Protoc.* **7**, 1694–1708.

Gomez-Lazaro, M., Freitas, A., and Ribeiro, C. C. (2017). “Confocal Raman microscopy.” Springer.

Gracia, A. and León, L. (2011). Non-destructive assessment of olive fruit ripening by portable near infrared spectroscopy. *Grasas y Aceites* **62**, 268–274.

Hartshorn, S. (2016). Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners. *Kindle Ed.*, 74.

Haturusihghe, L. S., Silva, D. S. M. De, and Wimlasena, S. (2004). Quantification of arsenic and phosphorus in calcium carbide treated mangoes. *Proc. Adv. Sci.* **60**.

Huang, Y., Kangas, L. J., and Rasco, B. A. (2007). Applications of Artificial Neural Networks (ANNs) in food science. *Crit. Rev. Food Sci. Nutr.* **47**, 113–126.

Kan, H., Wong, C.-M., Vichit-Vadakan, N., Qian, Z., and others (2010). Short-term association between sulfur dioxide and daily mortality: The Public Health and Air Pollution in Asia (PAPA) study. *Environ. Res.* **110**, 258–264.

Kathirvelan, J. and Vijayaraghavan, R. (2017). An infrared based sensor system for the detection of ethylene for the discrimination of fruit ripening. *Infrared Phys. Technol.* **85**, 403–409.

Kendrick, M. (2009). The Origin of Fruit Ripening. *Sci. Am.*, 1–3.

Kesse, S., Oti, K., Asim, M., Sied, M. F., and Bo, W. (2019). Analysis of Phosphorus as an Impurity from the Use of Calcium Carbide as an Artificial Ripening Agent in Banana (Musa acuminate). *Res. Pharm. Heal. Sci.* **5**.

Kiyohara, S., Miyata, T., Tsuda, K., and Mizoguchi, T. (2018). Data-driven approach for the prediction and interpretation of core-electron loss spectroscopy. *Sci. Rep.* **8**, 1–4.

Krenker, A., Bester, J., and Kos, A. (2011). Introduction to the Artificial Neural Networks. *Artif. Neural Networks - Methodol. Adv. Biomed. Appl.*

Lakade, A. J., Sundar, K., and Shetty, P. H. (2018). Gold nanoparticle-based method for detection of calcium carbide in artificially ripened mangoes (Magnifera indica). *Food Addit. Contam. - Part A Chem. Anal. Control. Expo. Risk Assess.* **35**, 1078–1084.

Larkin, P. J. (2017). "Infrared and Raman Spectroscopy: Principles and Spectral Interpretation." Elsevier.

Lee, S., Choi, H., Cha, K., and Chung, H. (2013). Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha. *Microchem. J.* **110**, 739–748.

Lee, S. G., Hyun, S. H., Sung, G. H., and Choi, H. K. (2014). Simple and Rapid Determination of Cordycepin in Cordyceps militaris Fruiting Bodies by Quantitative Nuclear Magnetic Resonance Spectroscopy. *Anal. Lett.* **47**, 1031–1042.

Liu, J., Osadchy, M., Ashton, L., Foster, M., Solomon, C. J., and Gibson, S. J. (2017). Deep convolutional neural networks for Raman spectrum recognition: A unified solution. *Analyst* **142**, 4067–4074.

Liu, Y. and Liu, T. (2011). Determination of pesticide residues on the surface of fruits using micro-Raman spectroscopy. *IFIP Adv. Inf. Commun. Technol.* **347 AICT**, 427–434.

Ma, Y. and Guo, G. (2014). "Support vector machines applications."

Madden, M. G. and Ryder, A. G. (2003). Machine learning methods for quantitative analysis of Raman spectroscopy data. *In* "Opto-Ireland 2002: Optics and Photonics Technologies and Applications," Vol. 4876, pp1130.

Maduwanthi, S. D. T. and Marapana, R. A. U. J. (2019). Induced ripening agents and their effect on fruit quality of banana. *Int. J. Food Sci.* **2019**.

McCreery, R. L. (2001). "Raman Spectroscopy for Chemical Analysis."

MehnazMursalat and Rony, A. H. (2013). A Critical Analysis of Artifical Fruit Ripening. *Chem. Eng. Sci. Mag.* **4**, 1–7.

Meier, R. (2003). Handbook of Vibrational Spectroscopy. *Spectrochim. Acta Part A Mol.*

Biomol. Spectrosc. **59**, 413–414.

Menges, F. (2017). Spectragryph v1. 2.9. *Opt. Spectrosc. Softw.*

Morris, M. D. (2008). “Review - Modern Raman Spectroscopy: A Practical Approach.”

Müller, J., Ibach, W., Weishaupt, K., and Hollricher, O. (2003). “Confocal Raman Microscopy.”

Nagel, M. C. (1989). The fruits of ethylene. *ChemMatters* **7**, 11–13.

Nakamoto, K. (2006). Infrared and R aman Spectra of Inorganic and Coordination Compounds. *Handb. Vib. Spectrosc.*

Naushad, M. and Khan, M. R. (2014). “Ultra performance liquid chromatography mass spectrometry: Evaluation and applications in food analysis.”

Nims, C., Cron, B., Wetherington, M., Macalady, J., and Cosmidis, J. (2019). Low frequency Raman Spectroscopy for micron-scale and in vivo characterization of elemental sulfur in microbial samples. *Sci. Rep.* **9**, 1–12.

Nowshad, F., Imtiaz, M. Y., Khan, M. S., Shadman, S. A., Islam, M. N., and Alam, S. S. (2018). Artificial ripening on banana (Musa Spp.) samples: Analyzing ripening agents and change in nutritional parameters. *Cogent Food Agric.* **4**, 1–16.

Pirson, J., Toussaint, P., and Segers, N. (2003). An unusual cause of burn injury: skin exposure to monochloroacetic acid. *J. Burn Care Rehabil.* **24**, 407–409.

Qin, J., Kim, M. S., Chao, K., Schmidt, W. F., Dhakal, S., Cho, B. K., Peng, Y., and Huang, M. (2017). Subsurface inspection of food safety and quality using line-scan spatially offset Raman spectroscopy technique. *Food Control* **75**, 246–254.

Ramachandra, B. L., Gumpu, M. B., Nesakumar, N., Krishnan, U. M., and Rayappan, J. B. B. (2016). Calcium carbide in mangoes: An electrochemical way for detection. *Anal. Methods* **8**, 4590–4599.

Roberts, T. R., Hutson, D. H., Lee, P. W., Nicholls, P. H., and Plimmer, J. R. (1998). “Metabolic Pathways of Agrochemical Part 1: Herbicides and Plant Growth Regulators.” Royal Society of Chemistry.

Sathyanarayana, S; Amarappa, S. V. (2014). Data classification using Support vector Machine (SVM), a simplified approaCH. *Int. J. Electron. Comput. S cience Eng. Vol. 3, Number 4, ISSN- 2277-1956*, 435–445.

Schrader, B. (2008). “Infrared and Raman spectroscopy: methods and applications.” John Wiley & Sons.

Shaver, J. M. (2001). Chemometrics for Raman spectroscopy. *Pract. Spectrosc. Ser.* **28**, 275–306.

Shrivastava, A. and Gupta, V. (2011). Methods for the determination of limit of detection and limit of quantitation of the analytical methods. *Chronicles Young Sci.* **2**, 21.

Singal, S., Kumud, M., and Thakral, S. (2012). Application of apple as ripening agent for banana. *Indian J. Nat. Prod. Resour.* **3**, 61–64.

Taylor, P., Huang, Y., Kangas, L. J., and Rasco, B. A. Food Science Applications of Artificial Neural Networks (ANNs) in Food Science. 37–41.

Thissen, U., Üstün, B., Melsseit, W. J., and Buydens, L. M. C. (2004). Multivariate calibration with least-squares support vector machines. *Anal. Chem.* **76**, 3099–3105.

Uhrovčík, J. (2014). Strategy for determination of LOD and LOQ values - Some basic aspects. *Talanta* **119**, 178–180.

Varmuza, K. and Filzmoser, P. (2016). “Introduction to Multivariate Statistical Analysis in Chemometrics.”

Wei, D., Chen, S., and Liu, Q. (2015). Review of fluorescence suppression techniques in Raman spectroscopy. *Appl. Spectrosc. Rev.* **50**, 387–406.

Wu, M., Cai, H., Cui, X., Wei, Z., and Ke, H. (2020). Fast inspection of fruits using nuclear magnetic resonance spectroscopy. *J. Chinese Chem. Soc.* **67**, 1794–1799.

Wyzgoski, F. J., Paudel, L., Rinaldi, P. L., Reese, R. N., Ozgen, M., Tulio, A. Z., Miller, A. R., Scheerens, J. C., and Hardy, J. K. (2010). Modeling relationships among active components in black raspberry (Rubus occidentalis L.) fruit extracts using high-resolution1H nuclear magnetic resonance (NMR) spectroscopy and multivariate statistical analysis. *J. Agric. Food Chem.* **58**, 3407–3414.

Yakubovskaya, E., Zaliznyak, T., Martínez Martínez, J., and Taylor, G. T. (2019). Tear Down the Fluorescent Curtain: A New Fluorescence Suppression Method for Raman Microspectroscopic Analyses. *Sci. Rep.* **9**, 1–9.

Zhang, P. X., Zhou, X., Cheng, A. Y. S., and Fang, Y. (2006a). Raman spectra from pesticides on the surface of fruits. *J. Phys. Conf. Ser.* **28**, 7–11.

Zhang, P. X., Zhou, X., Cheng, A. Y. S., and Fang, Y. (2006b). Raman spectra from pesticides on the surface of fruits. *J. Phys. Conf. Ser.* **28**, 7–11.

Zhang, T., Liang, L., Wang, K., Tang, H., Yang, X., Duan, Y., and Li, H. (2014). A novel approach for the quantitative analysis of multiple elements in steel based on laser-induced breakdown spectroscopy (LIBS) and random forest regression (RFR). *J. Anal. At. Spectrom.* **29**, 2323–2329.

APPENDICES

APPENDIX 1: PCA code in R for Exploratory Analysis

```
#load required package
library(ChemoSpec)

# Reading a matrix data file stored in the working directory
raw <- matrix2SpectraObject(gr.crit =  c("L","M", "H","N"), gr.cols =
c("auto"),
                                freq.unit = "Raman shift (cm^-1)",
                                int.unit = "Intensity",
                                in.file = "ARFNEW31.csv",
                                out.file = "Raman_data")

#Perfom classical PCA
pca<-c_pcaSpectra(raw, choice = "autoscale", cent = TRUE)

#Visualize the scores and loadings plots
plotScores(raw, pca, main ="pca", pcs = c(1,2), tol = 0.01)
plotLoadings(raw, pca, main ="pca",  loads = 1:3, ref = 1)

#Remove frequencies from both ends at once to remain with spectral ROI:
spec1 <- removeFreq(raw, rem.freq = raw$freq > 1200 | raw$freq < 200)

#Remove groups of spectra:
spec2 <- removeGroup(spec1, rem.group = "L")
spec3 <- removeGroup(spec2, rem.group = "M")

#PCA for sulfur ROI
spec4 <- removeFreq(spec3, rem.freq = spec3$freq > 500| spec3$freq <
450)
pca4<-c_pcaSpectra(spec4, choice = "autoscale", cent = TRUE)

#PCA for acetylene ROI
spec4 <- removeFreq(spec3, rem.freq = spec3$freq > 650| spec3$freq <
600)
pca4<-c_pcaSpectra(spec4, choice = "autoscale", cent = TRUE)

#PCA for clacium hydroxide ROI
spec4 <- removeFreq(spec3, rem.freq = spec3$freq > 800 | spec3$freq <
750)
pca4<-c_pcaSpectra(spec4, choice = "autoscale", cent = TRUE)
```

```
#PCA for phosphine ROI
spec4 <- removeFreq(spec3, rem.freq = spec3$freq > 1000| spec3$freq <
950)
pca4<-c_pcaSpectra(spec4, choice = "autoscale", cent = TRUE)

# getting the PC from PCA attributes and saving it as a csv
attributes(pca)
pca_scores <-pca[["x"]]
write.csv(pca_scores,'Raman_raw_pca_scores.csv')
```

APPENDIX 2: ANN, RF and SVM Codes in R for Classification (Qualitative Studies)

ANN Classifier

```
#load required packages
library(neuralnet)

# Load the PC saved data, these will be used as model inputs
data <- read.csv(file.choose(), header = T)

#Transform the samples into factors
data$samples <- as.factor(data$samples)

# Data Partition into training and test set
ind <- sample(2, nrow(data), replace = TRUE, prob = c(0.75, 0.25))
train <- data[ind==1,]
test <- data[ind==2,]

#Designing the neural network
nn = neuralnet(samples ~ .,
               data = train,
               hidden = c(10,5,3),
               linear.output = F)

#Making predictions on the training and test data
pred<- predict(nn, train)
predt<-predict(nn, test)

#confusion matrix for the training and test data
tab<-table( pred[,1]>0.5, train$samples==1)
tab2<-table( predt[,1]>0.5, test$samples==1)

#Calculating error (misclassification) rate for the training and test
data
1-sum(diag(tab))/sum(tab)
1-sum(diag(tab2))/sum(tab2)
```

RF Classifier

```
#load required packages
library(caret)
library(randomForest)

# Designing the Random Forest
rf <- randomForest(samples~.,
                   data=train[,1:21],
                   ntree = 2000,
                   mtry = 4)

# Prediction & Confusion Matrix - train data
p1 <- predict(rf, train[,1:21])
confusionMatrix(p1, train$samples)

# Tune mtry and ntree and save the best tuned parameters
t <- tuneRF(train[,2:21], train[,1],
            ntreeTry = 20,
            stepFactor = 0.5,
            improve = 0.05,
            doBest=TRUE)

# # Prediction & Confusion Matrix - test data
p3<-predict(t, test[,1:21])
confusionMatrix(p3, test$samples)

# Variable Importance plot
varImpPlot(t,sort = T,n.var = 5,main = "Top 5 - Variable Importance")
```

SVM Classifier

```
#Load required packages
library(e1071)

#Read data
data1 <- read.csv(file.choose(), header = T)

#Designing support vector machine model
mymodel <- svm(samples~., data=train)

#Confusion matrix and misclassification error for training data
pred <- predict(mymodel,train)
tab <- table(predicted=pred,actual=train$samples)
1-sum(diag(tab))/sum(tab)

#Fine tuning the SVM model
tmodel <- tune(svm,samples~.,data=train, kernel = "radial",
               ranges= list(cost = c(0.01,0.1, 1, 10,100,1000)))

#Best model from results above
mybestmodel <- tmodel$best.model
```

```
#confusion matrix and misclassification error for test data
predtest <- predict(mybestmodel,test)
tab <- table(predicted=predtest,actual=test$samples)
1-sum(diag(tab))/sum(tab)

#Decision surface (hyperplane) plot
plot(mybestmodel,data=test,
     PC2~PC1, slice = list(PC3=3,PC4=4))
```

APPENDIX 3: ANN, RF and SVR Codes in R for Regression (Quantitative Studies)

Artificial Neural network Regression

```
#Read the Data and transform the concentration values to numeric
data <- read.csv(file.choose(), header = T)
data$conc <- as.numeric(data$conc)

#Splitting the data into training and test set and then scaling using
min-max
samplesize = 0.8 * nrow(data)
index = sample( seq_len ( nrow ( data ) ), size = samplesize )
datatrain = data[ index, ]
datatest = data[ -index, ]
max = apply(data , 2 , max)
min = apply(data, 2 , min)
scaled = as.data.frame (scale(data, center = min, scale = max - min))
trainNN = scaled[index , ]
testNN = scaled[-index , ]

# Fit neural network
NN = neuralnet(conc ~ .,
               trainNN,
               hidden = c(5,5,5),
               threshold = 0.01,
               stepmax = 1e+5, rep = 10,
               learningrate.factor = list(minus = 0.2, plus = 1.7),
               lifesign.step = 1000, algorithm = "rprop+", err.fct =
"sse",
               act.fct = "logistic", linear.output = TRUE)

#plot neural network
plot(NN, rep="best")

# calibration using neural network
predict_trainNN = compute(NN, trainNN[,c(2:25)])
predict_trainNN = (predict_trainNN$net.result * (max(data$conc) -
min(data$conc))) + min(data$conc)
## Prediction using neural network
predict_testNN = compute(NN, testNN[,c(2:25)])
predict_testNN = (predict_testNN$net.result * (max(data$conc) -
min(data$conc))) + min(data$conc)

#Graphs of actual versus the predicted value with a 45 degree slope line
```

```
par(mfrow=c(1,2))
plot(datatrain$conc, predict_trainNN, col='blue', pch=16)
plot(datatest$conc, predict_testNN, col='blue', pch=16)

#Root Mean Square Error of calibration(RMSEC) and Root Mean Square Error
of prediction (RMSEP)
RMSEC(datatrain$conc, predict_trainNN)
RMSEP(datatest$conc, predict_testNN)

#Relative measure of fit for calibartion and validation data (R^2)
Rsqcal_ = 1 - sum((datatrain$conc-
predict_trainNN)^2)/sum(datatrain$conc-
(sum(datatrain$conc)/nrow(datatrain))^2)
Rsqpred_ = 1 - sum((datatest$conc-predict_testNN)^2)/sum(datatest$conc-
(sum(datatest$conc)/nrow(datatest))^2)

#Save calibration values, Predicition values and the ANN model for later
use
valuesc1 <-cbind(datatrain$conc, predict_trainNN)
write.csv(valuesc1, "tabulated_valuesc1.csv")
valuesc11 <-cbind(datatest$conc, predict_testNN)
write.csv(valuesc11, "tabulated_valuesp11.csv")
save(NN, file = "ANNR13.rda")

#Fit a linear model of the actual and the predicted calibration values
to obtain the slope and SD of the y-intercept of the psuedounivariate
graph to calculate LOD and LOQ.
Lm_Mod <- lm(datatrain$conc, predict_trainNN)
summary(Lm_Mod)

#Load the saved model for use in prediciting new data sets and saving
the results to a CSV file
load(file="ANNR13.rda")
df2 = read.csv(file.choose(), header = T)
DataPred <- compute(NN, df2[,c(1:24)])
new_predictions1 <- DataPred$net.result
write.csv(new_predictions1, file = "new_predictionsP.csv")
```

Random Forest Regression

```
# Develop the Random Forest model
rf <- randomForest(conc~.,
                   data=train[,1:26],
                   ntree = 2000,
                   mtry = 14)

# Tune RF parameters and save best model
t <- tuneRF(train[,2:26], train[,1],
            ntreeTry = 2000,
            stepFactor = 2,
            improve = 0.05,
            doBest=TRUE)

# Variable Importance plot
varImpPlot(rf, sort = T, n.var = 10, main = "Top 10 - Variable 
Importance")

# Calibration using tuned RF model
p1 <- predict(t, train[,1:26])
# Prediction using tuned RF model
p3 <- predict(t, test1)

#Graphs of actual versus the predicted value with a 45 degree slope line
par(mfrow=c(1,2))
plot(train$conc, p1, col='blue', pch= 16)
plot(test1$conc, p3, col='blue', pch=1)

#Root Mean Square Error of calibration(RMSEC) and Root Mean Square Error
of prediction (RMSEP)
RMSEC.rf = (sum((train$conc - p1)^2) / nrow(train)) ^ 0.5
RMSEP.rf = (sum((test1$conc - p2)^2) / nrow(test1)) ^ 0.5

#Relative measure of fit for calibartion and validation data (R^2)
Rsqcalib = 1 - sum((train$conc-p1)^2)/sum(train$conc-
(sum(train$conc)/nrow(train))^2)
Rsqpred = 1 - sum((test1$conc-p2)^2)/sum(test1$conc-
(sum(test1$conc)/nrow(test1))^2)

#Save calibration values, Predicition values and the RF model for later 
use
valuesc2 <-cbind(train$conc, p1)
write.csv(valuesc2, "tabulated_valuesc2.csv")
valuesy22 <-cbind(test1$conc, p3)
write.csv(valuesy22, "tabulated_valuesp22.csv")
save(t, file = "RFR13.rda")

#Load the saved model for use in prediciting new data sets and saving
```

```
the results to a CSV file
load(file="RFR13.rda")
df2 = read.csv(file.choose(), header = T)
DataPred2 <- predict(t, df2[,c(1:25)])
write.csv(DataPred2, file = "new_predictionsPR.csv")
```

Support vector Regression

```
#Develop SVR model
library(e1071)
SVR <- svm(conc ~ .,
           data = train,
           type = 'eps-regression',
           kernel = 'radial',
           cost=10, gamma=0.04)

#Fine tuning and saving model with optimized parameters
tmodel <- tune(svm, conc~ .,
               data=train, kernel = "radial",ranges= list(cost =
c(0.01,0.1, 1, 10)))
mybestmodel <- tmodel$best.model

# Calibration using tuned SVR model
predtrain <- predict(mybestmodel,train)
# Prediction using tuned SVR model
predtest1 <- predict(mybestmodel,test)

#Graphs of actual versus the predicted value with a 45 degree slope line
#prediction -train data
par(mfrow=c(1,2))
plot(train$conc, predtrain, col='blue', pch= 16)
plot(test$conc, predtest1, col='blue', pch=16)

#Root Mean Square Error of calibration(RMSEC) and Root Mean Square Error
of prediction (RMSEP)
RMSEC(predtrain, train$conc)
RMSEP(predtest1, test$conc)

#Relative measure of fit for calibartion and validation data (R^2)
Rsqcal1 = 1 - sum((train$conc-predtrain)^2)/sum(train -
(sum(train)/nrow(train))^2)
Rsqpred1 = 1 - sum((test$conc-predtest1)^2)/sum(test -
(sum(test)/nrow(test))^2)
```

```
#Save calibration values, Predicition values and the SVR model for later
use
valuesc3n <-cbind(train$conc, predtrain)
write.csv(valuesc3n, "tabulated_valuesc3n.csv")
valuesp3n <-cbind(test$conc, predtest1)
write.csv(valuesp3n, "tabulated_valuesp3n.csv")
save(mybestmodel, file = "SVR13.rda")

#Load the saved model for use in prediciting new data sets and saving
the results to a CSV file
load(file="SVR13.rda")
df2 = read.csv(file.choose(), header = T)
DataPred3 <- predict(mybestmodel, df2[,c(1:6)])
write.csv(DataPred3, file = "new_predictionsRS.csv")
```

APPENDIX 4: Model Predictions for Test Data Sets

Table A4.1: The model predicted concentrations in g/L alongside their mean and corresponding standard deviations. The test data set was the data hidden from the ML models during the calibration phase.

	ANN		RF		SVR	
observed	predicted	Mean±SD	predicted	Mean±SD	predicted	Mean± SD
0	-0.07, -0.07, -0.05, -0.03, -0.03	-0.05±0.02	0.01, 0.02, 0.04, 0.01	0.02±0.01	-0.01, -0.04, -0.05	-0.03±0.02
0.24	0.37	0.37±0.00	0.33, 0.30, 0.40, 0.35	0.35±0.04	0.35, 0.15, 0.49, 0.16	0.29±0.14
0.26	0.39	0.39±0.00	0.34, 0.58	0.46±0.12	0.64, 0.31	0.48±0.16
0.28	0.89, 0.27	0.58±0.31	0.34, 0.38	0.36±0.02	0.29, 0.54	0.41±0.13
0.30	0.42, 0.28	0.35±0.07	0.58, 0.38	0.48±0.10	0.55, 1.04	0.79±0.25
0.34	0.51	0.51±0.00	-	-	-	-
0.36	0.26, 0.42, 0.25, 0.44	0.34±0.09	0.34	0.34±0.00	0.37	0.37±0.00
0.40	0.52	0.52±0.00	0.79	0.79±0.00	0.76	0.76±0.00
0.60	0.70, 0.56, 0.33	0.53±0.15	0.58, 0.45, 0.56	0.53±0.05	0.84, 0.62, 0.51	0.66±0.14
1.00	0.91, 1.65	1.28±0.37	1.22, 0.59, 0.83	0.88±0.26	1.22, 0.59, 0.75	0.86±0.27
1.20	1.30, 1.01,	1.19±0.13	-	-	-	-

	1.27					
1.40	-	-	1.51, 1.46, 1.43	1.47±0.03	0.79, 1.11, 1.16	1.02±0.16
1.60	1.74	1.74±0.00	1.62, 1.56	1.59±0.03	1.30, 1.44	1.37±0.07
1.80	1.70, 1.47, 1.79	1.65±0.13	1.00, 1.76	1.38±0.38	1.37, 2.06	1.71±0.34
2.00	1.65, 1.71	1.68±0.03	1.79	1.79±0.00	1.38	1.38±0.00

APPENDIX 5: EDXRF Analysis of Calcium Carbide

Table A4.2: Elemental composition of the industrial grade calcium carbide used in ripening bananas artififically.

Analyte	Result (ppm)
S	518000
Ca	301000
Pb	95600
Al	43300
Si	15000
Mg	10100
Cd	7710
Fe	5530
P	1570
Hg	848