# RAPID DETECTION OF CHLORPYRIFOS IN KALE AND MILK USING MACHINE LEARNING-AIDED RAMAN SPECTROSCOPY

**MAINA THUKU JEREMIAH (B.Sc)**

**I56/28857/2019**

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE (CHEMISTRY) IN THE SCHOOL OF PURE AND APPLIED SCIENCES OF KENYATTA UNIVERSITY**

**DECEMBER, 2025**

## DECLARATION

This thesis is my original work and has not been presented for the award of any degree in any university or any other institution of higher learning.

Signature…….……. Date……

**Maina Thuku Jeremiah (B.Sc)**
I56/28857/2019
Department of Chemistry

## SUPERVISORS

We confirm that the work reported in this thesis was carried out by the candidate under our supervision.

Signature……. Date……

**Dr. Lucy Kiruri**
Department of Chemistry
Kenyatta University

Signature……. Date……...

**Dr. Ian Kaniu**
Department of Physics
University of Nairobi

## DEDICATION

This thesis is dedicated to my father, whose unwavering encouragement has been a constant source of strength throughout this journey. To my late mother, whose presence and support were with me as I embarked on this academic quest. To my loving siblings, whose support and understanding have been invaluable.

## ACKNOWLEDGMENTS

I am deeply grateful to Almighty God, whose strength and wisdom have guided me throughout this journey. I extend my sincere and heartfelt thanks to my supervisors, Dr. Lucy Kiruri and Dr. Ian Kaniu, for their unwavering support, insightful guidance, and mentorship. Your expertise in academic research has shaped my skills, and your encouragement has been invaluable throughout the research process. I sincerely thank Prof. Kenneth Kaduki, the thematic leader for the Laser Physics and Spectroscopy research group at the University of Nairobi (UoN), for providing valuable resources throughout my study. Special thanks to Dr. Ian and Dr. Ndung'u of the Department of Physics, UoN, for generously providing research samples and helping preserve them. I am indebted to Dr. Ian for offering me workspace at UoN and to Dr. Ndung'u for his friendly advice, continuous support, and technical guidance. Your efforts in ensuring I had access to work late into the night and offering transport are deeply appreciated.

I would also like to acknowledge the technical staff at the Department of Chemistry, KU, led by Mr. Njagi, for their assistance with laboratory equipment, and Mr. Omucheni, Chief Technologist, Department of Physics, UoN, for providing a supportive laboratory environment. Special thanks to Dr. Moraa and Carol for their guidance and generosity in facilitating my experimental work. I am incredibly thankful to my friends and colleagues, Joseph Auka and David Thairu, for their support throughout this journey.

Lastly, my deepest gratitude goes to my family for their endless encouragement, prayers, and financial support. Your love and faith have been my greatest motivation. I am forever grateful to God for you all.

# TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **ANOVA** | Analysis of Variance |
| **DEP** | Diethyl Phosphate |
| **GC-MS** | Gas Chromatography Mass Spectroscopy |
| **HPLC** | High Performance Liquid Chromatography |
| **LOD** | Limit of Detection |
| **LOQ** | Limit of Quantification |
| **ML** | Machine Learning |
| **MRL** | Maximum Residue Limit |
| **OPP** | Organophosphorus pesticide |
| **PC** | Principal Component |
| **PCA** | Principal Component Analysis |
| **ppm** | Parts per Million |
| **$R^2$** | Coefficient of Determination |
| **RF** | Random Forest |
| **RMSEP** | Root Mean Square Error of Prediction |
| **SERS** | Surface Enhanced Raman Spectroscopy |
| **SG** | Savitzky-Golay |
| **SNR** | Signal-to-Noise Ratio |
| **SVM** | Support Vector Machine |
| **SVR** | Support Vector Regression |
| **TCP** | 3,5,6-trichloro-2-pyridinol |

## ABSTRACT

Chlorpyrifos, a widely used organophosphorus pesticide in Kenya, is banned from use on vegetables due to its health risks; however, studies show it is still widely used and detected in food products. Conventional detection methods, such as gas chromatography (GC) and high-performance liquid chromatography (HPLC), are accurate but costly, time-consuming, and destructive, making them unsuitable for rapid on-site analysis. This study aimed to develop a fast, non-destructive method for detecting chlorpyrifos in milk and kale using Raman spectroscopy and machine learning (ML). ML involves computational algorithms that analyze complex data patterns, improving prediction accuracy and classification. These techniques were crucial for efficiently processing spectral data, recognizing patterns, and building predictive models for chlorpyrifos detection. Raman spectroscopy was chosen for its solvent-free, non-invasive nature. Spectral preprocessing steps, including baseline correction, smoothing, and normalization, improved signal quality. Analysis of Variance (ANOVA) was applied to identify Raman bands with statistically significant differences, and Principal Component Analysis (PCA) revealed the spectral fingerprint and reduced dimensionality. The 314-354 $cm^{-1}$ spectral band, centered at 342 $cm^{-1}$, was identified as the chlorpyrifos Raman fingerprint due to distinct C-Cl vibrational modes absent in untreated samples. Machine learning models, including Support Vector Machine (SVM), Support Vector Regression (SVR), and Random Forest (RF), were trained using Principal Components (PCs) from the fingerprint. These models were used to classify chlorpyrifos levels in the samples with respect to the Maximum Residue Limit (MRL), the highest permissible pesticide concentration in food for consumer safety, ensuring the models provided relevant food safety assessments. Classification models achieved high accuracy: SVM outperformed RF with 95.79% accuracy in milk and 92.61% in kale, while RF achieved 95.23% and 90.15%, respectively. In regression tasks, RF showed superior performance with a coefficient of determination ($R^2$) > 0.9997 and a root mean square of prediction (RMSEP) < 0.0231 ppm, compared to SVR's $R^2$ > 0.9961 and RMSEP < 0.0897 ppm. These results confirm that Raman spectroscopy combined with ML offers a highly accurate, rapid, and non-destructive alternative to conventional methods, enhancing real-time food safety monitoring and regulatory compliance.

# CHAPTER ONE

# INTRODUCTION

## 1.1 Background to the Study

Pesticide contamination of agricultural products, such as milk and vegetables, has recently received considerable attention worldwide (de Andrade *et al*., 2023). This results from the increased application of pesticides to control and kill pests in crops and milk-producing animals, increasing output. One category whose global use has grown substantially is organophosphorus pesticides (OPPs), owing to their market availability and low prices (Mali *et al*., 2023). This class of pesticides is estimated to account for 45% of global pesticide and herbicide usage (Mali *et al*., 2023). OPPs are esters of phosphoric acid, which are formed when phosphoric acid reacts with alcohol through esterification. These compounds often have complex chemical structures comprising heterocyclic, aliphatic, and phenyl derivatives. Examples are diazinon, chlorpyrifos, parathion, malathion, and dichlorvos (Zou *et al*., 2022).

O, O-diethyl O-3,5,6-trichloro-2-pyridyl phosphorothioate, commonly known as chlorpyrifos, is one of the most extensively used OPPs for controlling insects on food crops such as corn, wheat, and apples, as well as managing ticks in livestock (Ambreen and Yasmin, 2021; Wołejko *et al*., 2022). Rekha (2005) describes chlorpyrifos as the world's leading insecticide in terms of volume and effectiveness against many pests. Since its first registration in the US in 1965, chlorpyrifos has been instrumental in the US and worldwide pest control efforts (Raj and Kumar, 2022). It is estimated that 5.1 million pounds of chlorpyrifos were used annually between 2014 and 2018 in the US (US Environmental

Protection Agency (EPA), 2020), while by 2017, about one hundred countries globally registered chlorpyrifos, permitting its use on over 50 different crops (Wołejko *et al*., 2022). Chlorpyrifos is widely used by Kenyan farmers as a leading acaricide and an effective insecticide for maize, sweet potatoes, pineapples, rice, beans, cabbages, spinach, and tomatoes (Adum *et al*., 2021; Asamba *et al*., 2022a; Inonda *et al*., 2015). It then follows that the residues of this toxic chemical are present in most agricultural products, including vegetables and milk.

Conventional analytical methods for chlorpyrifos determination in food matrices, such as GC and HPLC, offer excellent sensitivity and selectivity but are often time-consuming, expensive, and laboratory-bound, requiring extensive sample preparation, trained personnel, and organic solvents (Elsaadani *et al*., 2025; Peris-Vicente *et al*., 2022). In contrast, Raman spectroscopy provides rapid, non-destructive, label-free analysis with minimal sample preparation (Pimenta and Correia, 2025), making it attractive for routine screening of pesticide residues directly on complex matrices such as milk and leafy vegetables. However, Raman spectra from such matrices often contain overlapping bands from multiple sample constituents (Ember *et al*., 2017), making visual interpretation and simple univariate analysis inadequate for reliable residue detection and quantification. To overcome these challenges, chemometric and machine learning methods can be used to extract subtle spectral features, handle high-dimensional data, and build robust models that discriminate between contaminated and uncontaminated samples and predict chlorpyrifos levels with improved accuracy.

### 1.1.1 Health Effects of Chlorpyrifos

Although chlorpyrifos is an important OPP that is used on crops and animals to boost production, it has widely been associated with toxic effects such as nervous system disorders, birth defects, immune system imbalances, impaired brain function in children, attention deficit disorder, lower birth weight, and leukemia (Ambreen and Yasmin, 2021; Christensen *et al*., 2009; Eskenazi *et al*., 2014; Foong *et al*., 2020; Sass, 2022). An investigation into the reproductive effects of chlorpyrifos established that children born to mothers exposed to high chlorpyrifos levels exhibit developmental delays, attention-deficit disorders, and hyperactivity (Christensen *et al*., 2009).

### 1.1.2 Regulatory Status of Chlorpyrifos

Due to the multiple health risks associated with chlorpyrifos, various governments and authorities worldwide have adopted measures to address the dangers of this chemical. Such measures include the establishment of a Maximum Residue Limit (MRL), which refers to the maximum pesticide residue levels allowable in foodstuffs (Vettorazzi, 1977). For chlorpyrifos, the established MRL is 0.01 ppm (Ma *et al*., 2020; Shaker and Elsharkawy, 2015; Vemuri, 2016).

Chlorpyrifos has been banned in many countries. For instance, based on accumulating evidence showing that the chemical posed an increased risk of multiple health problems, the US EPA reassessed chlorpyrifos' human health risk, banning its use in food production in 2016 (Sass, 2022). The European Commission (EC) also prohibited the use and marketing of pesticides with chlorpyrifos in 2020 (Wołejko *et al*., 2022). Similarly, in 2019, Australia suspended and canceled domestic and home garden uses of chlorpyrifos products

due to their environmental and health impacts (Waras *et al*., 2020). Further, chlorpyrifos use has also been restricted in Canada, India, China, Thailand, New Zealand, and Argentina (Schulte, 2021; The United Nations Environment Programme (UNEP), 2022). In Africa, South Africa, Morocco, and Egypt are among the countries that have banned the use of chlorpyrifos (UNEP, 2022).

### 1.1.3 Chlorpyrifos Utilization in Kenya

Chlorpyrifos is registered in 25 products in Kenya and categorized as a highly hazardous substance. Its use on vegetables is not permitted, but it can be used to control various pests in maize, barley, wheat, and pineapples (East Africa Natural History Society (EANHS), 2021). Chlorpyrifos appears on the Pest Control Products Board (PCPB) of Kenya's list of hazardous pesticides whose ban and immediate registration termination are unavoidable to shield the environment, agriculture, and citizens' health (Human Rights Watch, 2023). However, despite the health safety risks linked to chlorpyrifos and the regulatory efforts of Kenyan authorities, it remains one of Kenyan farmers' most commonly used pesticides for crops, including kale, spinach, avocado, coffee, cabbage, melon, tomatoes, sweet potatoes, rice, and maize (EANHS, 2021).

Significant evidence shows that chemical pesticides containing chlorpyrifos are still widely used in the country, notwithstanding their damaging effects on the environment and human health. A recent baseline study in Kilifi, Nakuru, and Kajiado Counties found that chlorpyrifos accumulation in water, spray race, dip wash, soil, and milk exceeds the recommended limits set by WHO at 0 to 0.01 mg/kg body weight (Adum *et al*., 2021). Also, a recent investigation has established that pesticides, such as chlorpyrifos, already

banned within the European Union markets, are still widely used in Kenya, despite the danger they pose to the health of Kenyans (Leskovac and Petrović, 2023; Mwendwa, 2023). Failure to comply with the recommended MRL increases the threat of exposure to dangerous chlorpyrifos levels, leading to adverse health effects. According to Mebdoua (2019), fresh produce, including vegetables, often contains high chlorpyrifos residues mainly because the chemical is repeatedly applied to combat pests. Equally, the extensive use of chlorpyrifos in dairy farms increases the probability of milk contamination (Adum *et al*., 2021; Asamba *et al*., 2022b). Chlorpyrifos milk contamination also arises from consuming animal feeds and water containing the pesticide (Bedi *et al*., 2018). Thus, there is a need to analyze milk and vegetables, such as kale, in the market, to determine whether their chlorpyrifos residue levels comply with international standards.

### 1.2 Statement of the Problem

Chlorpyrifos continues to be used extensively in Kenya despite the high human health risks it poses to unsuspecting citizens. As a result, its residues are present in many freshly harvested products in Kenyan markets, posing a significant threat to food safety in the country. Therefore, it is paramount to constantly monitor chlorpyrifos residue levels in fresh farm produce and milk to reduce both the chemical's short- and long-term harmful effects. Previous studies have employed conventional pesticide residue analysis methods, such as GC, HPLC, and GC-MS, to examine chlorpyrifos levels in vegetables and milk (Adum *et al*., 2021; Asamba *et al*., 2022a; Asamba *et al*., 2022b; Inonda *et al*., 2015). While these methods are accurate and highly reliable, they have limitations in terms of being destructive (causing irreversible damage to samples), time-consuming (multiple sample preparation steps), and expensive (solvents and reagents, regular instrument maintenance).

Raman spectroscopy offers a rapid, nondestructive, and solvent-free alternative that can be applied directly to complex matrices such as kale and milk. However, the resulting spectra often contain overlapping bands, making simple visual inspection and peak-based analysis insufficient for reliable residue assessment. Coupling Raman spectroscopy with machine learning techniques enables the extraction of relevant spectral features and the development of robust models for the classification and quantification of chlorpyrifos in such matrices.

### 1.3 Justification

The continued use of chlorpyrifos in crop production and dairy farming in Kenya, despite its associated risks, underscores the need for reliable monitoring of residues in commonly consumed products such as kale and milk. Conventional analytical methods provide high accuracy and sensitivity but are destructive, require complex sample preparation, and are restricted to laboratory settings, making them unsuitable for the routine screening of large numbers of samples. This necessitates alternative approaches that are rapid, non-destructive, and less resource-intensive, yet sufficiently reliable for detecting chemical residues on agricultural products. Raman spectroscopy is one such technique, as it provides non-destructive, rapid, and relatively low-cost analysis with minimal or no sample preparation (Ayvaz *et al*., 2017; Kucha *et al*., 2018). However, the complex and overlapping nature of Raman spectra from nonhomogeneous matrices such as kale and milk hinders straightforward interpretation. By coupling Raman spectroscopy with chemometric and machine learning techniques, it is possible to extract relevant spectral features and develop models for the classification and quantification of chlorpyrifos residues. The present study is warranted by the potential of integrating Raman spectroscopy with machine learning to enable rapid screening for chlorpyrifos in kale and milk, thereby

providing regulators and laboratories with information to support residue monitoring and ensure food safety.

## 1.4 Objectives

### 1.4.1 General Objective

The main objective of this work was to develop a rapid detection method for chlorpyrifos in kale and milk using Raman spectroscopy coupled with machine learning.

### 1.4.2 Specific Objectives

(i) To identify statistically significant Raman bands for chlorpyrifos detection in kale and milk.

(ii) To determine a Raman spectral fingerprint that differentiates chlorpyrifos-contaminated kale and milk samples from controls.

(iii) To evaluate the performance of machine learning algorithms in classifying samples and quantifying their chlorpyrifos levels using the identified Raman spectral fingerprint.

## 1.5 Significance

This study demonstrates the feasibility of using Raman spectroscopy combined with machine learning as a rapid, non-destructive approach for chlorpyrifos residue screening in kale and milk. The ability to analyze samples with minimal preparation and without destroying them has practical benefits for quality control laboratories and regulatory agencies that need to handle many samples within limited time and resource constraints. In addition, the approach reduces reliance on solvent-intensive chromatographic methods and the associated costs, waste generation, and environmental impacts. By presenting a Raman spectroscopy-based method for classification and quantification of chlorpyrifos in two

representative food matrices, this study contributes to the broader development of spectroscopic and data-driven approaches that can be adapted for routine monitoring of pesticide residues.

### 1.6 Limitations of the Study

This study focused on a single active compound, chlorpyrifos, and did not include other pesticides or their degradation products. The experimental work was carried out using spiked kale and milk to ensure controlled concentrations and repeatability. Consequently, the performance of the developed models on field samples with unknown residue levels was not directly assessed. In addition, the machine learning models deployed in the study required substantial amounts of data to yield viable results and precise predictions, which necessitated the collection of a large number of spectra (100 spectra per sample). Furthermore, the study used milk and kale samples, which are highly perishable; therefore, the samples were temporarily refrigerated before analysis, introducing additional constraints on sample handling.

# CHAPTER TWO

# LITERATURE REVIEW

## 2.1 Chemical Properties and Degradation of Chlorpyrifos

Chlorpyrifos, a colorless to white crystalline solid with low water solubility, contains a trichlorinated pyridine ring and exposed chlorine (Cl) and phosphorothioate (P=S) groups, as shown in Figure 2.1. This implies that the duo can readily interact with other molecules, thus significantly determining the compound's reactivity.

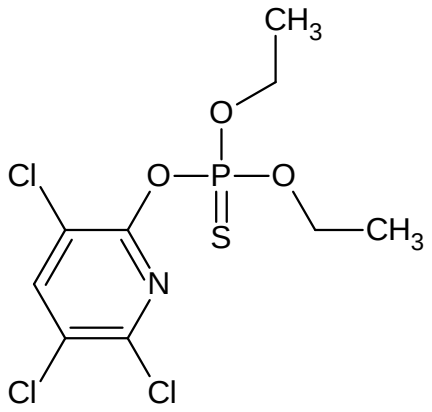


**Figure 2.1: Structural formula of chlorpyrifos**

According to Chen *et al*. (2012), chlorpyrifos undergoes natural degradation in the environment, mainly through oxidation and hydrolysis. The atmospheric OH radicals enhance the oxidation of chlorpyrifos, converting the P=S moiety into P=O (Kharabsheh *et al*., 2017). The major metabolite during this process is diethyl (3,5,6-trichloropyridin-2-yl) phosphate, commonly known as chlorpyrifos-oxon, which is more toxic than chlorpyrifos (Dhiraj *et al*., 2020). Chlorpyrifos-oxon further degrades through hydrolysis to form 3,5,6-trichloro-2-pyridinol (TCP) and diethyl phosphate (DEP), with the former

being the major metabolite (Chen *et al*., 2012; Rahman *et al*., 2021). The degradation process of chlorpyrifos is summarized in Figure 2.2.

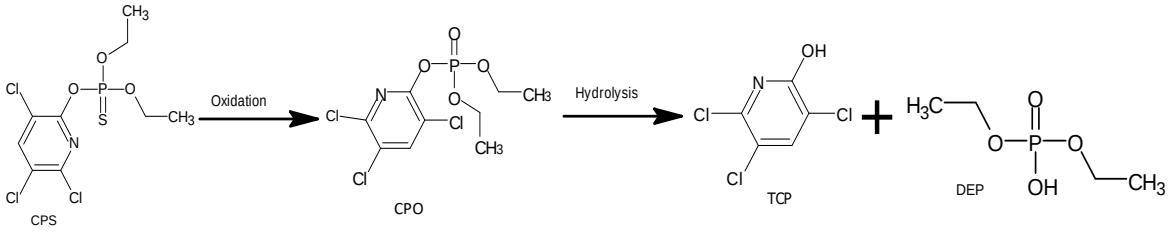


**Figure 2.2: Major degradation products of chlorpyrifos**

From Figure 2.2, it is evident that the pyridine ring remains unchanged during degradation. Chlorpyrifos also undergoes biodegradation in the presence of microorganisms such as fungi and bacteria, with TCP being the primary metabolite (Dhiraj *et al*., 2020). TCP is more persistent, toxic, and soluble than chlorpyrifos, implying that this particular metabolite causes extensive aquatic and soil environment contamination (Kharabsheh *et al*., 2017). Due to its antibacterial properties, TCP rapidly accumulates in plant tissues and inhibits the biodegradation of chlorpyrifos (Abraham and Silambarasan, 2016). Consequently, this metabolite and its parent compound occur widely in crops and the environment where chlorpyrifos is used, threatening ecological systems and public health. Depending on conditions such as climate and soil type, the half-life of chlorpyrifos can extend between 2 weeks to more than a year, but typically ranges from 10 to 120 days (Abraham and Silambarasan, 2016). Studies show that the half-life of chlorpyrifos in cattle tissues ranges from 4 to 7 days (International Programme on Chemical Safety, 1972). Similarly, the half-life of TCP in soil has been found to range from 65 to 360 days (Hou *et al*., 2022; Li *et al*., 2010). The long half-lives of chlorpyrifos and TCP further explain their persistence in the environment and animal fat tissues.

### 2.1.1 Mode of Action of Chlorpyrifos

Chlorpyrifos is active against various insect pests in household, agricultural, and veterinary settings. Like other organophosphate insecticides, chlorpyrifos damages or blocks acetylcholinesterase, an enzyme critical in controlling nerve signals in pests (Christensen *et al*., 2009). Acetylcholinesterase (AChE) breaks down acetylcholine (ACh), a neurotransmitter, preventing its accumulation at the synaptic junction for the body's proper functioning. Exposing insects to chlorpyrifos causes acute neurotoxicity through the inhibition or suppression of AChE, leading to ACh buildup at the synaptic junction. Chlorpyrifos binds to the enzyme's active site, preventing the ACh's breakdown. The accumulation of Ach causes uncontrolled muscle and nerve stimulation, resulting in muscle tetany, exhaustion, and eventually death (Ambreen and Yasmin, 2021). In humans, acute exposure to chlorpyrifos exhibits within minutes to hours with initial signs including tearing of the eyes, nausea, increased sweat and saliva production, headache, and runny nose (Christensen *et al*., 2009).

### 2.1.2 Chlorpyrifos Residues in Vegetables and Milk

Multiple studies have been conducted to investigate the presence of chlorpyrifos in vegetables and milk. For instance, Momtaz and Khan (2024) analyzed chlorpyrifos residues in cabbage, eggplant, and cauliflower samples and compared the residue levels with the MRL. The study detected chlorpyrifos residues in 66% of the cabbage samples, with 65% having levels above the MRL (Momtaz and Khan, 2024). Similarly, 80% of the eggplant samples contained chlorpyrifos, with levels higher than the MRL in 65% of the samples. For cauliflower samples, chlorpyrifos was found in 44%, and 91% of the samples had residues above the MRL (Momtaz and Khan, 2024). Hongsibsong *et al*. (2020) also

conducted a study to monitor chlorpyrifos residue levels in vegetable samples. The researchers reported that over 33.8% of all the samples used in the study tested positive for chlorpyrifos, with levels as high as 275, 145, and 35.8 ppm being detected in cucumbers, coriander, and morning glory, respectively. Significantly high chlorpyrifos residue levels of 26.95 and 332 ppm were also detected in Chinese kale and Chinese cabbage, respectively (Hongsibsong *et al*., 2020).

In a study to assess pesticide residue levels in peri-urban milk, Gill *et al*. (2020) identified chlorpyrifos as one of the leading sources of milk contamination. Some milk samples collected from different sites during the study contained chlorpyrifos levels above the MRL, with the leading group comprising 11.2% of contaminated milk. The least contaminated group had 6.3% of the samples with chlorpyrifos residue levels higher than the MRL (Gill *et al*., 2020). Another study by Dasriya *et al*. (2021) aimed to detect pesticides in milk, cereals, and fruit juices. The researchers found chlorpyrifos and other pesticides in 33 of the 125 collected milk samples, with chlorpyrifos levels in most samples being above the MRL. Further, Asamba *et al*. (2022b) investigated chlorpyrifos' effect on calcium levels in milk. The study found significant levels of chlorpyrifos in 53% of the raw milk samples collected from different counties, with all the chlorpyrifos-contaminated samples exceeding the MRL (Asamba *et al*., 2022b).

According to the available literature, milk is a crucial primary product and a leading protein source in the food industry (Górska-Warsewicz *et al*., 2019; Jonas *et al*., 1976; Ma *et al*., 2019; Pereira, 2014). It is used in the production of multiple secondary products, including

yogurt, cream, and cheese. Among dairy products, milk remains the most widely consumed due to its rich content of fat, protein, and essential minerals, qualifying it as a nearly complete food (Zheng *et al*., 2014). Milk is an essential constituent in many people's daily diet, including vulnerable groups, such as infants, the sick, and the elderly. For this reason, milk was chosen as one of the samples in the present study.

Similarly, kale is among the most widely consumed leafy vegetables due to its nutritional value and high availability worldwide (Šamec *et al*., 2019; Satheesh and Fanta, 2020). It belongs to the *Brassicaceae* family, which includes other crops like cabbage, broccoli, and cauliflower, and is scientifically known as *Brassica oleracea var. sabellica* (Satheesh and Fanta, 2020: Khalid *et al*., 2023). Studies have linked the regular consumption of vegetables such as kale to reduced risk of obesity and chronic diseases such as cancer and cardiovascular diseases (Alfawaz *et al*., 2022; Sheats and Middlestadt, 2013). Consequently, an increasing trend in the consumption of leafy vegetables has been recorded recently (Satheesh and Fanta, 2020). In Kenya, kale is popularly known as *sukuma wiki*. It is highly popular in many households due to low selling prices in the market and low production costs. In many cases, kale is for fresh consumption and is also widely eaten raw in the form of salads, increasing the risk of pesticide residues. Due to its popularity and vast consumption, kale was the vegetable of choice in the current study. Furthermore, Wang *et al*. (2021) report that milk and green leafy vegetables such as kale are among the most affected by adulteration processes, contributing to over 14% of all foodborne illnesses.

## 2.2 Analytical Methods for Pesticide Residue Analysis in Vegetables and Milk

Analytical methods are integral in monitoring pesticide residues in food, including milk and vegetables. Pesticide monitoring and analysis include detecting, identifying, and quantifying certain pesticides or their degradation products. Previous studies have achieved this using spectroscopic, chromatographic, immunoassay, and mass spectrometry techniques. For example, in their investigation of the effects of washing and cooking vegetables on chlorpyrifos and its metabolites, Ling *et al*. (2011) used gas chromatography coupled with triple quadrupole mass spectrometry (GC/MS/MS). Although they were able to detect chlorpyrifos and its primary metabolite TCP in cabbage, eggplant, cucumber, and tomato, they reported a complex sample preparation procedure that involved blending the samples, causing irreversible damage, and the use of acetone, a hazardous solvent (Ling *et al*., 2011). The researchers found that washing and cooking could not entirely remove chlorpyrifos residues and their metabolites from vegetables (Ling *et al*., 2011). Their findings underline the need for the present work, whereby it develops a simple and non-destructive method for pesticide residue analysis.

Sheridan and Meola (1999) used GC/MS/MS to identify pesticides in agricultural samples, including vegetables, fruits, and milk. The study points out some of the key benefits of this analytical method, which include high sensitivity and selectivity, as well as the ability to analyze multiple pesticides simultaneously (Sheridan and Meola, 1999). Nevertheless, the researchers also identified sample matrix interference as a major challenge. A problem was encountered in detecting two pesticides (methamidophos and acephate), which was attributed to the capillary columns' active sites (Sheridan and Meola, 1999). Moreover, the

study reported lab contamination from extraction reagents as a significant obstacle to detection limit determination (Sheridan and Meola, 1999). By employing Raman spectroscopy, the present study overcomes these challenges since it does not involve columns or extraction reagents necessary for traditional separation methods. Also, the fact that Raman spectroscopy relies on bond vibrations rather than compound separation minimizes the problem of matrix interferences.

Further, to analyze pesticide residues in vegetables and fruits, Stocka *et al*. (2016) used gas chromatography with an electron capture detector (GC-ECD) after sample preparation using QuEChERS (Quick, Easy, Cheap, Effective, Rugged and Safe). After optimizing the experimental parameters such as sample amounts, solvent and sorbent types, and extraction time, detection limits of between 0.003 and 0.011 ppm were established for all analytes (Stocka *et al*., 2016). The researchers emphasize the method's high sensitivity and selectivity, making it reliable and convenient for routine pesticide residue monitoring in fruits and vegetables. Using the GC-ECD, the study detected chlorpyrifos residues in apple and tomato samples (Stocka *et al*., 2016). Some of the challenges reported in the study include solvent issues, whereby ethyl acetate resulted in the co-extraction of interfering compounds that contaminated the chromatography column. Additionally, the researchers used a small sample size of about 5g while noting that their method could be cumbersome for routine analysis. The current study circumvents the solvent issues associated with GC-ECD and other traditional analytical methods by employing Raman spectroscopy, a solvent-free technique. Besides, the direct and non-destructive approach adopted by the

present research study helps overcome the challenge that comes with handling large sample volumes.

In another study by Bedi *et al*. (2018), pesticide residues in milk and their connection to pesticide contamination of dairy cattle's feedstuffs were investigated. Using gas chromatography (GC), the researchers detected chlorpyrifos as the primary contaminant in fodder and milk samples from 55 dairy farms, affecting 45.7% of the samples (Bedi *et al*., 2018). The results were also confirmed using a gas chromatography-mass spectrometer (GC-MS) (Bedi *et al*., 2018). However, the researchers highlight the challenge of residue extraction, whereby incomplete extraction due to complex sample matrices such as milk could result in inaccurate measurements (Bedi *et al*., 2018). Notably, the current study overcomes such challenges since it adopts a direct and non-destructive approach that requires minimal to no sample preparation.

Although conventional pesticide residue analysis techniques have proved superior in sensitivity, selectivity, and reliability, they are associated with limitations such as complex sample preparation procedures that are often time-consuming, possible interference from solvents and extraction reagents, and inefficiency in settings with larger sample volumes. However, these limitations can be avoided by adopting simple, cost-effective, and portable methods such as Raman spectroscopy.

## 2.3 Raman Spectroscopy

### 2.3.1 Fundamentals of Raman Spectroscopy

Raman spectroscopy is one of the increasingly popular vibrational techniques with multiple applications in chemistry (Shipp *et al*., 2017). The technique includes shining a monochromatic laser light on a sample and detecting the scattered light. When a sample's molecule scatters light, the photon's oscillating electromagnetic field causes temporary polarization of the molecular electron cloud (Baiz *et al*., 2020). The electric field (E), induces a dipole moment (P), which is proportional to the field as represented in Equation 2. 1.

$$P = \alpha E \tag{2.1}$$

The molecule's polarizability is represented by α (the proportionality constant), which is a measure of the ease with which the molecular electron cloud can be distorted (Tandon *et al*., 2019). The bond's change in polarizability leads to Raman scattering, whose intensity is directly proportional to the square of the induced dipole moment (Keresztury, 2006). A vibration that causes little change to the molecule's polarizability is linked to low Raman band intensity. For instance, the vibrations of a highly polar moiety, such as the O-H bond, are typically weak (Baiz *et al*., 2020). This is because an external field can't induce a significant dipole moment change; hence, bending or stretching vibrations do not change, resulting in a weak or no Raman signal (Seki *et al*., 2020). In contrast, moieties with distributed electron clouds, such as C=C, have a strong Raman scattering effect. An external field easily distorts the double bond's pi-electron cloud, and stretching or bending causes a substantial change in the electron density distribution (Zedler *et al*., 2014). This ultimately gives rise to a significant change in the induced dipole moment.

During the interaction between the photon and molecule, the photon's energy is transferred to the molecule, leaving it in a higher energy state. This phenomenon results in a short-lived complex between the molecule and the photon, often known as the molecule's virtual state (Terrones *et al*., 2023). Notably, the virtual state has an arbitrary energy, which does not have to be an actual allowed state (Shipp *et al*., 2017). This state is typically unstable, and the photon is re-emitted almost immediately (10-15 seconds) as scattered light (Terrones *et al*., 2023). Timing is crucial as it distinguishes Raman scattering and other slower phenomena, such as fluorescence. In most scattering events, the molecule's energy remains unchanged after interacting with the photon. The molecule decays from the virtual state into the ground state, a process that is influenced by temperature and state distribution in the molecule (Shipp *et al*., 2017). Therefore, the energy, wavelength, and frequency of the scattered photon equal those of the incident photon, a phenomenon known as Rayleigh scattering (Li *et al*., 2024). Since Rayleigh scattering conserves the particle's energy, it is also called elastic scattering, and it is often the dominant process (Surzhykov *et al*., 2015).

In contrast, only a significantly small amount of scattering (about one in 10 million) involves an energy shift, which gives rise to Raman scattering (Jones *et al*., 2019). This results in a weak signal, which is one of the significant downsides of Raman spectroscopy. Raman scattering is a rare and inelastic process that involves energy transfer between the photon and molecule (Jones *et al*., 2019). The molecule may gain energy after interacting with the photon and is excited to a higher vibrational level. In contrast, the scattered photon loses energy, increasing its wavelength and leading to Stokes scattering or Stokes shift (Shipp *et al*., 2017). Conversely, the sample molecules may lose energy, falling to a lower

vibrational level. At the same time, the photon gains energy, decreasing its wavelength and leading to anti-Stokes scattering or anti-Stokes shift (Shipp *et al*., 2017). According to quantum mechanics, Stokes scattering dominates over anti-Stokes scattering (Kauffmann *et al*., 2019; Shipp *et al*., 2017). For anti-Stokes scattering to happen, the molecules should already be in an excited vibrational state before interaction with the incident photon, making it rare (Li *et al*., 2024). The populations between the excited ($N_v$) and ground ($N_g$) states can be predicted using equation 2.2.

$$\frac{N_v}{N_g} = e^{\frac{hw_v}{k_B T}} \tag{2.2}$$

In equation 2.2, h is the Planck's constant, $w_v$ represents the vibration energy's frequency, $k_B$ is the Boltzmann constant, and T represents the sample's temperature. Since most molecules will be in the ground vibrational level at room temperature, Boltzmann distribution predicts that Stokes scattering is statistically more likely compared to anti-Stokes (Thyr and Edvinsson, 2023). Consequently, Stokes Raman scattering is more intense than anti-Stokes, which explains why Raman spectroscopy almost entirely measures Stokes scattering (Li *et al*., 2024). Brewer and Kirkwood (2013) note that the anti-Stokes shift is often weaker than Stokes; thus, it is generally ignored and filtered out. Figure 2.3 shows the energy diagrams for Rayleigh and Raman light scattering.

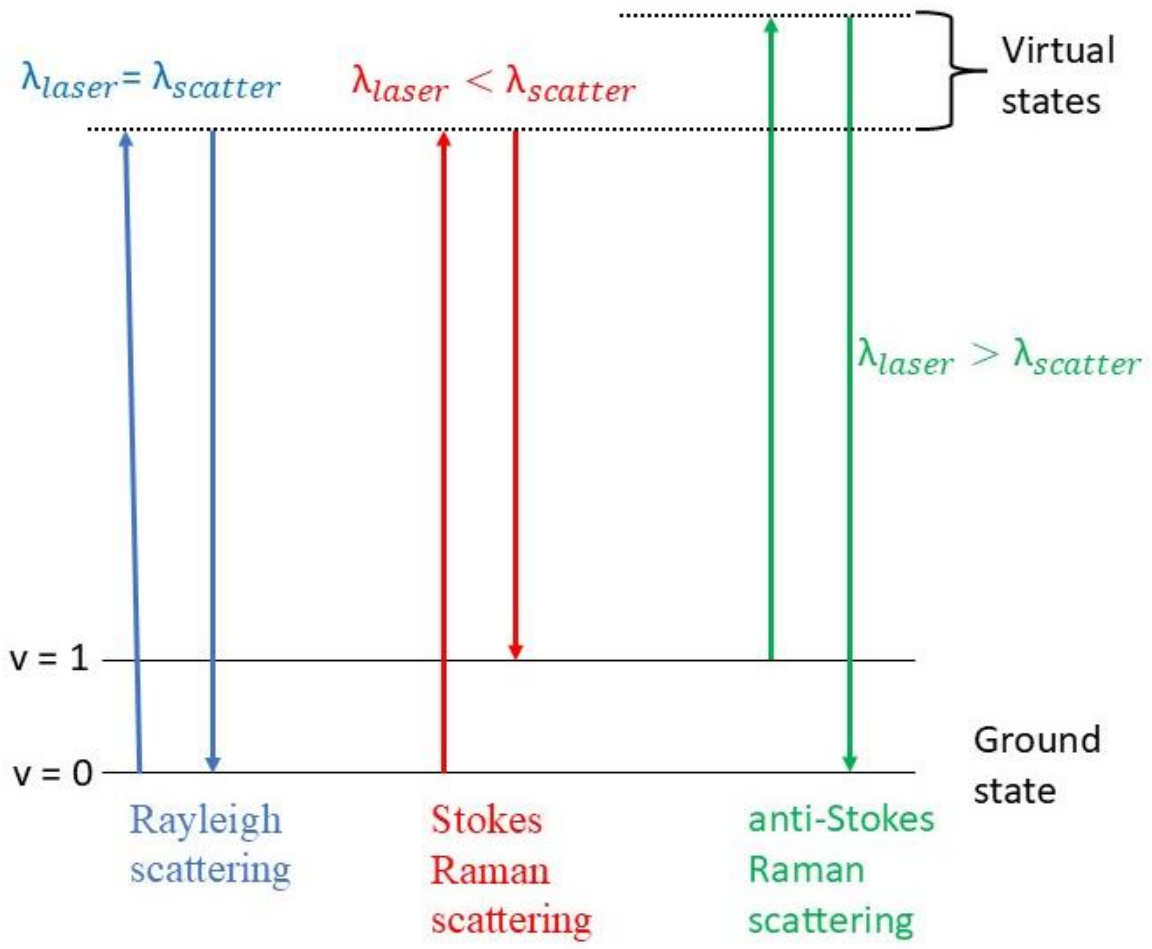


**Figure 2.3: Energy diagrams for Rayleigh and Raman light scattering (Source: Li et al., 2024)**

Therefore, Raman scattering involves incident photons losing or gaining energy after interaction with a sample's vibrating molecules. This technique uses light scattering to reveal valuable information about the molecule's bonding, electronic environment, and structure, enabling substance identification and characterization (Singh *et al*., 2014). As aforementioned, only a small fraction of photons undergoes Raman scattering. This has necessitated techniques to enhance the scattering, giving rise to a method known as surface-enhanced Raman spectroscopy (SERS) (Li *et al*., 2024).

Traditionally, SERS involves using nanostructured materials, such as gold and silver, amplifying the Raman signal by up to $10^{14\text{-}15}$ and enabling the detection of single molecules (Contreras-Caceres *et al*., 2011; Tao *et al*., 2022). Two accepted mechanisms, chemical and electromagnetic, explain the signal enhancement effect (Contreras-Caceres *et al*., 2011). Electromagnetic enhancement involves the adsorption of analyte molecules onto a metal surface followed by an excitation which forms a plasmon whose energy initiates the Raman process in the analyte molecules (Contreras-Caceres *et al*., 2011; Langer *et al*., 2020). On the other hand, chemical enhancement entails the interaction between the analyte molecules and the substrate, modifying the polarizability of the molecules and, thus, the vibrational modes' Raman cross-section (Pilot *et al*., 2019).

Recent studies have demonstrated the utility of aluminium foil as a readily available and inexpensive SERS substrate. For instance, Bukasov *et al*. (2023) applied aluminum foil in the detection of biological compounds and compared it to gold film substrate. The authors reported that aluminium foil performed equally well with conventional SERS substrates such as gold. In agreement with these findings, Sultangaziyev *et al*. (2020) found that aluminium foil has a high potential as a low-cost metal substrate for Raman signal enhancement, sometimes outperforming gold film in terms of reproducibility. The availability of low-cost signal amplification methods qualifies Raman spectroscopy as a suitable vibrational technique. In addition, it is a direct determination method requiring minimal or no sample preparation, allowing for the rapid, non-destructive analysis of solid and liquid samples. For this reason, Raman spectroscopy is ideal for both qualitative and

quantitative analysis of chlorpyrifos residues in solid and liquid agricultural products. The current study uses SERS to analyze chlorpyrifos in milk.

### 2.3.2 Application of Raman Spectroscopy in Pesticide Residue Analysis

Various studies have employed Raman spectroscopy to examine chlorpyrifos and other pesticide residues in diverse agricultural products. Mikac *et al*. (2021) used SERS to detect a range of pesticides in food samples. The researchers highlighted some of the method's notable advantages, including its ability to handle samples in different states (solid and liquid), simplicity, and rapidness (Mikac *et al*., 2021). The researchers add that the discovery of SERS has increased the application of Raman spectroscopy, including the analysis of multiple compounds in situ, with increased sensitivity and repeatability (Mikac *et al*., 2021).

In a different study to detect pesticide residues on fruit surfaces, Chen *et al*. (2018) used SERS to obtain information about the distribution of chlorpyrifos and omethoate residues on the surface of apples. The method allowed for the detection of the peaks of interest, which helped to quantitatively analyze the pesticide residues, exemplifying the rapidness of Raman spectroscopy in detecting pesticides in food and agricultural products.

Further, Tao *et al*. (2022) conducted a study using SERS to detect several pesticides, including chlorpyrifos, carbendazim, and thiabendazole. Using a substrate, the researchers achieved a low detection limit, enabling the trace analysis of pesticide residues on tomato peel (Tao *et al*., 2022). These researchers identify the preparation of flexible substrates as a major limitation of SERS since the process involves complicated procedures and the

nanoparticles' morphology is not easily controlled (Tao *et al*., 2022). Yet, the study demonstrates that high accuracy can be achieved using Raman spectroscopy, allowing for rapid detection of pesticide residues on agricultural products such as tomatoes.

Similarly, Li *et al*. (2014) used SERS to detect fenthion and phorate, organophosphate pesticides, in apple skin. The study's results demonstrated that the Raman peaks of the two pesticides could be easily identified using SERS, further depicting this spectroscopic technique as simple, fast, and convenient in pesticide residue detection. Li and colleagues also noted limitations, such as background noise, outpower stability requirement, and the influence of noises.

A recent study by Chen *et al*. (2023) used SERS to detect residual pyrimethanil and chlorpyrifos on fruit surfaces. The application of SERS enabled the researchers to detect trace amounts of the chemicals in vegetables and fruits within minutes. This demonstrated the possibility of achieving rapid detection of pesticide residues and high sensitivity using Raman spectroscopy. The researchers draw attention to other advantages of Raman spectroscopy, such as its non-destructive analytical ability and the capacity to provide rich information regarding a sample's molecular vibrations and chemical structure (Chen *et al*., 2023).

Wang *et al*. (2024) also employed SERS in pesticide residue analysis of the pericarp. The researchers reported SERS's ability to detect trace amounts of ten different pesticides (10 ppt) in the sample, illustrating the sensitivity of this approach. Like in other previous

studies, Wang *et al*. (2024) described the advantages of SERS, such as broad adaptability, unique fingerprint peaks, and amplified sensitivity. However, the study linked Raman spectroscopy to limitations such as background interference and the need for improved differentiation ability when it comes to in-situ detection.

Furthermore, Pham *et al*. (2022) conducted a study to rapidly determine multiple pesticide residues (acephate, imidacloprid, and carbaryl) in mango fruits using SERS. In their study, the characteristic peaks of the pesticides under investigation were clearly observed, with the lowest detection limit recorded as $5 \times 10^{-5}$mg/kg (Pham *et al*., 2022). In line with the observation made by Tao *et al*. (2022), the researchers noted that the most significant challenge of SERS occurs in the design of the substrate Pham *et al*. (2022). Nevertheless, the study identified the strengths of Raman spectroscopy, including improved sensitivity, rapid analysis due to the absence of traditional time-consuming sample preparation, and the capacity for multi-residue analysis.

A comprehensive review of previous studies that use Raman spectroscopy to detect pesticide residues shows that the method offers several advantages over wet-chemistry techniques. These benefits include its simplicity, rapidness, non-destructiveness, ability to handle samples in different states, and in-situ analysis of multiple compounds. In terms of analytical performance, reported studies also show that Raman, and particularly SERS, can achieve low detection limits for a range of pesticides, with limits of detection typically spanning from the low mg/L (ppm) level down to µg/L or even sub-µg/L (ppb or lower),

depending on the analyte, substrate, and measurement conditions (Dowgiallo and Guenther, 2019; Mayorga *et al*., 2025; Vandenabeele *et al*., 2012; Wang *et al*., 2025).

Notably, the current study leverages these advantages to develop a quick method for detecting chlorpyrifos residues in milk and vegetables. On the other hand, the present study mitigates the limitations of Raman spectroscopy, as identified in prior research. For instance, previous works identify the preparation of SERS substrates as a significant limitation of the method. However, the current study uses aluminium foil as the SERS substrate, thereby eliminating the need for expensive nanostructured materials such as gold and silver. Additionally, background noise and other types of noise have been acknowledged as possible drawbacks of Raman spectroscopy. However, the current study addresses these limitations by using preprocessing methods such as baseline correction and smoothing. Finally, some studies have identified the method's differentiation ability as an area for improvement. In this regard, the present study applies chemometric techniques, such as PCA, to the spectral data, compressing the data and reducing noise, ultimately improving sample differentiation.

## 2.4 Principal Component Analysis in Spectroscopic Data

### 2.4.1 Theory of Principal Component Analysis

Principal component analysis is a chemometric technique that transforms a large set of correlated variables into a smaller set of uncorrelated variables (principal components) that explain most of the original data’s variability. PCA is considered an unsupervised ML technique because it involves a set of predictors or features $X_1$, $X_2$, …, $X_P$, without the associated response Y (Sen and Das, 2023). This technique extracts meaningful

information, removing noise and reducing the dimensionality of a large data set by computing principal components (PCs) and subsequently using these components to reveal data patterns (Kherif and Latypova, 2019). In spectroscopy, PCA clusters data according to their spectral characteristics. Therefore, PCA serves as a data visualization tool in addition to producing variables for supervised ML. Moreover, PCA can be employed for data imputation, which entails filling missing values in a data set (James *et al*., 2013).

Standardizing data before performing PCA is recommended to ensure all variables under consideration are on the same scale. Without standardization, the PCs obtained tend to be biased toward the variables with high variance, which ultimately affects the results of ML models with the PCs as inputs (James *et al*., 2013). PCA finds a linear combination of the features that help explain the maximum data variance. For a matrix X, with *n* observations and $p$ features, up to $p$ distinct PCs can be constructed. For example, a maximum of two PCs is constructed for a two-dimensional (2D) data set. The first PC represents the direction in which the data varies the most and captures most of the information (Mishra *et al*., 2017). It can be represented as shown in equation 2.3.

$$Z_1 = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \cdots + \emptyset_{P1}X_P \qquad 2.3$$

The coefficients $\emptyset_{11}, \ldots, \emptyset_{p1}$ form the loading vector $\emptyset_1 = \emptyset_{11}\ \emptyset_{21} \ldots \emptyset_{p1})^T$ while the constraint $\sum_{j=1}^{p} \emptyset_{j1}^2 = 1$ is introduced by normalization to prevent arbitrarily large variance (James *et al*., 2013). The first PC is computed by optimizing the problem in 2.5, which can be solved by eigen decomposition.

$$maximize\ \emptyset_{11}, \ldots, \emptyset_{p1} \left\{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} \emptyset_{j1}x_{ij}\right)^2\right\} \text{ subject to } \sum_{j=1}^{p} \emptyset_{j1}^2 = 1 \qquad (2.5)$$

The second PC captures the second highest variability in the data and takes the form shown in equation 2.6.

$$Z_{i2} = \emptyset_{12}x_{i1} + \emptyset_{22}x_{i2} + \cdots + \emptyset_{P2}x_{iP} \quad (2.6)$$

The first and second PCs are uncorrelated (Mishra *et al*., 2017), implying that the two are orthogonal or perpendicular. All other PCs are computed by following similar steps while ensuring that each is orthogonal to all the previous PCs. The cumulative variances of the PCA scores determine the optimal number of PCs, with PCs explaining low variances representing noise.

### 2.4.2 Utility of Principal Component Analysis as a Dimensionality Reduction Technique

According to Shi *et al*. (2022), Raman spectroscopy data often involves complex spectra with thousands of dimensions and redundant information. Consequently, subsequent analysis demands more computational resources and negatively influences the accuracy and robustness of the developed models (Shi *et al*., 2022). Using the entire spectrum for modeling is discouraged to optimize ML models and boost their prediction accuracy. Instead, the characteristic range spectrum should be selected for processing and analysis, followed by the extraction of the variables whose contribution rate is high and the use of these variables to model relationships (Shi *et al*., 2022). This process is referred to as feature extraction, and PCA is one of the most commonly used techniques. PCA is a popular technique for simplifying complex spectral datasets. It achieves this by finding recurring data patterns while maintaining minimal information loss (Beattie and Esmonde-White, 2021).

Walse *et al*. (2016) used PCA to reduce the number of raw data features for the development of optimal artificial neural networks (ANN). The researchers investigated how the average classification performance of the model varied with the number of PCs as input. The optimal performance was realized with 70 PCs, reducing the number of features from 561, while retaining most discriminative information (Walse *et al*., 2016). As a result, an overall prediction accuracy of 96.17% was achieved, while the training time was significantly reduced to 128.00 s from 658.53 s when the model was built using all the 561 features (Walse *et al*., 2016). Also, the study compared these results with those of other techniques with more predictive features, such as RF with 186 features, which gave an accuracy of 95% (Walse *et al*., 2016).

In a different study by Shi *et al*. (2022), PCA was used for non-linear feature extraction in processing SERS data for pesticide residue detection on fruit surfaces. The researchers reduced the number of raw data features to only two PCs, accounting for 98.9% of the data variance. Similarly, Chen *et al*. (2023) performed PCA on SERS data and used the first three PCs, cumulatively accounting for 75.1% of the data variance. The PCs were used to build an SVM classification model, achieving an accuracy of 96%. Furthermore, Ma *et al*. (2020) performed PCA on SERS data in a study to rapidly determine chlorpyrifos pesticide residues in tomatoes. PCA reduced the number of features from 24 to 2, enhancing their model's quantification accuracy. Also, using Raman spectroscopy data, Jiang *et al*. (2021) employed PCA as a dimensionality reduction technique in their study to differentiate milk samples from different species, such as cow, human, goat, and buffalo. Their study used the first 8 PCs as inputs to RF, improving their classification model's accuracy to 93.7%.

These studies demonstrate PCA as an essential feature extraction and data compression technique whose output helps improve the performance of ML models.

### 2.5 Supervised Machine Learning

Machine learning is a subset of artificial intelligence (AI) (Wang *et al*., 2024), the capacity of machines to perform human-like tasks. ML allows applications and systems to learn from experience, enhancing performance automatically without programming (Sarker, 2021). Instead of writing a program to solve a problem, ML collects many examples specifying the correct output for a given input. An ML algorithm then learns from the examples and produces a program that processes new input to give the corresponding output. This capacity to perform on new and previously unseen data makes ML models robust (Ren *et al*., 2021). The algorithms' ability to extract meaningful information from big and complex datasets has prompted their wide application in analytical sciences, including mass spectrometry, chromatography, and spectroscopy (Lussier *et al*., 2020).

Supervised learning is the most common ML approach and utilizes labeled data for training in order to map specific inputs to their respective outputs (Sarker, 2021). Labeled data implies that the outcome is known, which informs the algorithm of the correct output from a given dataset. Classification and regression are the main types of supervised learning (Nasteski, 2017). While classification predicts categories/classes and has discrete outcomes, regression involves identifying relationships within multiple variables, which often have a continuous range (Alnuaimi and Albaldawi, 2024). Examples of supervised learning ML algorithms include random forests (RF) and support vector machines (SVM).

### 2.5.1 Random Forest

Random forest is a classifier that employs multiple de-correlated trees for training and predicting samples (Ye *et al*., 2022). As one of the advanced tree-based methods of ML, RF is an improved version of bagging, a procedure that reduces variance by creating multiple sub-samples of the original data and building separate prediction models (Janitza and Hornung, 2018). Like bagging, RF uses bootstrapping (sampling from a single data set with replacement) to create *B* distinct training sets. The model is then trained on the *b*th bootstrapped training set, resulting in the prediction $f^{*b}(x)$. An average of all the predictions is then obtained, as shown in equation 2.7.

$$f_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{*b}(x) \tag{2.7}$$

B trees are constructed and grown deep without pruning, leading to high variance but low bias in individual trees. By combining many trees, RF reduces the variance and significantly improves accuracy (Sun *et al*., 2024). Each decision tree's construction involves recursive binary splitting. A root node, which typically embodies the whole set of inputs or a portion of them, is separated into two or more groups (sub-nodes) (Janitza and Hornung, 2018). The recursive process continues until a terminal node or leaf node, which often represents a minimum number of observations, is reached. Such a complete tree's segment is known as a sub-tree or branch. Notably, each binary split is governed by a set of rules depending on the type of input variables, and the split's quality is determined using either entropy or Gini impurity (James *et al*., 2013). Entropy measures the uncertainty in the dataset at a given split, with lower entropy values indicating purer nodes. Given *i* input features, entropy can mathematically be expressed as:

$$entropy = \sum_i(-P_i log_2 P_i) \quad (2.8)$$

where $P_i$ is the probability that a data point belongs to class *i*. The choice of the split should be such that it minimizes entropy, thus reducing uncertainty and improving the homogeneity of the group. After the first split, the decision to determine the side of the split a particular variable goes is made, and the process unfolds iteratively. Gini impurity, another commonly used metric in RF, measures the probability of misclassifying a randomly chosen data point within a node. It is given by:

$$Gini = 1 - \sum_i P_I^2 \quad (2.9)$$

Thus, a low Gini impurity value of zero means perfect classification, where all values belong to a single class. Like entropy, Gini impurity seeks to minimize entropy, informing the decision tree on the best possible split. Constructing and growing decision trees is repeated across all bootstrapped samples. After each tree is fully grown, it is used to predict outcomes for the out-of-bag (OOB) data, which refers to the subset of training data not included in the bootstrapped set for a particular tree. The final prediction is obtained by aggregating the output of all trees, often through a majority vote in classification problems or averaging in regression tasks. RF offers benefits such as reduced risk of overfitting, flexibility, and low sensitivity to outliers (Ali *et al*., 2012), which is why it was adopted in the current study to classify samples and quantify their chlorpyrifos levels.

### 2.5.2 Support Vector Machine and Support Vector Regression Model

A support vector machine is an algorithm primarily used for classification problems but can be employed in regression tasks through its variant, support vector regression (SVR). SVM generates the optimal hyperplanes or decision boundaries that best segregate n-dimensional space (feature space) into classes, making it easy to put the new data points in

the correct category (Lussier *et al*., 2020). In $p$-dimensional space, a separating hyperplane can be expressed as:

$\beta_0 + \beta_i x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} > 0$ , if $y_i = 1$,

and (2.10)

$\beta_0 + \beta_i x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} < 0$ if $y_i = -1$

Therefore, the hyperplane has the following property:

$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > 0$ , for all $i = 1, \ldots, n.$ (2.11)

This hyperplane (which results in a linear decision boundary) will separate a set of samples into two classes, where an input will be assigned to a class based on the side of the hyperplane on which it falls. The model identifies the extreme vectors or points, known as support vectors, to create the hyperplane. The hyperplane is typically constructed to maximize the margin, the distance between the hyperplane and support vectors (Yang *et al*., 2022), as shown in Figure 2.4, while solving the optimization problem 2.12 for linear class boundaries.

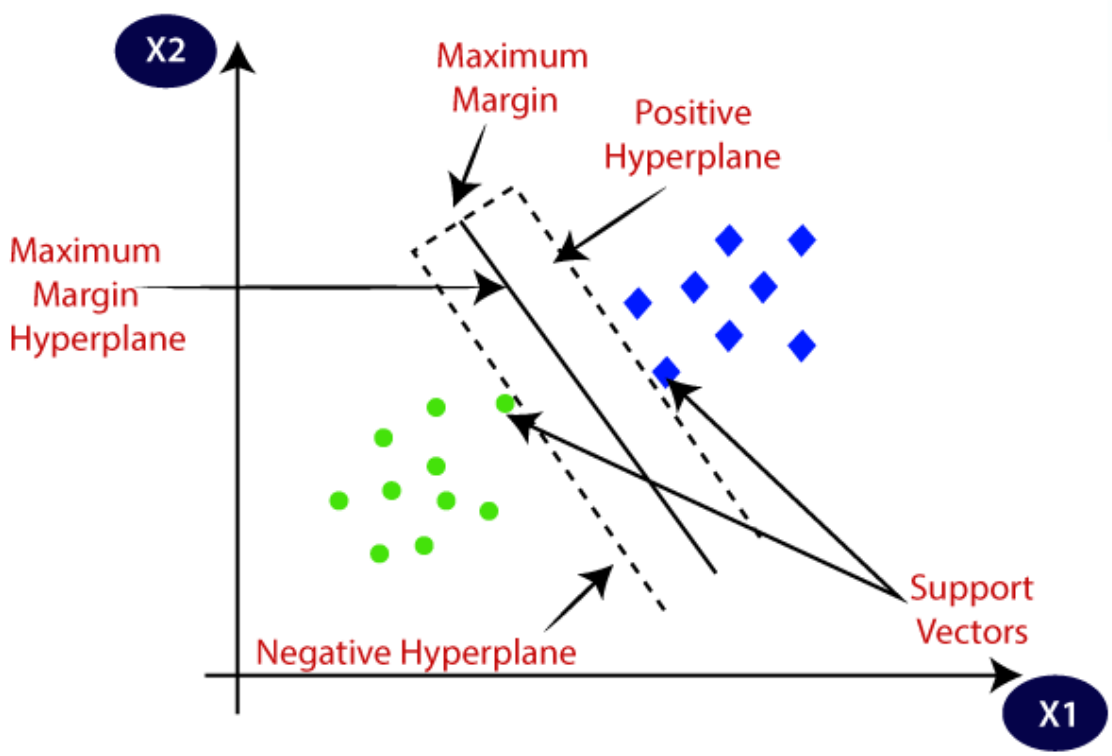


**Figure 2.4: A two-dimensional space hyperplane (Source: Ma and Guo, 2014)**

Solving the optimization problem allows for the classification of inputs depending on the side of the hyperplane on which they lie.

$$\underset{\beta_0,\beta_1,\ldots,\beta_p \epsilon_1,\ldots,\epsilon_n,M}{maximize} \quad M$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1, \tag{2.12}$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \geq M(1 - \epsilon_I),$$

$$\epsilon_i \geq 0, \ \sum_{i=1}^{n} \epsilon_i \leq C,$$

where $M$ is the margin's width; $\beta_0, \beta_1, \ldots \beta_P$ are maximal margin coefficients; and $\sum_{j=1}^{p} \beta_j^2 = 1$ is the constraint. $x_i$ and $y_i$ are the inputs and their associated labels, respectively. Additionally, $C$ is a regularization parameter, and $\epsilon_1, \ldots, \epsilon_n$ are variables that permit observations to fall on the wrong side of the margin or hyperplane. A large $C$ results

in a wide margin, low variance, and high bias, while a small *C* narrows the margin, lowering bias and increasing variance (James *et al*., 2013). Therefore, *C* is a positive tuning parameter that controls the trade-off between bias and variance.

For non-linear class boundaries, SVM uses kernels to enlarge the feature space. This applies to spectral data, which normally contains hundreds or thousands of columns. Adding a kernel function transforms the data into a higher-dimensional space, achieving basic segregation. The radial kernel is a commonly used kernel function for nonlinearly separable classes and takes the form shown in equation 2.13.

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2) \quad (2.13)$$

where γ, another crucial tuning parameter, is a positive constant.

SVR is based on the same working principles as SVM, with the primary difference being in the model inputs. The former uses continuous variables as inputs to predict continuous outputs. In contrast, categorical inputs are used in SVM to predict classes. SVM and SVR are known for their effectiveness in high-dimensional spaces, versatility, and capacity to model non-linear decision boundaries (Yang *et al*., 2022). Consequently, the present study adopts these ML models to solve multiclassification and regression problems.

### 2.5.3 Machine Learning Application in Spectroscopy

Traditionally, spectroscopic data interpretation entailed establishing one-to-one relationships between spectrum and structure and matching spectra to existing libraries. This process is effective when involving data with well-defined peaks but can be tedious, especially when dealing with the Raman spectra of complex samples, such as milk and

vegetables. Due to the heterogeneous nature of such complex samples, their Raman spectra are characterized by overlapping peaks that make their conventional interpretation difficult. ML has the capacity to handle complex spectral data, overcoming the limitations of univariate data analysis (Sarker, 2021).

A study by Du *et al*. (2020) used RF to analyze chlorpyrifos residues in pears using Raman spectra. The researchers observed that combining Raman spectroscopy with ML techniques, such as RF helps expedite the accurate detection of pesticide residues. On the other hand, Zhu *et al*. (2021) employed SVM in the multivariate analysis of SERS spectra, enabling the rapid detection of chlorpyrifos residues in tea. The detection took approximately 15 minutes and a t-test demonstrated no difference existed between the observed and predicted values. The results confirmed the method's rapidity and accuracy in pesticide detection. Lee *et al*. (2020) employed SERS combined with k-nearest neighbor (KNN), linear discriminant analysis (LDA), and partial least squares discriminant analysis (PLSDA) to classify animal feed samples and quantify the levels of chlorpyrifos and aldicarb. The study demonstrated the effectiveness of ML in classifying spectroscopic data and the possibility of the method as an effective and efficient analytical technique for analyzing pesticides in complex matrices.

Furthermore, Liu *et al*. (2016) used partial least squares (PLS) regression to model SERS data for the detection and quantification of pesticides (chlorpyrifos and phosmet) in fruit samples, with the optimal prediction model having a correlation coefficient (r) of 0.843 and a root mean square error of prediction (RMSEP) of 2.992 mg/L. The researchers

recommend further development of the quantitative method to increase accuracy based on the relatively poor results. The current study addresses this challenge by identifying the pesticide's Raman fingerprint and adopting more sophisticated ML models.

A study by Weng *et al*. (2019) used various models, including KNN, PLSDA, and SVM, to detect pesticide residues in paddy water. A 100% classification accuracy of Raman data was achieved using KNN, while the SVM provided optimal quantitative analysis results, with an RMSEP of 0.207 and a coefficient of determination ($r^2$) of 0.99952. The study concluded that using SERS with ML techniques can provide a simple and convenient approach to pesticide residue analysis. Similarly, Zhu *et al*. (2018) employed classification models, including KNN and the genetic algorithm-partial least squares (GA-PLS), to model Raman data, enabling the quantitative and qualitative detection of chlorpyrifos in tea. The models exhibited high prediction performances as reflected by $r^2$ values of between 0.96 and 0.98 as well as RMSEP values of 0.29 and 0.31. Consequently, the study highlights the power of ML models as effective analytical tools for chlorpyrifos residue analysis.

Moreover, Yazgan *et al*. (2019) utilized PLSDA in analyzing Raman data to discriminate milk samples based on their authenticity. The researchers were able to also discriminate the milk samples based on their species (ewe, goat, cow, and mixture) with a success rate of over 91.5%. Based on the results, the study emphasizes the utility of Raman spectroscopy coupled with models like PLSDA in the analysis of milk to determine its origin. The present study builds on this line of thought, employing ML models for pesticide detection and quantification in milk. The study further uses ML to process the Raman data from

chlorpyrifos-free (control) and chlorpyrifos-contaminated (treated) kale samples and predict the categories and contamination levels using unseen data.

### 2.5.4 Literature Gaps Addressed by the Current Study

Previous studies have demonstrated the effectiveness of conventional pesticide residue analysis techniques, such as GC-MS and HPLC. However, these methods require extensive sample preparation, hazardous solvents, and prolonged analysis times, making them impractical for rapid screening and large-scale monitoring (Bedi *et al*., 2018; Ling *et al*. 2011; Stocka *et al*., 2016). Other studies have explored spectroscopic methods like SERS for pesticide residue detection, highlighting its sensitivity and rapidity (Mikac *et al*., 2021; Tao *et al*., 2022). Despite these advantages, issues such as matrix interference and the high cost of conventional SERS substrates like gold and silver nanoparticles remain key limitations (Chen *et al*., 2023; Wang *et al*., 2024). Additionally, while machine learning models have been utilized in spectroscopy-based pesticide analysis, their application has largely involved using the entire spectral range, which often reduces efficiency and increases computational time and resource demands (Du *et al*., 2020; Zhu *et al*., 2021; Weng *et al*., 2019).

The current study directly addresses these gaps by employing Raman spectroscopy as a solvent-free and non-destructive technique for rapid pesticide detection in both milk and vegetables. Unlike conventional analytical methods, this approach eliminates the need for hazardous solvents and extensive sample preparation, making it more practical for routine monitoring. To further enhance sensitivity and reduce background noise, aluminum foil is introduced as a cost-effective and efficient SERS substrate, overcoming the high-cost

barrier associated with traditional metal nanostructures. Moreover, PCA is applied to identify a concise Raman fingerprint, which is then used for machine learning instead of the entire spectral range, thereby optimizing data interpretation and improving computational efficiency. The study also advances ML applications in pesticide detection by integrating RF and SVM/SVR models to classify and quantify chlorpyrifos contamination in both liquid (milk) and solid (kale) samples.

In summary, this work fills key information gaps in the literature by: (i) demonstrating a rapid, non-destructive, and solvent-free Raman-based approach for monitoring chlorpyrifos in two relevant food matrices (milk and kale); (ii) proposing a low-cost aluminum foil SERS strategy as an alternative to expensive noble metal substrates; (iii) introducing a fingerprint-based machine learning workflow that uses PCA-selected bands rather than the full spectrum; and (iv) systematically evaluating RF and SVM/SVR models for both classification and quantification of chlorpyrifos levels. These contributions extend current knowledge on spectroscopy and machine learning approaches for pesticide residue analysis and provide a framework that can be adapted to other pesticide-matrix combinations.

# CHAPTER THREE

# MATERIALS AND METHODS

## 3.1 Pesticide and Sample Collection

Chemical-free milk and uncontaminated kale leaves were sourced from Back to Nature Organic Farm (see Appendix II) and transported to the laboratory. The milk was collected in clean, sterilized bottles and sealed before being placed in a refrigerator to avoid contamination. Similarly, kale samples were packed in sealed plastic bags and stored in the crisper to maintain an optimal humidity level and freshness before analysis.

The pesticides used in this study were Duodip and Ranger (see appendix III). Duodip was chosen for spiking milk because it is a commonly used acaricide by Kenyan dairy farmers to control ticks in cattle due to its availability. It is applied by spraying or dipping in areas prescribed by the directorate of veterinary services. Duodip contains 50% chlorpyrifos and 5% cypermethrin. On the other hand, Ranger has its active ingredient as chlorpyrifos (48%) and was chosen for spiking kale samples because it is used to control a wide range of pests, including aphids, termites, whiteflies, berry borer, and thrips. Notably, kale leaves are affected by aphids and whiteflies, prompting farmers to use Ranger on this class of vegetables despite the manufacturer's target crops being coffee and roses. The two pesticides were obtained from a retail outlet and transported to the laboratory to prepare different concentrations. Using commercially available pesticides enabled us to simulate the actual process carried out by farmers.

### 3.2 Sample Preparation for Chlorpyrifos Treatment in Milk and Kale Samples

The milk and kale samples were divided into two groups: the first group was treated with varying concentrations of chlorpyrifos, while the second was used as the control group. The range of chlorpyrifos concentrations was based on the established MRL (0.01 ppm) and random concentrations above and below the value were generated. The concentrations were also restricted to previously detected chlorpyrifos levels in milk and vegetables, with the highest value of 5ppm and the lowest of 0.003ppm (Esturk *et al*., 2014; Vemuri, 2016). The dilution process used to prepare the concentrations was guided by equation 3.1.

$$C_1V_1 = C_2V_2 \tag{3.1}$$

where $C_1$ and $C_2$ are the initial and final concentrations, respectively, and $V_1$ and $V_2$ represent the initial and final volumes, respectively.

In the case of milk, sample preparation included filtering to remove physical impurities and large molecules that could interfere with the Raman signal. The milk was then divided into two groups, the first for treated samples and the second for the control group. Duodip was added to a portion of the first milk sample to prepare a stock solution (1000 ppm) according to the instructions of the pesticide manufacturer. From the stock solution, equation 3.1 was applied to prepare 21 concentrations labeled using A, B, C, ... up to U, as listed in Table 3.1. The prepared solutions were vigorously shaken for maximum homogenization based on the manufacturers' recommendations. The second group of milk was left untreated for control purposes. Figure 3.1 shows the prepared milk samples for analysis.


**Figure 3.1: Milk samples treated with varying levels of chlorpyrifos**

Similarly, kale samples were grouped into two: one group was treated with chlorpyrifos concentrations, and the remaining kale leaves were used as control samples. Following the steps used for milk, Ranger was used to prepare 21 concentrations using distilled water for application on kale leaves. Each solution was labeled using A, B, C ... up to U. Vigorous shaking was also done to ensure maximum homogenization. For convenience, kale samples were cut into small pieces using a scalpel and mounted on a flat surface, as shown in Figure 3.2. Using a piece of clean cotton wool, each of the 21 pesticide concentrations was carefully applied to one of the kale leaf cuttings, which was pinned to a fixed position. After application, the treated leaves were left to stand for about 5 minutes to allow the applied chlorpyrifos solution to dry on the leaf surface at room temperature before spectral measurements were taken. High-concentration spiking solutions (1000 ppm) for milk and kale samples were also prepared.

**Figure 3.2: Kale samples containing different concentrations of chlorpyrifos. Spacing between the samples was carefully chosen to prevent the different chlorpyrifos concentrations applied from interfering with one another.**

Notably, all precautions for using pesticides were observed during the experiments, including wearing protective clothing such as a face mask and rubber gloves, handwashing, and proper waste disposal. Moreover, the experiments were conducted in a well-ventilated area to minimize the risk of inhalation.

**Table 3. 1 Concentrations of chlorpyrifos prepared for milk and kale samples**

| Sample label | Concentration (ppm) of solution for kale/milk | Above/Below MRL |
|---|---|---|
| A | 5 | Above |
| B | 4.1 | Above |
| C | 3 | Above |
| D | 2.5 | Above |
| E | 1 | Above |
| F | 0.8 | Above |
| G | 0.7 | Above |
| H | 0.5 | Above |
| I | 0.3 | Above |
| J | 0.1 | Above |
| K | 0.06 | Above |
| L | 0.03 | Above |
| M | 0.02 | Above |
| N | 0.015 | Above |
| O | 0.01 | MRL |
| P | 0.009 | Below |
| Q | 0.008 | Below |
| R | 0.007 | Below |
| S | 0.006 | Below |
| T | 0.0045 | Below |
| U | 0.003 | Below |

### 3.3 Raman Spectra Acquisition

The Raman spectra were acquired at room temperature and in a dark room (to avoid ambient lighting) using a portable Raman spectrometer (EZRaman- NP-785, Enwave Optronics, USA) with a laser excitation wavelength of 785 nm. To optimize the method for

reliable Raman measurements, several adjustments were made. First, the laser power was varied to identify the optimal level that provided sufficient spectral intensity while avoiding sample degradation. Excessive laser power can destroy sensitive samples (Yazdanpanah *et al*., 2024), hence the need to optimize laser power. The laser was set to 300 mW as better-quality Raman spectra were generated using this laser power than when using lower laser powers. Additionally, the laser power did not cause any damage, such as burning, to the samples.

The distance between the instrument's fiber-optic probe end and the sample surface was varied during preliminary experimental trials, with the optimal working distance being established at approximately 7 mm. This working distance was adopted and utilized throughout the study. Optimal focusing ensured effective interaction between the laser and the samples, enhancing signal quality. Since the Raman measurements were taken over multiple days, calibrating the spectrometer was crucial to ensure spectral accuracy. A polystyrene reference standard with a known characteristic peak at 1000 $cm^{-1}$ was used for calibration before collecting any measurements. This procedure minimized the risk of axis drift, which could otherwise lead to misaligned spectral peaks, affecting the reproducibility of the results (Rusu and Baia, 2023). A 200 – 2000 $cm^{-1}$ spectral coverage was chosen since it is considered sufficient for the Raman identification of biological samples (Bumbrah and Sharma, 2016; He *et al*., 2020; Yang, 2022). Finally, laser safety goggles were worn throughout the experiments to protect against accidental exposure to stray laser beams.

Raman spectroscopy measurements were collected in two phases. The first phase included milk as the sample, while the second used kale leaves. For milk spectral data collection, the experiment used a clean aluminum foil to amplify the signal (Ondieki *et al*., 2023; Sultangaziyev *et al*., 2020). The aluminum foil spectrum was collected and saved for background subtraction. Drops of about 50µL of the control and spiked milk samples were carefully smeared on an aluminum-wrapped glass slide (as shown in Figure 3.3 (a)), and 100 spectra were acquired from 20 random spots on five portions of each sample. The integration time (time duration of each measurement) was set to 10 seconds, and an average of 10 consecutive scans was collected and smoothed with a boxcar (a smoothing window used to reduce noise) of 1 scan. It took approximately 1 minute and 43 seconds to obtain a single spectrum.

For kale samples, the dried leaf cuttings were placed directly on the Raman instrument's solid sample holder for spectra acquisition (Figure 3.3 (b)). The Raman spectra of the kale samples were also acquired over the wavelength range of 200 – 2000 cm-1 using a 785nm laser. The integration time was set to 5 seconds, and each sample spectrum was averaged over five consecutive scans and smoothed with a boxcar of 5 scans. It took approximately 30 seconds to obtain a single spectrum for each kale sample (representing a particular chlorpyrifos concentration). 20 random spots on five leaf cuttings of each sample were acquired, totaling 100 spectra for each concentration.

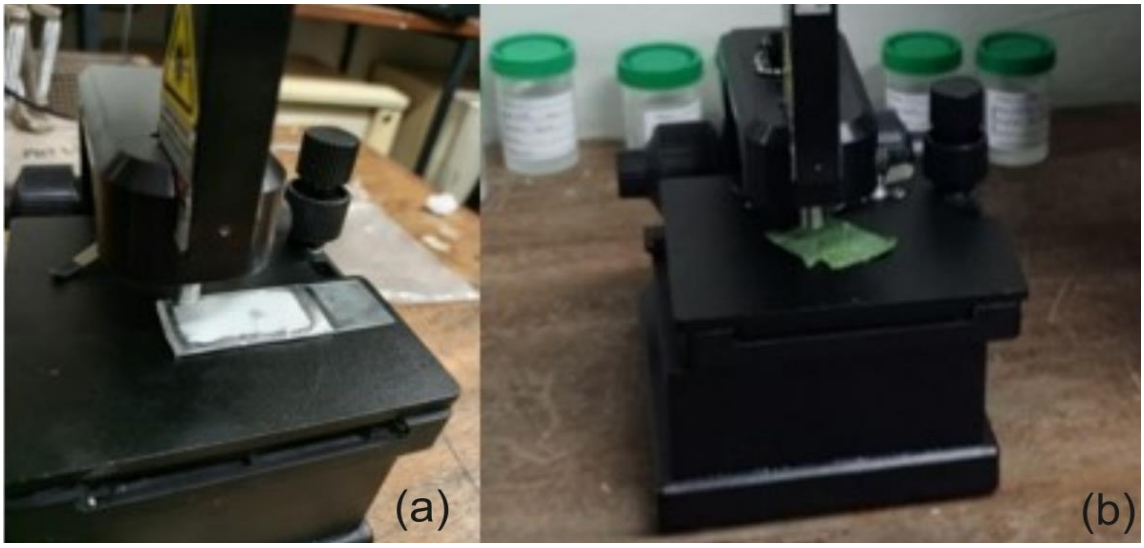


**Figure 3.3: Photograph of how milk (a) and d kale (b) samples were mounted on the Raman spectrometer's sample holder**

### 3.4 Spectra Data Preprocessing

Raw Raman spectra from biological samples, such as those utilized in this study, are prone to fluorescence background contributions, which are often extracted through data preprocessing (Bocklitz *et al*., 2011). The current study's raw spectral data underwent baseline correction to eliminate baseline drift. The Raman spectrometer had built-in parameters that allowed it to fit a least squares line to each of the spectrum's regions, subtract the spectrum, and move the baseline close to the zero line. These operations enabled the instrument to perform baseline correction, automatically eliminating the broad fluorescence baseline.

Smoothing of the spectral data followed, which was realized using a third-order Savitzky-Golay algorithm in spectragryph software (Menges, 2017). Smoothing removed noisy artefacts in the Raman spectra. Finally, normalization was done using Origin software (OriginPro, 2024). Normalization between 0 and 1 corrected intensity level disparities in

the data, ensuring that the Raman bands of a particular sample (concentration) were as similar as possible.

### 3.5 ANOVA of Spectra from Milk and Kale

Using the Data Analysis ToolPak in MS Excel, one-way ANOVA (Ostertagova and Ostertag, 2013) was performed on the average preprocessed spectra of the control and treated (1000 ppm) milk and kale samples to identify the spectral regions with the highest variation. The spectra of the control and treated samples were transferred to an Excel sheet, and the mean intensity at each wavenumber was calculated for each group. One-way ANOVA was then applied at each wavenumber to test for statistically significant differences in the mean intensities of the control and treated samples. Wavenumbers with p-values below 0.05 were considered statistically significant and indicative of sensitivity to the presence of chlorpyrifos (Thuku *et al*., 2025b). The resulting variance profile was plotted as a function of wavenumber to visualize the regions showing the largest differences, which were used to identify the spectral regions most informative for the subsequent exploratory analysis.

### 3.6 Exploratory Analysis of Raman Spectra using PCA

As explained in section 2.4, PCA uses an orthogonal transformation to convert a set of linearly correlated variables into a set of nonlinearly correlated variables, expressed as percentage variances (PC scores). The method retains the most essential features directly related to the outcome of interest while eliminating noise, redundancy, and irrelevant features. The results are represented on a set of orthogonal axes known as principal components, with PC1 describing the greatest variance, PC2 the second-highest, and so on. Guided by the regions with high variance, as revealed by ANOVA, an exploratory analysis

was done using PCA in open-source R (R Core Team, 2024) and the Chemospec package (Hanson, 2024). In this study, PCA helped select the most influential band (fingerprint) with the highest explained variance due to the pesticide and the best separation between control and treated sample groups. Additionally, PCA was used to reduce the dimensionality of the pre-processed Raman data, detect outliers, visualize the data, and extract PCs, which served as inputs to the ML models. Using PCs instead of wavelengths reduced the training time and accuracy of the developed models. A detailed description of the commands and algorithms used for the PCA can be found in Appendix I.

### 3.7 Application and Evaluation of Machine Learning Algorithms

The current study used two ML models, RF and SVM/SVR, to classify samples into three categories (Below MRL, MRL, and Above MRL) and quantify chlorpyrifos levels. Ideally, a fit for all ML model does not exist, and for this reason, this study employed two models for classification and regression tasks and compared their performance across each sample type. The PC data from the identified fingerprint were split into a training set (70%) for model calibration and a test set (30%) for evaluating the classification and regression models. After training the models with different hyperparameter values, the hyperparameter settings that yielded the best predictive performance were selected and reported.

#### 3.7.1 Application of Random Forest in Spectra Classification and Chlorpyrifos Quantification

Random forest employs decision trees to model both linear and non-linear data. This study used random forests to develop classification and regression models for milk and kale samples with varying concentrations of chlorpyrifos, using PC data from their Raman

spectra. The PCs served as the predictors, while the associated categories and real values represented the target labels in classification and regression problems, respectively. For classification tasks, chlorpyrifos concentrations were categorized as below MRL, MRL, or above MRL, yielding three classification categories. On the other hand, regression problems considered samples with concentrations between 0.003 and 5 ppm, as listed in Table 3.1.

The PC data was split into a training set and a test set. Bootstrapping (sampling with replacement) was used to select samples randomly and use them to construct classification and regression trees (Hayes *et al*., 2015). The splitting criteria (as discussed in section 2.5.1) were repeatedly applied to build the trees until each selected sample was classified in each tree, assigning each sample to its corresponding output. *K* independent trees were constructed by repeating the tree-building steps, forming the RF model. For classification, the model output for each sample was the class selected by most trees, i.e., the most frequent categorical variable. In contrast, the average across all independent trees was the regression model's output. The models were run several times with different numbers of PCs to determine the number that yields optimal classification and quantification performance.

To avoid overfitting of the RF model, different cross-validation parameters were used. One of these parameters was the number of variables selected at each split (*mtry*). When building trees in an RF, *m* predictors are typically randomly selected as split candidates from the entire set of *p* predictors (James *et al*., 2013). The number of predictors is chosen

so that m ≈ √p, which helps decorrelate the trees and makes the final average less variable and more reliable (James *et al*., 2013). Additionally, the number of trees to grow (*ntree*) is another important parameter used in this study to optimize the RF model's performance. Finally, a confusion matrix was used to determine the model's performance in the case of classification tasks, while regression performance was determined using $R^2$ and RMSEP. Appendix I provides a detailed description of the commands and algorithms used for the RF.

### 3.7.2 Application of SVM and SVR in Spectra Classification and Chlorpyrifos Quantification

This study also used a support vector machine and a support vector regression model to predict the class/category to which each test sample belongs, as well as the concentration of each sample. The PC data of the samples' spectra were extracted and split into a training and a test set before feeding the data into the SVM and SVR models (Yang *et al*., 2022). Model tuning was done by considering a range of cost (*C*) and gamma (γ) values (Adugna *et al*., 2022; Lorena and de Carvalho, 2008). The best combination of the two hyperparameters, which yielded the optimal hyperplanes, was adopted. Similarly, different SVM kernels (linear, polynomial, and radial basis function (RBF)) were used, and the best performing was adopted. The ability of the developed model to correctly classify samples according to their chlorpyrifos levels (with respect to the MRL) was crucial, as it indicated the practicability of the developed methods in the rapid identification of milk and kale samples with chlorpyrifos above the established MRL. A detailed description of the commands and algorithms used for SVM and SVR is provided in Appendix I.

### 3.7.3 Model Performance Evaluation Metrics

Several evaluation metrics, including accuracy, precision, recall, and Cohen's Kappa (Vujović, 2021), were used to assess the performance of the classification models developed in the current study. Each metric provided a different insight into the model's performance, ensuring a more comprehensive evaluation. These metrics were derived from the confusion matrix, a table that defines and summarizes a classification model's performance (Vujović, 2021). The present study handles a multi-class classification problem, with the confusion matrix including three classes: below MRL, MRL, and above MRL. All three classes provided values for True Positives (TP), False Positives (FP), and False Negatives (FN), as shown in Table 3.2. While TP refers to the number of correctly classified samples, FP is the number of misclassifications.

**Table 3. 2 Confusion matrix for this study's multi-class classification**

| | | **Actual** | | |
|---|---|---|---|---|
| | Group | Below MRL | MRL | Above MRL |
| **Predicted** | Below MRL | $TP_1$ | $FN_{12}$ | $FN_{13}$ |
| | MRL | $FP_{21}$ | $TP_2$ | $FN_{23}$ |
| | Above MRL | $FP_{31}$ | $FP_{32}$ | $TP_3$ |

$TP_1$, $TP_2$, and $TP_3$ represent the true positives for each group. $FP_{21}$, $FP_{31}$, and $FP_{32}$ represent false positives (instances from another class misclassified into a particular class), while $FN_{12}$, $FN_{13}$, and $FN_{23}$ represent false negatives (instances from one class misclassified into another). A false positive instance is known as a Type 1 Error, while a false negative refers to a Type 2 Error (Vujović, 2021).

In a classification problem, accuracy represents the proportion of correct predictions among the model's total predictions. Thus, accuracy in the current study was the total number of correct predictions across the three classes divided by the total number of predictions (Vujović, 2021), calculated as:

$$Accuracy = \frac{TP_1 + TP_2 + TP_3}{TP_1 + TP_2 + TP_3 + FP_1 + FP_2 + FP_3 + FN_1 + FN_2 + FN_3} \quad (3.2)$$

Notably, using accuracy alone to assess a model's performance can be misleading, especially in cases where the datasets are imbalanced. Precision was calculated for each group of chlorpyrifos concentration, and it expressed the proportion of samples predicted correctly for a given group divided by all the predictions belonging to that group, as shown in the expressions 3.3, 3.4, and 3.5.

$$Precision\ for\ Below\ MRL = \frac{TP_1}{TP_1 + FP_2 + FP_3} \quad (3.3)$$

$$Precision\ for\ \ MRL = \frac{TP_2}{TP_2 + FP_1 + FP_3} \quad (3.4)$$

$$Precision\ for\ Above\ MRL = \frac{TP_3}{TP_3 + FP_1 + FP_2} \quad (3.5)$$

Achieving a high precision value indicated that the model did not repeatedly classify other groups into the target group. Similarly, recall, also called sensitivity, was calculated for each group. This metric measured the proportion of samples predicted correctly for a particular group in relation to all the actual samples of that group. Recall for the three groups were defined as expressed in 3.6, 3.7, and 3.8.

$$Recall\ for\ Below\ MRL = \frac{TP_1}{TP_1 + FN_1} \quad (3.6)$$

$$Recall\ for\ MRL = \frac{TP_2}{TP_2 + FN_2} \quad (3.7)$$

$$Recall\ for\ Above\ MRL = \frac{TP_3}{TP_3+FN_3} \tag{3.8}$$

A high recall value indicated that the model correctly predicted most of the samples for a specific group. Finally, Cohen's Kappa was used to measure the agreement between the actual and predicted groups (Vujović, 2021). It is defined as:

$$K = \frac{P_o-P_e}{1-P_e} \tag{3.9}$$

where $P_o$ is the observed accuracy and $P_e$ represents the model's random accuracy. A Kappa value of zero indicates no agreement between the actual and predicted groups, while a 1 indicates complete agreement (McHugh, 2012; Warrens, 2011).

In the case of the regression models built in the current study the predictive performance was evaluated using the Root Mean Square Error of Prediction (RMSEP) and the coefficient of determination ($R^2$). The RMSEP identified the optimum model and was determined using the expression:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \tag{3.10}$$

where $y_i$ is the model's predicted concentration, $\hat{y}_i$ is the actual concentration and $n$ represents the number of milk and kale test samples. On the other hand, $R^2$ was used to determine the proportion of total variance explained by the model, with the remaining variance accounting for the errors. $R^2$ was calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{3.11}$$

where $y_i$ and $y_i$ hold the same meaning as provided in equation (3.10). $\bar{y}$ is the actual mean concentration. Figure 3.4 shows a schematic overview of this study's research design and methodology.

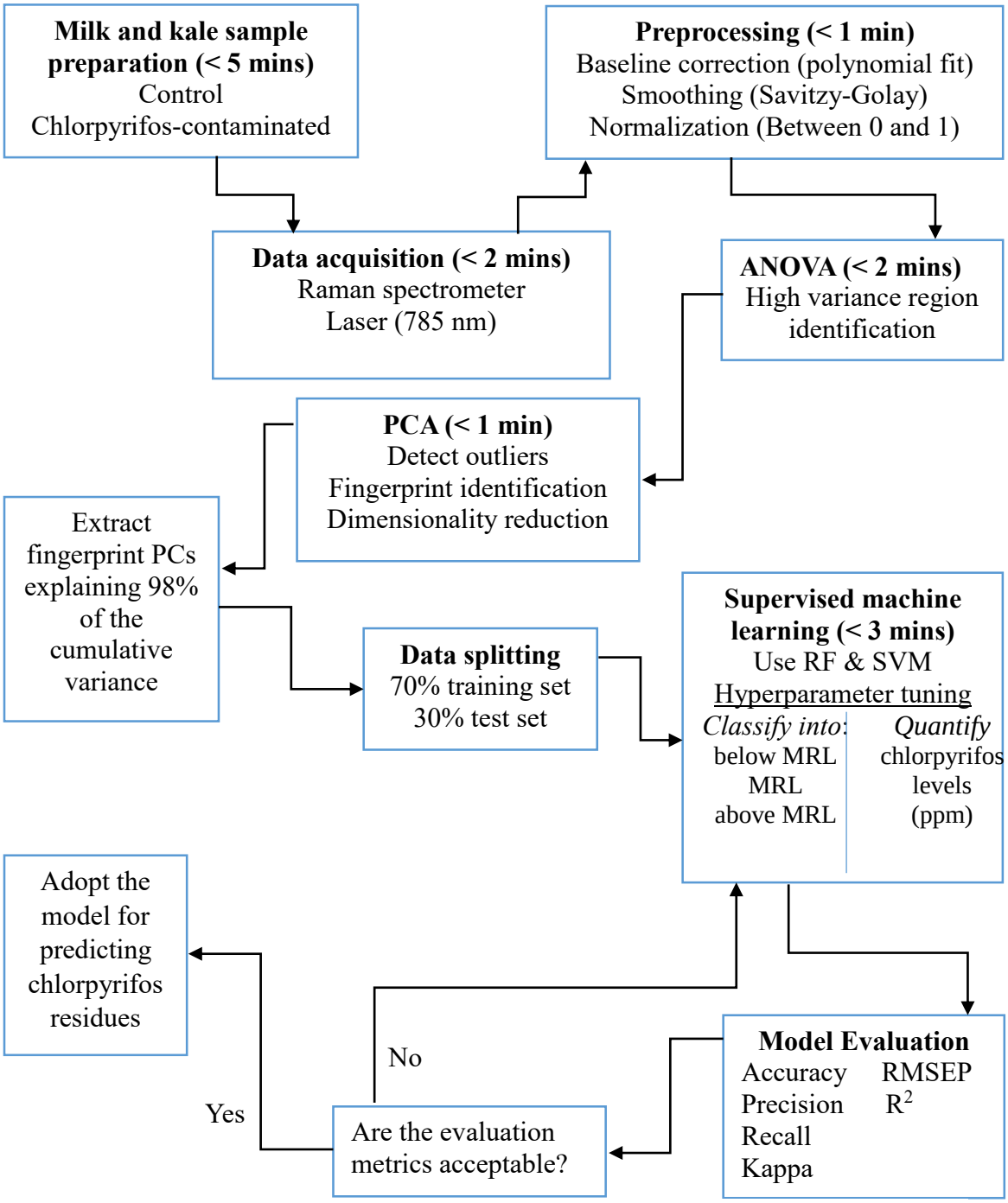


**Figure 3.4: A flowchart of the main data collection and analysis steps used in this study**

### 3.7.4 Determination of Limits of Detection and Limits of Quantification

The limit of detection (LOD) and limit of quantification (LOQ) were determined as the lowest chlorpyrifos concentrations the ML models could detect and quantify, respectively. By definition, the LOD is the minimum concentration of an analyte that can be reliably detected with a certain degree of confidence (Boqué and Heyden, 2009). In contrast, the LOQ is the lowest analyte concentration that can be reliably quantified with established precision, accuracy, and uncertainty (Bayona and Pawliszyn, 2012). For univariate calibration, the LOD and LOQ can easily be extracted from the univariate calibration line (Allegrini and Olivieri, 2014). However, the current study utilizes Raman spectra, which are multivariate and characterized by redundant and overlapped bands. As such, univariate approaches to determining the two metrics cannot be used as they do not sufficiently account for the multiple spectral responses, including noise.

Consequently, this work adopted a multivariate approach to evaluating the LOD and LOQ. To achieve this, the calibration curves of the built models were used, which included a plot of the predicted chlorpyrifos concentrations against the samples' actual concentrations, as recorded in Table 3.1. This is in agreement with the pseudo-univariate LOD concept. According to this approach, a pseudo-univariate calibration graph can be created by plotting the estimated analyte concentrations against their measured or nominal concentrations (Allegrini and Olivieri, 2014). In the pseudo-univariate approach, the LOD and LOQ are calculated using equations 3.12 and 3.13, respectively (Ahuja, 2005; Oleneva *et al*., 2019)

$$LOD = \frac{3\sigma}{S} \tag{3.12}$$

$$LOQ = \frac{10\sigma}{S} \tag{3.13}$$

where $\sigma$ is the standard deviation of the y-intercepts of the regression lines and $S$ is the calibration curve's slope.

# CHAPTER FOUR

# RESULTS AND DISCUSSION

## 4.1 Raman Spectra of Control Samples

Since the samples used in this study have complex chemical compositions, it is important to know the chemical components of milk and kale detected by Raman spectroscopy. This information is crucial in interpreting the Raman bands that distinguish treated from control samples. Using Raman spectroscopy, the chemical characterization of milk and kale leaves was possible, revealing the present biochemical compounds such as amino acids, mono and oligosaccharides, carotenoids, fatty acids, and flavonoids.

### 4.1.1 Characteristic Raman Spectra of Milk

As a complex food matrix, milk contains a wide variety of biological molecules, including primary metabolites like lipids, proteins, and carbohydrates. Figure. 4.1 shows the average Raman spectrum of the chlorpyrifos-free milk sample. The Raman bands in the milk spectrum appeared at 354, 444, 516, 644, 724, 872, 1080, 1120, 1264, 1300, 1362, 1440, 1652, and 1742 $cm^{-1}$.

The peaks at 354 and 444 $cm^{-1}$ were related to the C-C stretching of lactose, while the peak at 516 $cm^{-1}$ was attributed to glucose (Khan *et al*., 2023; Li *et al*., 2024). The peak at 644 $cm^{-1}$ was associated with tyrosine, while protein's *ν*(C-S) gave rise to the peak at 724 $cm^{-1}$ (Kuhar *et al*., 2018; Li *et al*., 2024). Phospholipids in milk led to the observed peak at 872 $cm^{-1,}$ while *ν*(C-C) in lactose and fat was responsible for the Raman peaks at 1080 and 1120 $cm^{-1}$(Amjad *et al*., 2018; Khan *et al*., 2023; Silva *et al*., 2021). Further, the peaks at 1264,

1300, and 1362 $cm^{-1}$ were attributed to protein's amide III, $\delta(CH_2)$ twisting of fat, and tryptophan, respectively (Amjad *et al*., 2018; Khan *et al*., 2023; Yazgan *et al*., 2020). Finally, milk fat was responsible for the peaks at 1440, 1652, and 1742 cm-1 due to $\delta(CH_2)$ scissoring, *v*(C=C) and *v*(C=O), respectively (Khan *et al*., 2023; Silva *et al*., 2021).

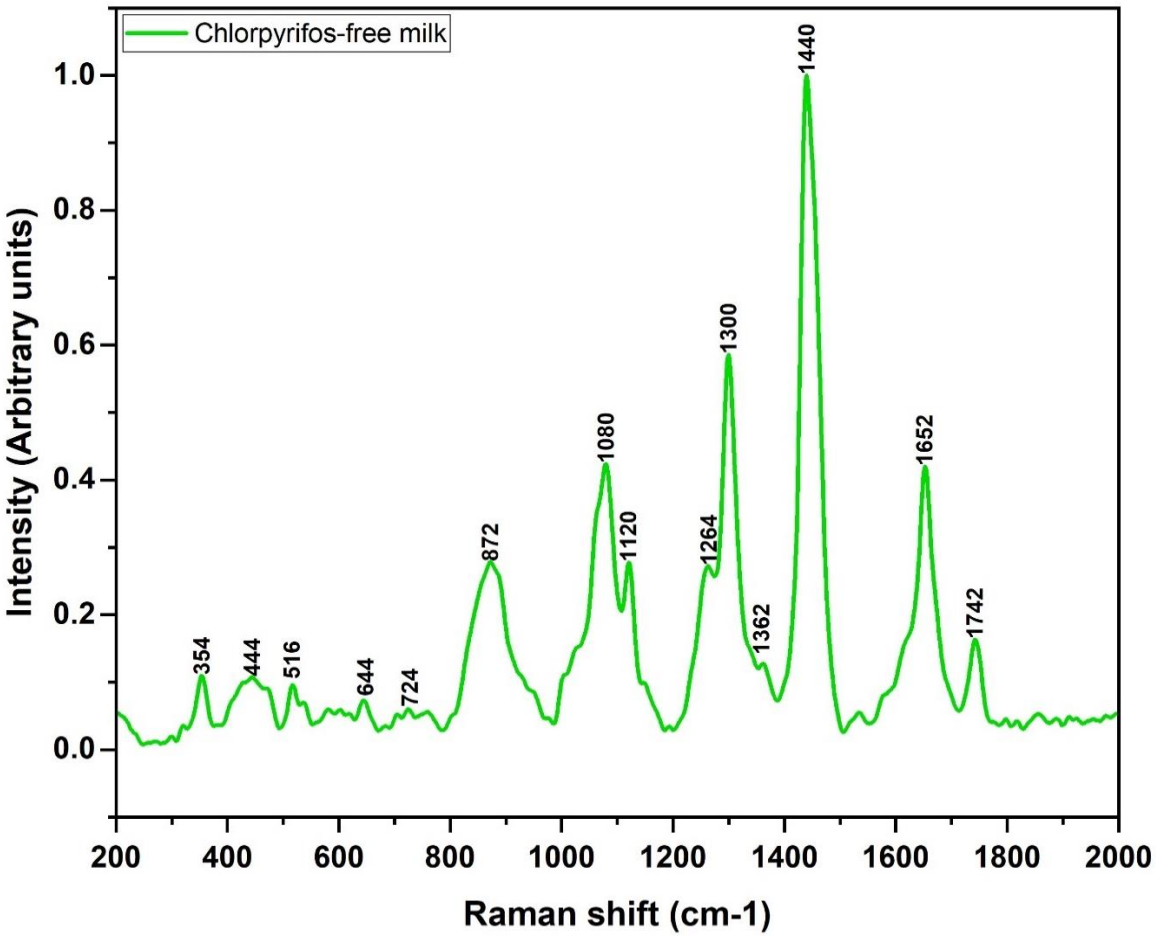


**Figure 4.1: Raman spectrum of chlorpyrifos-free milk**

Therefore, natural bovine milk consists of multiple compounds that are primarily made up of hydrogen, nitrogen, phosphorus, sulfur, and oxygen. Precisely, sulfur is contained in milk's sulfur-containing amino acids, such as methionine and cysteine, while phosphorous occurs in phospholipids (Landi *et al*., 2021; Silva *et al*., 2021). Additionally, Oxygen and

nitrogen are present in fatty acid esters and amides, respectively (Amores and Virto, 2019; Grewal *et al*., 2018).

### 4.1.2 Characteristic Raman Spectra of Kale

Akin to milk, kale leaves are biological samples containing numerous primary and secondary metabolites. The Raman spectrum of chlorpyrifos-free kale is shown in Figure 4.2.

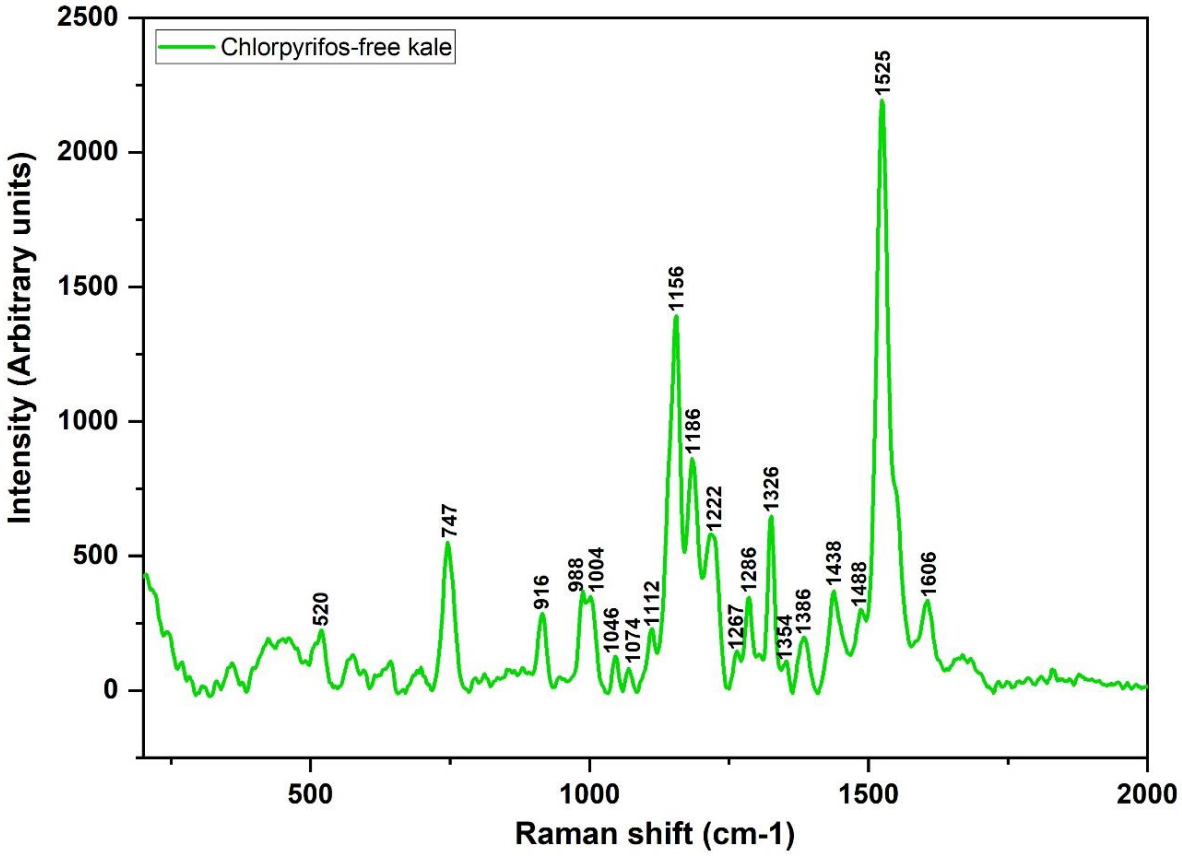


**Figure 4.2: Raman spectrum of chlorpyrifos-free kale leaves**

The peaks at 520 and 747 $cm^{-1}$ were due to ν(C–O–C) glycosidic and γ(C–O–H) of COOH vibrational models of cellulose and pectin, respectively (Edwards *et al*., 1997; Synytsya *et al*., 2003). N-C-C and C-C-C deformations of chlorophyll gave rise to the peak at 916 cm-1, while phenol compounds were responsible for the peak at 988 $cm^{-1}$ (Andreev *et al*., 2001;

Chen *et al*., 2004). As a result of ν3 (C–$CH_3$ stretching) phenylalanine ring stretching mode, a peak was observed at 1004 $cm^{-1}$. The peaks at 1046, 1076, 1112, and 1156 $cm^{-1}$ were assigned to $NO_3$ stretching, C-S stretching vibrations, carotenoids, and C–C stretching; ν(C–O–C), respectively (Baranska *et al*., 2005; Gupta *et al*., 2020; Sun *et al*., 2021). Additionally, the peaks at 1186, 1222, 1267, 1286, and 1326 $cm^{-1}$ were attributed to ***β***-carotene, C-H deformation, C–O stretching (aromatic) of Phenylpropanoids, $CH_2$ or $CH_3$ vibrations of aliphatic groups, and δ$CH_2$ bending vibration of Cellulose, respectively (Cao *et al*., 2006; Dhanani *et al*., 2022; Edwards *et al*., 1997; Frąszczak *et al*., 2023; Jacob *et al*., 2022). Furthermore, the peaks at 1354, 1386, and 1438 $cm^{-1}$ were due to carbohydrates, chlorophyll, and aliphatic δ($CH_2$) + δ($CH_3$), respectively while 1488, 1525, and 1606 $cm^{-1}$ were correlated with aliphatic δ$CH_2$ bending vibration, C=C stretching vibration (ν1) of carotenoids, and ν(C–C) aromatic ring of lignin (Barron *et al*., 2006; Devitt *et al*., 2018; Gierlinger *et al*., 2012; Gierlinger and Schwanninger, 2006; Sanchez *et al*., 2019; Schulz and Baranska, 2007; Yu *et al*., 2007).

Notably, kale leaves also contain nitrogen, sulfur, and oxygen, in addition to hydrogen and carbon. They also contain phosphorus, one of the minerals supplied by green leafy vegetables (Opazo-Navarrete *et al*., 2021).

**4.2 ANOVA of Raman Spectra and Chlorpyrifos Peak Assignment**

Analysis of Variance was used to assess the differences between the spectra of control milk and kale samples and those of the samples spiked with chlorpyrifos. However, the mean Raman spectra of chlorpyrifos-free milk and that of spiked milk were similar at low concentrations. The average spectrum of control milk samples was plotted along with the

average of samples spiked with chlorpyrifos concentrations: below MRL (0.003 ppm), MRL (0.01 ppm), and above MRL (5 ppm), the highest concentration prepared, as shown in Figure 4.3.

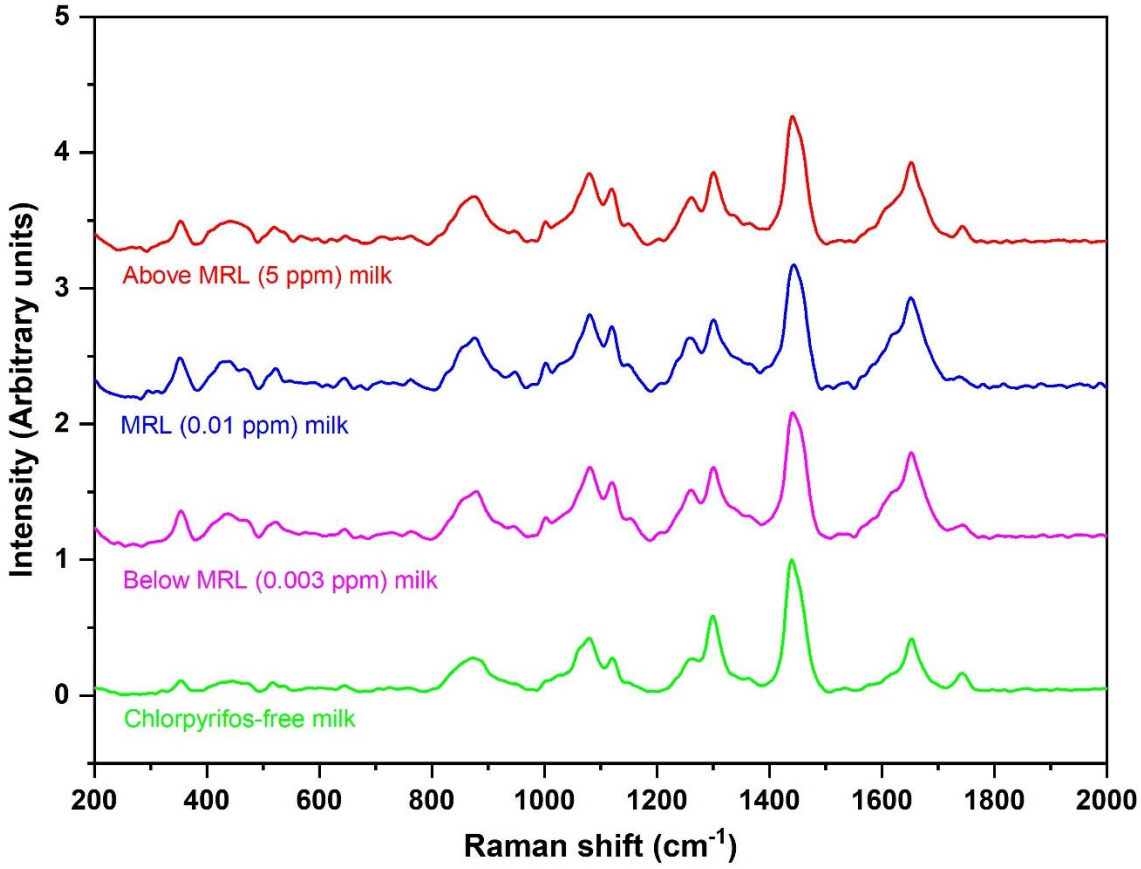


**Figure 4.3: Plot showing the common peaks in the averaged Raman spectra of milk samples at low concentrations**.

Further, a comparative analysis was done to differentiate the control kale samples from those containing low-level chlorpyrifos (0.003, 0.01, and 5 ppm). Similar results were obtained for kale samples, as shown in Figure 4.4. Notably, all the mean spectra contained common peaks. This implies that the chlorpyrifos molecules at low concentrations produced weak signals obscured by the samples' natural components, such as lipids, carbohydrates, and proteins. Therefore, the Raman spectra of milk and kale control samples

looked very similar to those from low chlorpyrifos-level treated samples and thus could not provide sufficient information about the existing spectral differences.

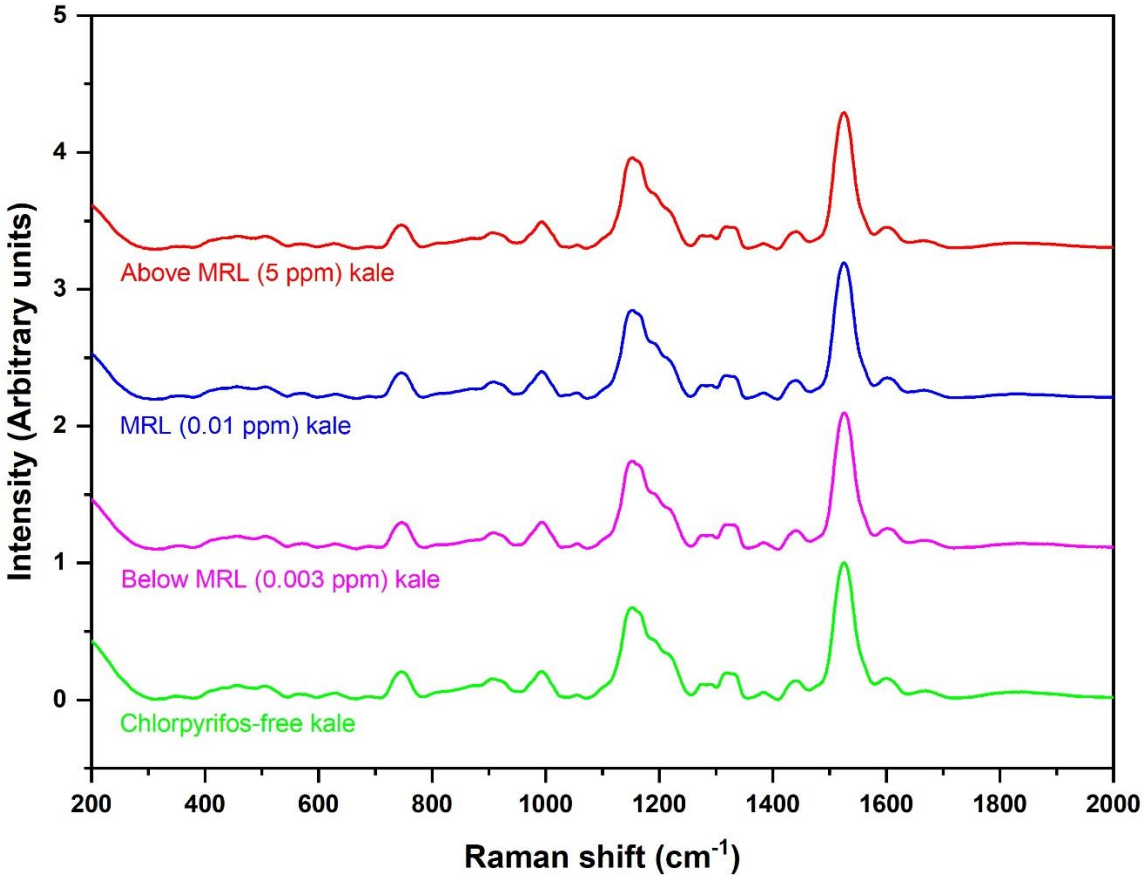


**Figure 4.4: Plot showing the common peaks in the averaged Raman spectra of kale samples at low concentrations.**

It follows that visually discriminating milk and kale samples' Raman spectra was implausible at low chlorpyrifos concentrations. Theoretically, the intensity of the Raman signal increases with an increase in the analyte's concentration. Consequently, spectra from samples spiked with the stock solution (1000 ppm) were helpful for ANOVA. At high concentrations, chlorpyrifos's Raman signal was significantly increased, making the peaks easily noticeable, and the corresponding regions with significant differences between the mean spectra of control and treated samples were evident. The high chlorpyrifos

concentration amplified the molecule's vibrational modes, allowing apparent spectral differences between the control and treated samples. The standard error of the difference was determined and used to evaluate the samples' variability (Ikedi *et al*., 2023). The mean spectra of control and treated samples, along with the variance from ANOVA, were plotted to identify the bands exhibiting significant variance. It turns out that the bands with significant variance could be correlated with a corresponding chlorpyrifos characteristic peak, which was responsible for the difference.

### 4.2.1 Statistically Significant Bands in Milk Samples

As depicted in Figure 4.5, the regions of Raman milk spectra with significant variance were identified using ANOVA. These regions served as potential bands that can be used to detect chlorpyrifos residues in milk.

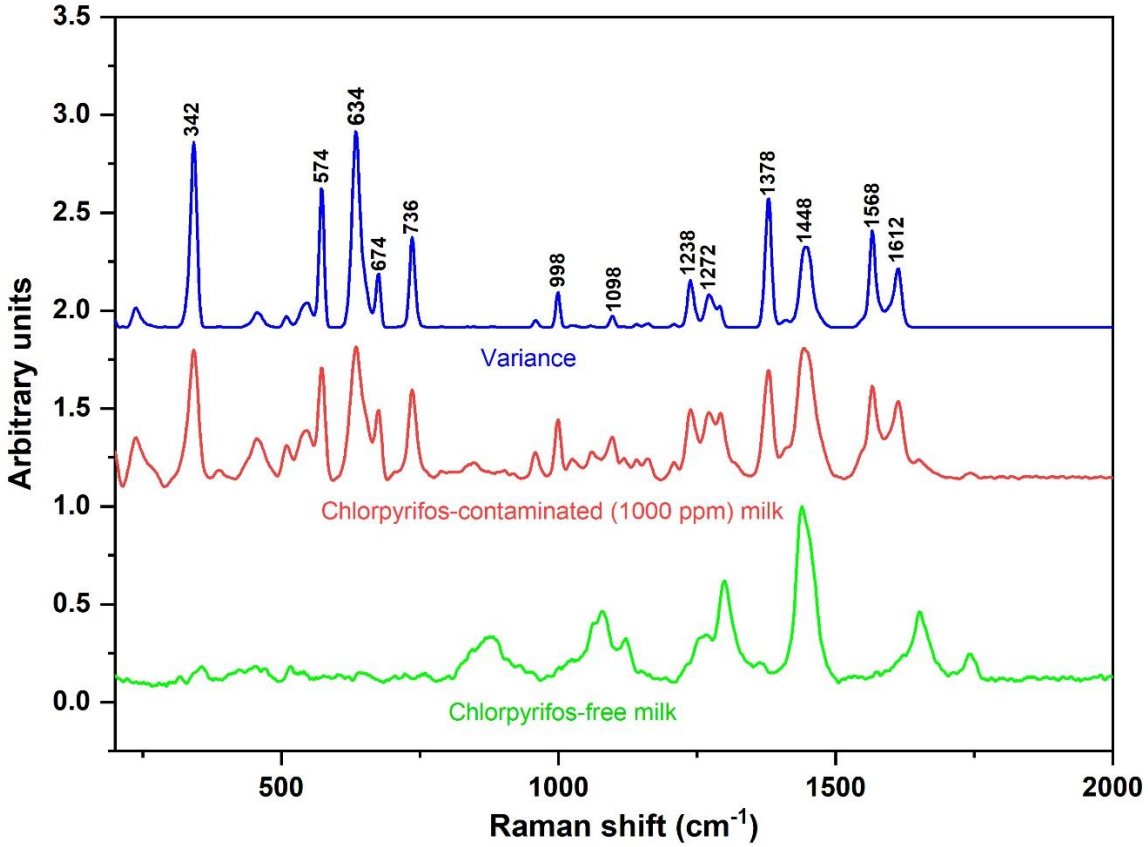


**Figure 4.5: Average Raman spectra of chlorpyrifos-free milk, average Raman spectra of chlorpyrifos-contaminated milk; ANOVA plot showing the variance of the two groups.**

The Raman bands exhibiting substantial variance in milk include those centered at 342, 574, 634, 674, 736, 998, 1240, 1272, 1378, 1448, 1568, and 1612 $cm^{-1}$. Notably, the 342 and 634 $cm^{-1}$ peaks are intense, indicating higher variance than others.

### 4.2.2 Statistically Significant Bands in Kale Samples

Similar results were obtained after plotting the mean spectra of control and treated kale samples, together with their variance. As Figure 4.6 shows, the bands exhibiting significant differences between the spectra of chlorpyrifos-free and chlorpyrifos-contaminated kale samples are those centered at 342, 574, 632, 672, 736, 1100, 1146, 1170, 1244, 1278, 1322, 1376, 1450, 1516, and 1568 $cm^{-1}$.

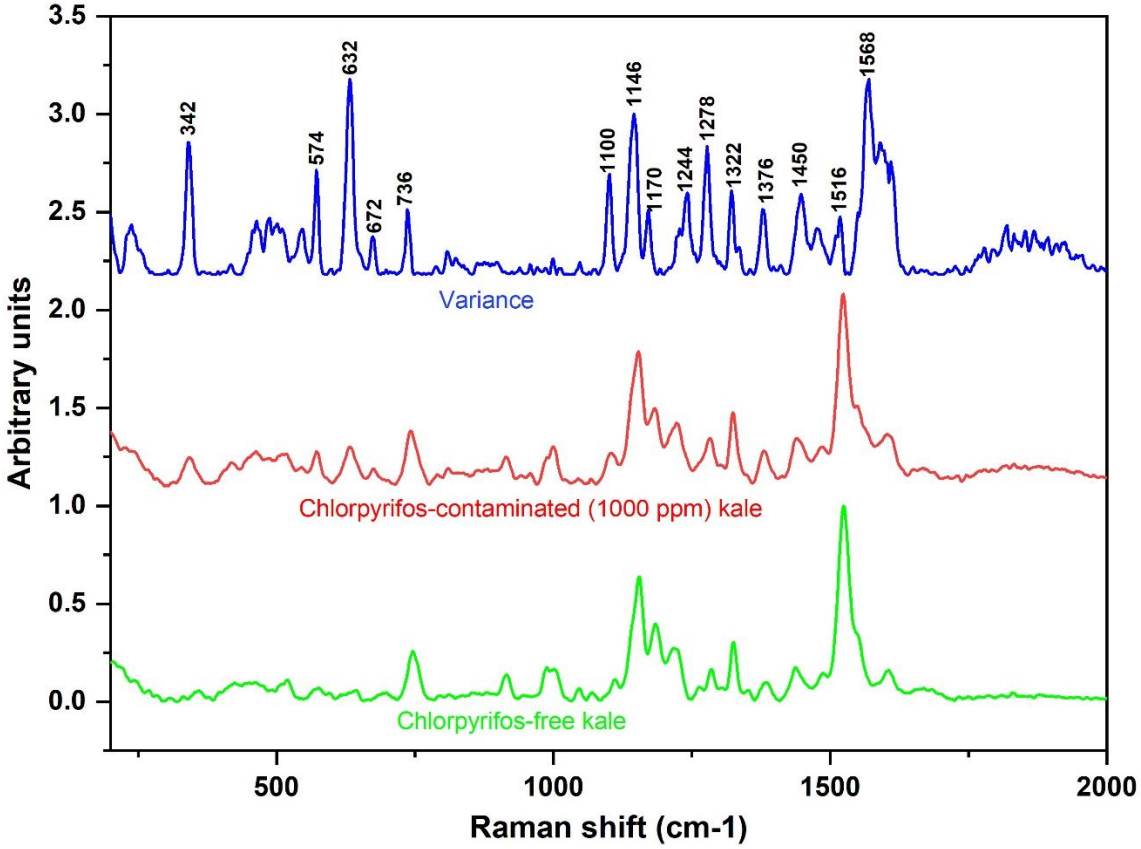


**Figure 4.6: Average Raman spectra of chlorpyrifos-free kale, average Raman spectra of chlorpyrifos-contaminated kale; ANOVA plot showing the variance of the two groups**

The plot of variance for kale closely matches that of milk but exhibits more regions of high variance. This suggests that the variance in both cases arises from the same compound, present in the treated sample but absent in the control groups. Consequently, the ANOVA-informed bands with significant variance can be linked to chlorpyrifos peaks, enabling the detection of residues in milk and kale leaves.

### 4.2.3 Correlation of Raman Bands with Chlorpyrifos Characteristic Peaks

As mentioned, the variant bands identified through ANOVA can be correlated with the respective chlorpyrifos peak responsible for the difference. Table 4.1 summarizes the

identified peak regions in milk and kale samples, as well as the associated chlorpyrifos Raman characteristic peak and attribution.

**Table 4.1: Correlation of peak regions with characteristic peaks of chlorpyrifos**

| Raman shift ($cm^{-1}$) | | Chlorpyrifos characteristic peak | Peak assignment | Reference |
|---|---|---|---|---|
| **Milk** | **Kale** | | | |
| 342 | 342 | 342 | C-Cl stretch | Zhang *et al*., 2011 |
| 574 | 574 | 568 | *v*(P=S) | Zhu *et al*., 2021 |
| 634 | 632 | 630 | P=S stretch | Du *et al*., 2020 |
| 674 | 672 | 676 | P=S stretch | Ma *et al*., 2020 |
| 736 | 736 | 736 | P-O-C stretch | Tao *et al*., 2022 |
| 1098 | 1100 | 1100 | P-O-C stretch | Chen *et al*., 2023 |
| - | 1146 | 1150 | w (C-H) | Zhu *et al*., 2018 |
| - | 1170 | 1168 | δ(C–H) | Chen *et al*., 2023 |
| 1240 | 1244 | 1236 | *δ*(C–H) | Ma *et al*., 2020 |
| 1272 | 1278 | 1276 | *δ*(C–H), $v_{as}$(C=C) | Ma *et al*., 2020 |
| - | 1322 | 1326 | *v* (C = C) | Zhu *et al*., 2018 |
| 1378 | 1376 | 1378 | $CH_3$ bending mode | Hongsibsong *et al*., 2020 |
| 1448 | 1450 | 1452 | C-H deformation | Xu *et al*., 2017 |
| 1568 | 1568 | 1568 | C-H wagging | Zhu *et al*., 2021 |

From Table 4.1, it is evident that some variance bands resulting from chlorpyrifos peak contribution (1150, 1168, and 1326 $cm^{-1}$) were present in kale leaves but absent in milk. This phenomenon can be attributed to peak overlap, which is common in Raman spectroscopy, especially when dealing with complex samples (Schulze *et al*., 2022). One possible reason for peak overlap is that the molecular vibrations of milk in these regions are significantly close to those of chlorpyrifos moieties. As a result, spectral overlap obscures individual peaks from chlorpyrifos and complicates the differentiation between control and spiked milk samples. This implies that the bands around these regions are

unreliable in detecting chlorpyrifos residues across diverse sample types. The broad peaks in the Raman spectrum of milk, centered around 1120 and 1300 $cm^{-1}$ (see Figure 4.1), can also explain the overlap of the chlorpyrifos peaks at 1150, 1168, and 1326 $cm^{-1}$. Further, the overlap can result from weak chlorpyrifos Raman peaks in these regions, diminishing their prominence.

A comparative analysis of the characteristic Raman peaks of chlorpyrifos and those obtained from ANOVA shows slight deviations in some cases. This can be ascribed to the interaction between the sample matrix and chlorpyrifos molecules, which causes the shifts. More specifically, the chlorpyrifos characteristic peaks reported in existing literature are obtained from the pure form of the pesticide (Ma *et al*., 2020; Tao *et al*., 2022; Xu *et al*., 2017; Zhang *et al*., 2011). Therefore, the peaks reported mirror the molecule's vibrational modes in the unperturbed state without matrix interactions. However, in the current study, chlorpyrifos Raman measurements were made in the presence of solvents and matrices, including water and milk, which altered the chemical environment of chlorpyrifos, causing the shift.

Despite the slight shift in ANOVA-identified bands, the regions with high variance for milk and kale samples were selected as potential bands for chlorpyrifos detection. These bands were further explored using PCA for chlorpyrifos fingerprint determination. Additionally, PCA helped overcome the challenge of overlapping peaks while revealing the relevant hidden information in the spectra data.

### 4.3 PCA of Raman Spectra using ANOVA-identified Bands for Fingerprint Identification

Principal component analysis was performed on each band to ascertain the effectiveness (the ability of a band to visually separate the sample groups in PCA space) of the bands identified through ANOVA as potential Raman fingerprints of chlorpyrifos. The bands investigated included 314-354, 558-588, 614-686, 720-750, 1086-1192, 1218-1348, 1366-1396, 1424-1530, and 1542-1630 $cm^{-1}$. Table 4.2 summarizes the PCA results of the bands.

**Table 4. 2: PCA results for the ANOVA-identified bands**

| | Milk | | | Kale | | |
|---|---|---|---|---|---|---|
| **Band ($cm^{-1}$)** | **Explained variance in PC1 & PC2 (%)** | **No. of misclassified spectra** | | **Explained variance in PC1 & PC2 (%)** | **No. of misclassified spectra** | |
| | | **Control** | **Treated** | | **Control** | **Treated** |
| 314-354 | 94 | 0 | 0 | 93.6 | 2 | 0 |
| 558-588 | 85 | 0 | 0 | 78 | 7 | 11 |
| 614-686 | 81.9 | 1 | 0 | 65 | 3 | 0 |
| 720-750 | 93.3 | 0 | 1 | 90 | 4 | 8 |
| 1086-1192 | 92 | 0 | 0 | 92.8 | 2 | 4 |
| 1218-1348 | 91.5 | 0 | 2 | 93.1 | 2 | 4 |
| 1366-1396 | 93 | 0 | 0 | 89 | 2 | 2 |
| 1424-1530 | 85 | 3 | 0 | 93 | 3 | 5 |
| 1542-1630 | 93 | 0 | 0 | 92.8 | 5 | 8 |

From these PCA results, the band at 314-354 cm-1, centered at 342 cm-1, was the most effective discriminator between chlorpyrifos-contaminated milk and kale samples and those without contamination. The band explained the highest cumulative variance for the first two PCs compared to other bands, as depicted in Table 4.2. The first two PCs were extracted from the PCA results for this band, explaining cumulative variances of 94% in

the Raman spectral data of milk and 93.6% in the spectra of kale leaves. This implies that the selected band contains spectral features that effectively distinguish the chlorpyrifos-free samples from those containing the pesticide. The band centered at 342 $cm^{-1}$ is associated with the vibrational modes of C-Cl bonds present in chlorpyrifos but naturally absent in milk and kale samples. Thus, the treated samples exhibited a peak in this region, resulting in higher variance in their spectra compared to the control group. This allowed for the definitive differentiation of chlorpyrifos-containing milk and kale leaves from chlorpyrifos-free samples. As discussed in section 2.1, the pyridine ring Cl is attached to in the chlorpyrifos molecule remains unchanged during degradation. This implies that the identified band can also detect the presence of TCP, the primary metabolite of chlorpyrifos (Thuku *et al*., 2025a).

While other bands were promising for milk, with most showing high explained variance in PC1 and PC2 and accurately distinguishing between control and treated samples, they performed poorly on kale. Most bands could not discriminate between treated and untreated kale samples, as evidenced by their high misclassification rates. In practical terms, the higher number of misclassified spectra for kale indicates that, across many bands, the projected scores of control and treated kale overlap substantially in PCA space. This means that those bands do not carry sufficiently distinct information to separate the two groups in this matrix. By contrast, the near-zero misclassifications for milk across most bands suggest that the same spectral regions provide much clearer separation between contaminated and uncontaminated milk samples. This phenomenon can be explained by attributing these bands to groups such as phosphorus, oxygen, carbon, and sulfur, which are present in both

chlorpyrifos and kale leaves. Because structurally similar groups are present in both control and treated samples, their Raman contributions overlap in these spectral regions, reducing contrast between the two classes and limiting separation in PCA space, particularly in the kale matrix.

In contrast, using the first two PCs, the band 314-354 $cm^{-1}$ captures high variability in the milk and kale Raman spectra, which is necessary to effectively discriminate sample groups based on chlorpyrifos residues. Additionally, this band exhibited a low rate of spectral misclassification, with only two spectra from untreated kale leaves misclassified. Based on these features, the band 314-354 $cm^{-1}$ stands out among the ANOVA-selected bands, qualifying it as the chlorpyrifos fingerprint. Figure 4.7 (a) shows the PCA results for the control and highly concentrated (1000 ppm chlorpyrifos) milk samples, and (b) the kale samples using the identified fingerprint.

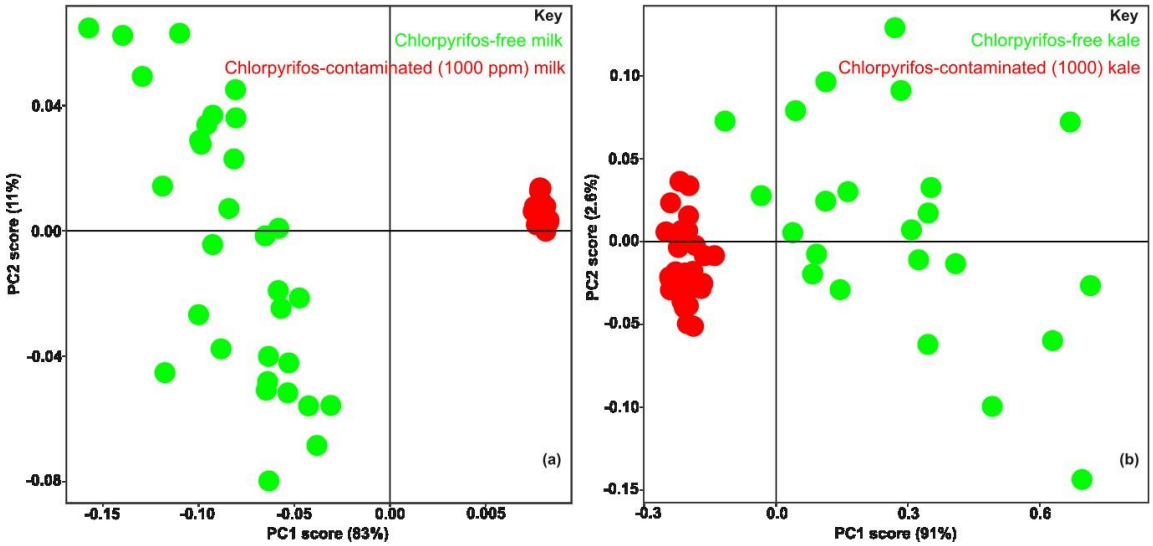


**Figure 4.7: PCA scores plots showing the fingerprint's ability to distinguish chlorpyrifos-contaminated (1000 ppm) from chlorpyrifos-free (a) milk and (b) kale**

In the case of milk, Figure. 4.7 (a) shows that PC1 separates the spectra of milk samples into two clusters: the control/chlorpyrifos-free group and the treated/chlorpyrifos-contaminated group. The control milk samples group along the negative PC1, while those containing chlorpyrifos cluster on the positive PC1. On the other hand, PC1 also groups the Raman spectra from kale leaves into control and treated samples, with the treated kale samples clustering on the negative PC1 and the control samples clustering on the positive PC1. Notably, unlike the milk samples, which were all correctly classified, two kale samples were misgrouped by PCA. The PCA results also show that the Raman fingerprint effectively discriminated milk and kale samples with high chlorpyrifos levels (1000 ppm) from those without. However, it is imperative to appreciate that the food safety problem of pesticide contamination is interested in low residue levels around the MRL. Additionally, the fingerprint band at 342 $cm^{-1}$ is not always prominent in individual raw spectra, and the differences between control and contaminated samples are subtle when visually inspected. However, PCA applied to this restricted spectral region can still exploit minor but systematic intensity variations across spectra, thereby allowing the fingerprint region to contribute to sample discrimination. As such, the fingerprint was tested at low chlorpyrifos concentrations above and below the 0.01 ppm MRL. For illustration purposes, 0.003 ppm was chosen to represent the below MRL group, while 0.3 ppm represented samples with concentrations above the MRL. The PCA results obtained when focusing on the fingerprint were compared with those of the entire spectrum and depicted in Figure 4.8.

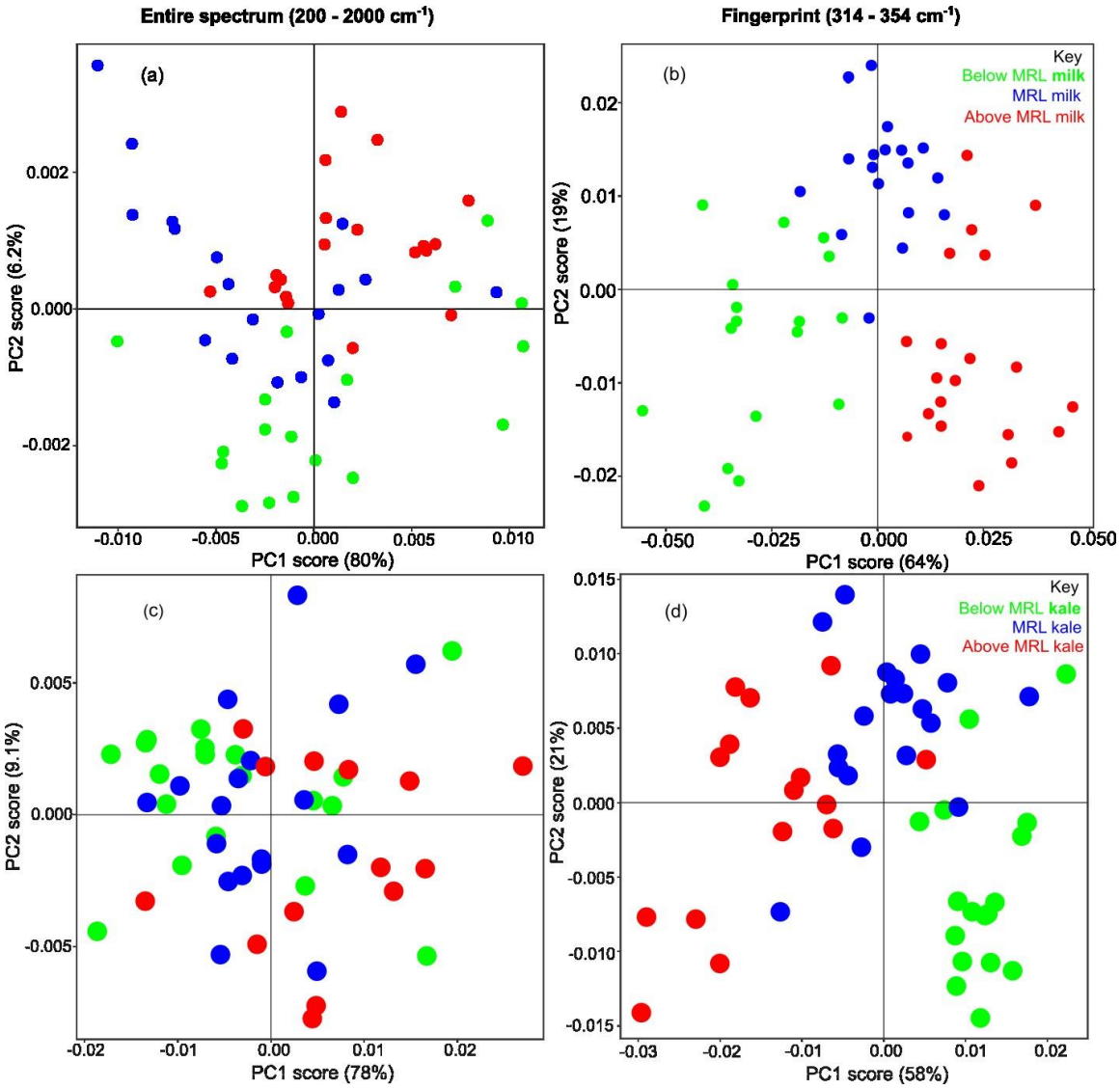


**Figure 4.8: PCA scores plots showing (a, c) significant overlap among Below MRL, MRL, and above MRL milk and kale samples using the entire spectrum (200-2000 $cm^{-1}$) versus (b, d) distinct separation of the three groups using the fingerprint (314-354 $cm^{-1}$).**

The PCA results presented in Figures 4.8 (a) and (c) reveal the challenge of using the entire spectrum for analysis, as it results in significant overlap between the low-concentration groups. Thus, PCA based on the full spectral range cannot clearly distinguish the three sample groups: below MRL, MRL, and above MRL. On the other hand, Figures 4.8(b) and (d) demonstrate the effectiveness of using the identified Raman fingerprint to discriminate

milk and kale samples based on their chlorpyrifos concentrations relative to the MRL. A clearer separation is observed using the fingerprint, particularly for the milk samples. PCA performed using the fingerprint groups the spectra of the samples according to chlorpyrifos concentration along PC1. For milk, Figure 4.8 (b) shows the samples containing chlorpyrifos levels below the MRL cluster on the negative PC1, while those whose concentration exceeds the MRL cluster on the positive PC1. Milk samples with chlorpyrifos levels at MRL cluster between the below MRL and the above MRL.

Similarly, Figure 4.8 (d) shows kale samples separated according to chlorpyrifos concentration along PC1, with samples containing levels above the MRL group on the negative PC1, while those below the MRL cluster on the positive PC1. Just as in milk samples, the groups whose chlorpyrifos concentration is at the MRL cluster between the below-MRL and above-MRL groups. In both instances, a positive trend is evident, with chlorpyrifos concentration increasing from left to right and from right to left for milk and kale samples, respectively. While using the fingerprint improves the classification of kale samples compared to the entire spectral range, the separation is characterized by several misclassifications. These PCA limitations necessitated the adoption of ML models that significantly enhanced our method's classification capacity. Such sophisticated methods could better handle complex patterns and non-linear relationships in the spectral data of milk and kale samples, thereby improving classification performance.

**4.4 Evaluating Models for Sample Classification and Chlorpyrifos Quantification**

One of the current work’s objectives was to assess the performance of ML algorithms in classifying samples based on their chlorpyrifos concentration (relative to the established

MRL) and quantifying the levels using the identified fingerprint. The models were first developed using preprocessed data across the entire spectral range, resulting in poor predictive performance. This was likely because using the full spectrum could have led to overfitting, in which the models learned both important and irrelevant patterns/noise, thus performing poorly on the test set (unseen data). Notably, the Raman spectral data obtained in the current study contained a wide range of wavelengths. Thus, training a model using the entire set of features limited its ability to focus and learn the most informative aspects of the data. Consequently, modeling was done using the variables with the essential information. This was realized using PCA, which transformed the original features in the Raman datasets into PCs, reducing dimensionality while removing noise and redundant information.

The PCA results from the Raman fingerprint (314-354 $cm^{-1}$) were used as inputs for the ML models. Scree plots were used to determine the appropriate number of PCs for modeling. To demonstrate the utility of ML models in resolving the misclassifications observed in the PCA section, RF and SVM models were built using data from section 4.3 to classify samples into treated and control groups, representing chlorpyrifos-contaminated and chlorpyrifos-free samples, respectively.

Further, RF, SVM, and SVR models were used to quantify chlorpyrifos levels and classify the preprocessed Raman spectra from the 21 concentrations into three groups: below MRL, MRL, and above MRL. PCA of the 21 concentrations, along with the control samples, was performed using the identified Raman fingerprint of chlorpyrifos. Based on the scree plot

results (Figure 4.9), the first six and ten PCs were extracted for milk and kale samples, respectively. These PCs explained 98% of the cumulative variance in the samples' spectral data. The PCs were used as inputs in developing the classification and regression models.

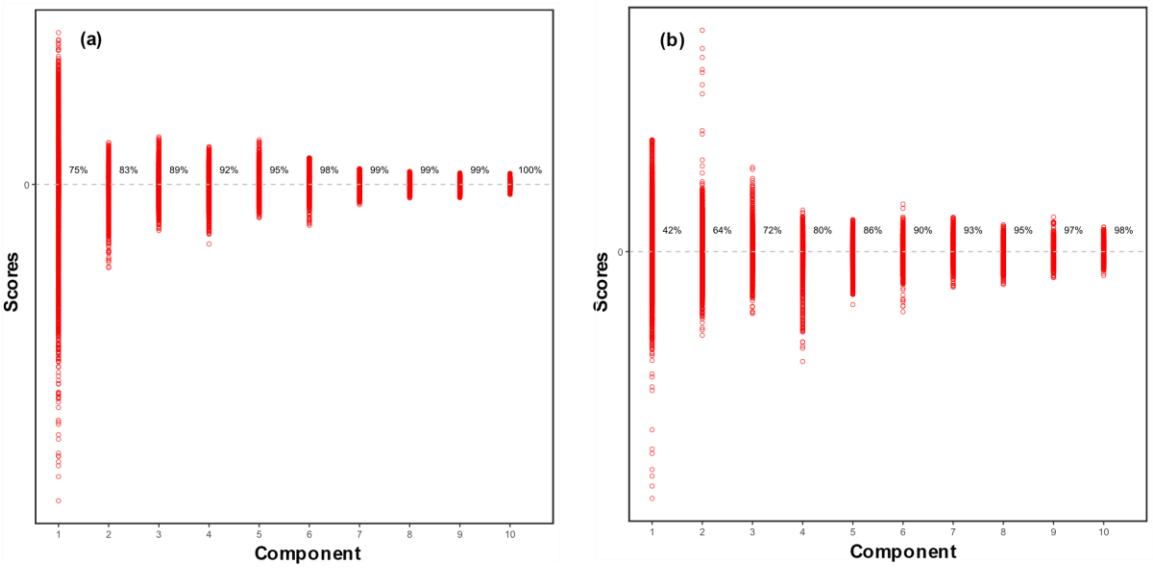


**Figure 4.9: Scree Plots for Principal Component Selection in (a) Milk and (b) Kale Samples**

### 4.4.1 Evaluation of RF in Classifying Milk and Kale Samples

Using the data from section 4.3 and the first two PCs as inputs to the RF model, all milk and kale samples were correctly classified into their respective classes, including the kale samples grouped wrongly by PCA. No misclassifications occurred, implying that the model had an accuracy of 100% on the test set.

An RF model was then developed to predict the classes to which test samples, which the model had not yet seen, belong. As described in section 3.7.1, the two crucial parameters that were varied during the training process included *mtry* values ranging from 1 to 6 for milk and 1 to 10 for kale samples as well as *ntrees.* The best models had 500 trees and an *mtry* value of 1. It was noted that choosing a low *mtry* value reduced overfitting and

increased the model's generalization. The optimal model was adopted to predict the three classes of kale and milk samples. Table 4.3 presents the confusion matrices for the RF model on the test data from milk and kale. An overall classification accuracy of 95.23% and 90.15% was achieved for milk and kale samples, respectively.

**Table 4. 3: Confusion matrix of milk and kale test data for RF**

| **RF performance** | | **Milk** | | | **Kale** | | |
|---|---|---|---|---|---|---|---|
| Accuracy (%) | | 95.23 | | | 90.15 | | |
| Kappa | | 0.93 | | | 0.85 | | |
| | | **Actual Group** | | | **Actual Group** | | |
| | | Below | MRL | Above | Below | MRL | Above |
| **Predicted Group** | Below | 397 | 0 | 28 | 347 | 0 | 43 |
| | MRL | 3 | 415 | 5 | 0 | 415 | 4 |
| | Above | 19 | 5 | 387 | 72 | 2 | 373 |
| Precision (%) | | 93.41 | 98.11 | 94.16 | 88.97 | 99.05 | 82.89 |
| Recall (%) | | 94.75 | 98.81 | 92.14 | 82.82 | 98.81 | 88.81 |

Table 4.3 shows that RF performed better on milk, with an overall accuracy of 95.23% compared to 90.15% on kale samples. The higher Kappa of 0.93 associated with RF's performance on milk versus a Kappa of 0.85 for kale samples indicates better classification agreement between milk's actual group and the predicted group. RF's precision and recall were generally higher for milk samples, particularly the "Below" class. In contrast, the model did not perform well at classifying kale samples, as reflected in lower precision and recall of 82.82% and 88.97%, respectively.

However, despite the model's performance varying between the two sample types, RF demonstrated a high ability to classify milk and kale samples according to their chlorpyrifos levels: below MRL, at MRL, and above MRL. It achieved a high accuracy above 90% in both instances, indicating that the model performed reasonably well (Chu *et al*., 2012; Foody, 2009; Olusanya *et al*., 2022). Furthermore, the Kappa values of RF were

above 0.80, which indicates a strong model performance (Hallgren, 2012; Marchevsky *et al*., 2020). Thus, there was generally significant agreement between the actual and the RF-predicted classes. Despite performing better in classifying milk, RF still classified kale samples relatively well, especially in the "MRL" and "Above" groups, with high precision and recall values above 80% in most cases. RF's structure explains these results. The model combines multiple decision trees, trains them independently, and aggregates their results, reducing overfitting. Also, the model used 500 trees, and splitting was based on a single feature, as indicated by the *ntree* and *mtry* values of 500 and 1, respectively. This parameter combination enabled the RF model to reduce variance, thereby enhancing its overall robustness. Therefore, the results of this study demonstrate that RF is a supervised ML model that remains stable even in scenarios involving complex classifications.

### 4.4.2 Quantitative Analysis of Chlorpyrifos in Samples Using RF

As in the classification tasks, using the identified fingerprint, RF achieved remarkable performance in predicting chlorpyrifos levels in test kale and milk samples. Figure 4.10 illustrates the calibration performance of the RF models in predicting chlorpyrifos concentrations in milk and kale samples. The RF calibration plots show strong alignment between predicted and actual concentrations, indicating high model reliability. The error bars, representing standard deviations, highlight the variability of actual chlorpyrifos concentrations around the regression line, further emphasizing the model's precision. The models exhibited exceptional predictive accuracy, reflected in $R^2$ values of 0.9997 for milk and 0.9998 for kale samples. These values, close to unity, indicate that the RF models effectively captured the relationship between the input data and chlorpyrifos concentrations. Additionally, the Root Mean Square Error of Prediction (RMSEP) values

were remarkably low at 0.0231 ppm for milk and 0.0182 ppm for kale, indicating minimal average prediction errors.

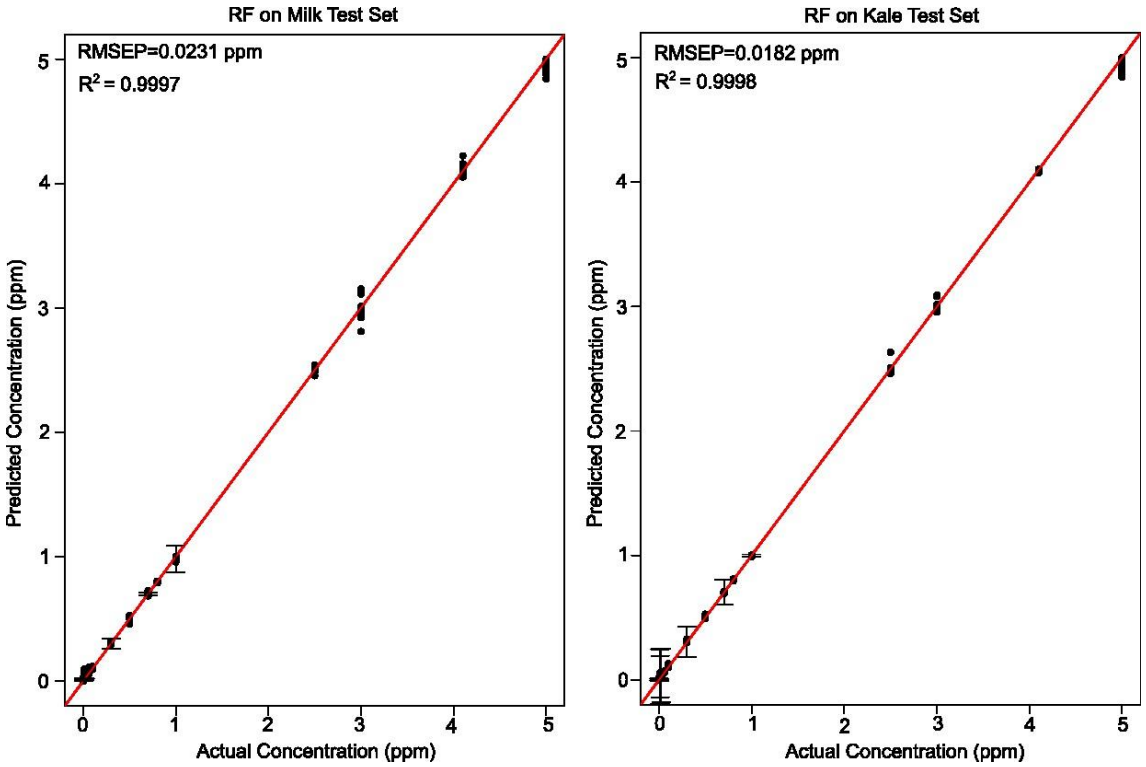


**Figure 4.10: RF calibration plots for test data sets. The error bars use standard deviation to show the variability in the actual chlorpyrifos concentration around the regression line.**

In addition to calibration, the RF models' sensitivity was assessed by determining the Limits of Detection (LOD) and Limits of Quantification (LOQ) using the pseudo-univariate approach discussed in Section 3.7.4. A linear fit function was applied to derive the standard deviation and slope, which were then used to calculate LOD (3σ) and LOQ (10σ). The results, summarized in Table 4.4, demonstrate significantly low LOD and LOQ values for both milk and kale samples, underscoring the models' capability to detect and quantify trace levels of chlorpyrifos.

**Table 4.4: RF models performance based on LOD and LOQ**

| RF | LOD (ppm) | LOQ (ppm) |
|---|---|---|
| Milk | 0.00290 | 0.00966 |
| Kale | 0.00212 | 0.00708 |

The predictive performance of the RF models was further evaluated by comparing predicted chlorpyrifos concentrations with actual values in the test samples. Table 4.5 presents these predictions for both milk and kale samples. The predicted values closely match the actual concentrations, demonstrating the models' high accuracy. For example, the lowest measured concentration of 0.003 ppm was predicted as 0.003005 ppm in milk and 0.003011 ppm in kale, with minimal errors of 0.000719 ppm and 0.000522 ppm, respectively.

**Table 4.5: Chlorpyrifos concentration predictions in milk and kale using RF**

| RF-predicted chlorpyrifos concentrations (ppm) | | |
|---|---|---|
| **Actual** | **Milk** (Predicted ± SD) | **Kale** (Predicted ± SD) |
| 0.003 | 0.003005 ± 0.000719 | 0.003011 ± 0.000522 |
| 0.01 | 0.010001 ± 0.000242 | 0.010009 ± 0.000225 |
| 0.03 | 0.030015 ± 0.005550 | 0.030028 ± 0.003400 |
| 0.06 | 0.05997 ± 0.002010 | 0.060080 ± 0.003100 |

| | | |
|---|---|---|
| 0.1 | 0.099888 ± 0.001930 | 0.100399 ± 0.045100 |
| 0.3 | 0.3006 ± 0.0024100 | 0.301206 ± 0.031600 |
| 0.5 | 0.500027 ± 0.008030 | 0.5006 ± 0.013400 |
| 0.8 | 0.7998 ± 0.004470 | 0.80016 ± 0.037600 |
| 1.0 | 0.996583 ± 0.025400 | 1.00186 ± 0.147000 |
| 2.5 | 2.483576 ± 0.451000 | 2.497843 ± 0.221000 |

To statistically validate the agreement between the predicted and actual concentrations, paired t-tests were conducted. For milk samples, the t-test indicated no significant difference ($t(9) = 0.064$, $p = 0.950$), confirming that the RF model predictions closely matched the actual values. Similarly, for kale samples, the t-test also showed no significant difference ($t(9) = 0.062$, $p = 0.952$), further supporting the high predictive accuracy of the RF models. These results reinforce the conclusion that the RF models provide reliable predictions across both sample types. The RF models' high predictive accuracy and sensitivity underscore their strong generalizability and reliability in quantifying chlorpyrifos in complex matrices such as milk and kale. The high number of decision trees utilized in the RF models reduced overfitting, thereby enhancing model stability and performance on previously unseen Raman spectroscopy data.

### 4.4.3 Evaluation of SVM in Classifying Milk and Kale Samples

The first two PCs extracted in section 4.3 were also used to build an SVM model, which gave an accuracy of 100%. The model correctly classified all the sample spectra, including the kale samples that were incorrectly clustered by PCA.

The ability of SVM to distinguish milk and kale samples across the three groups (Below MRL, MRL, and above MRL) was also assessed. The critical parameters considered in

training SVM classification models included kernel type, gamma, and cost. For each sample (milk and kale), optimization involved a grid search using predefined values between $10^{-2}$ and $10^{-3}$ for the cost, while gamma values were chosen from the set (0.001, 0.01, 0.1, 1, 5, 10). The radial basis kernel provided the best model due to its high hyperplane flexibility. The optimal model had a cost of 10 and a gamma of 0.1 for milk, while a cost of 100 and a gamma of 1 provided the best performance for kale spectra. The confusion matrices for the SVM models on the test data from milk and kale are shown in Table 4.6. An overall classification accuracy of 95.79% was achieved for milk samples, while the optimal SVM model achieved 92.61% accuracy for kale samples.

**Table 4. 6: Confusion matrix of milk and kale test data for SVM**

| **SVM performance** | | **Milk** | | | **Kale** | | |
|---|---|---|---|---|---|---|---|
| Accuracy (%) | | 95.79 | | | 92.61 | | |
| Kappa | | 0.94 | | | 0.89 | | |
| | | **Actual Group** | | | **Actual Group** | | |
| | | Below | MRL | Above | Below | MRL | Above |
| **Predicted Group** | Below | 402 | 0 | 21 | 354 | 1 | 22 |
| | MRL | 3 | 417 | 12 | 0 | 415 | 1 |
| | Above | 14 | 3 | 387 | 65 | 4 | 397 |
| Precision (%) | | 95.01 | 96.53 | 95.79 | 93.90 | 99.76 | 85.19 |
| Recall (%) | | 95.94 | 99.29 | 92.14 | 84.49 | 98.81 | 94.52 |

Like RF, SVM also performed better on milk samples than on kale samples, with accuracies of 95.79% and 92.61%, respectively. This performance was also supported by the association of milk classification, with a Kappa of 0.94, compared with 0.89 for kale samples. Generally, SVM achieved high precision and recall values across milk and kale samples, outperforming RF in most cases.

### 4.4.4 Quantitative Analysis of Chlorpyrifos in Samples Using SVR

Figure 4.11 illustrates the calibration performance of the Support Vector Regression (SVR) models in predicting chlorpyrifos concentrations in milk and kale samples. The SVR

calibration plots exhibit a strong correlation between predicted and actual chlorpyrifos concentrations, demonstrating the model's high reliability. The error bars, which represent standard deviations, illustrate the dispersion of actual concentration values around the regression line, providing further evidence of the model's precision.

Like the RF regression models, SVR exhibited high $R^2$ values (0.9961 and 0.9981) and low RMSEP values (0.0897 and 0.0658), implying that a high degree of chlorpyrifos quantification accuracy was ultimately achieved on both sample types. The high $R^2$ and low RMSEP values can be attributed to the models' architecture, specifically the radial kernel, which enhances their generalization capacity and regression performance across different samples.

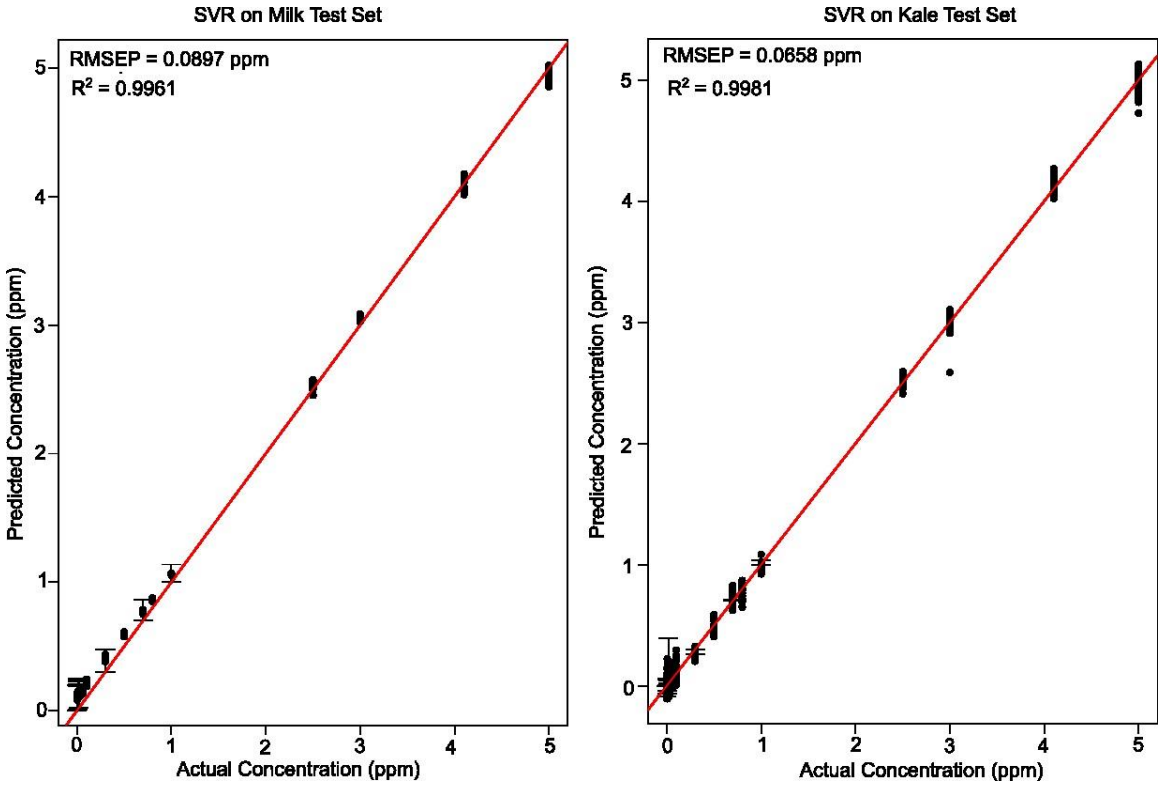

**Figure 4.11: SVR calibration plots for test data sets. The error bars use standard deviation to show the variability of the actual concentration of chlorpyrifos around the regression line.**

Furthermore, pseudo-univariate graphs were plotted, and the (3σ) and (10σ) approach was used to determine the Limits of Detection (LOD) and Limits of Quantification (LOQ) using the standard deviation of the y-intercept and the calibration curve's slope. The LOD and LOQ calculated for the SVR models are summarized in Table 4.7.

**Table 4.7: SVR models performance based on LOD and LOQ**

| SVR | LOD (ppm) | LOQ (ppm) |
|---|---|---|
| Milk | 0.00307 | 0.01530 |
| Kale | 0.00847 | 0.02820 |

The low LOD and LOQ values provide further insight into the SVR models' sensitivity when detecting chlorpyrifos levels across diverse samples. The slightly higher LOQ of the SVR models explains the lower prediction accuracy at lower concentrations (see Table 4.8), which improves at higher chlorpyrifos concentration levels.

The predictive performance of the SVR models was further evaluated by comparing predicted chlorpyrifos concentrations with actual values in the test samples. Table 4.8 presents these predictions for both milk and kale samples. The predicted values closely match the actual concentrations, although slight deviations were observed at lower concentrations. For example, the lowest measured concentration of 0.003 ppm was predicted as 0.005943 ppm in milk and 0.003061 ppm in kale, with errors of 0.022500 ppm and 0.021700 ppm, respectively.

**Table 4.8: Chlorpyrifos concentration predictions in milk and kale using SVR**

| SVR-predicted chlorpyrifos concentrations (ppm) | | |
|---|---|---|
| **Actual** | **Milk** (Predicted ± SD) | **Kale** (Predicted ± SD) |
| 0.003 | 0.005943 ± 0.022500 | 0.003061 ± 0.021700 |
| 0.01 | 0.026408 ± 0.024800 | 0.021153 ± 0.022900 |
| 0.03 | 0.04829 ± 0.028400 | 0.034332 ± 0.043300 |
| 0.06 | 0.053565 ± 0.020700 | 0.065359 ± 0.017300 |
| 0.1 | 0.191264 ± 0.015000 | 0.106248 ± 0.029000 |
| 0.3 | 0.38359 ± 0.012900 | 0.296675 ± 0.023400 |
| 0.5 | 0.577849 ± 0.011500 | 0.490417 ± 0.019800 |
| 0.8 | 0.846956 ± 0.012400 | 0.803687 ± 0.020200 |
| 1.0 | 1.03078 ± 0.014400 | 0.988029 ± 0.036100 |
| 2.5 | 2.503069 ± 0.008220 | 2.525368 ± 0.024800 |

Notably, SVR-predicted chlorpyrifos levels at lower concentrations showed slight deviations from the measured concentrations. For example, the lowest measured concentration, 0.003 ppm, was predicted as 0.005943 ppm in milk. However, the models' prediction accuracy improved at higher concentrations, with 2.5 ppm of the pesticide predicted as 2.503069 ppm and an error of 0.008220 ppm. The lower prediction accuracy of SVR compared to RF gave rise to the slightly lower $R^2$ values for milk and kale samples.

To further assess the agreement between the predicted and actual chlorpyrifos concentrations, paired t-tests were performed. For milk samples, the results showed no statistically significant difference between actual and predicted values ($t(9) = 1.389$, $p = 0.198$), indicating good predictive accuracy. Similarly, for kale samples, no significant difference was observed ($t(9) = 0.998$, $p = 0.344$). Although the mean predictions are closely aligned with the actual concentrations, the considerably larger standard deviations, particularly at lower concentration levels, reveal reduced precision of the SVR model

compared to the RF model. This increased variability explains the wider spread observed in the SVR predictions despite comparable mean accuracy.

### 4.4.5 Comparison of Model Performance in Classifying Samples

Comparing the performance of RF and SVM in classifying milk and kale samples based on their chlorpyrifos concentration using the identified spectral fingerprint was essential to understanding the models' strengths and limitations. Consequently, a comparative analysis of the models helped identify the model better suited for classifying milk and kale samples based on chlorpyrifos residues. Starting with the accuracy and Kappa achieved on milk, SVM outperformed RF, as indicated by accuracy and Kappa of 95.79% and 0.94 against RF's 95.23% and 0.93. Generally, SVM had higher precision and recall than RF, suggesting that SVM performed slightly better at classifying new Raman data from milk and kale samples. SVM also outperformed RF in categorizing kale samples according to the three groups: Below MRL, MRL, and above MRL. In this case, SVM achieved an accuracy of 92.61%, while RF achieved a lower accuracy of 90.15% on the kale test data. The Kappa associated with SVM (0.89) was higher than RF's 0.85. SVM outperformed RF in terms of precision, particularly for the "Below" (93.90%) and "Above" (85.19%) groups compared to RF's 88.97% and 82.89%. Additionally, the recall was significantly higher with SVM, especially for the "Above" group at 94.52%, compared with RF's 88.81%. Thus, these results indicate that SVM outperformed RF in all aspects.

Several features distinguish the two classification models, resulting in SVM's slightly higher performance. Starting with the model structure, SVM maximizes the separation of the three classes (below MRL, MRL, and above MRL) by finding a hyperplane. After

model optimization, this study selected a radial basis function kernel, which is known for its excellent performance in pattern classification of non-linear data (Aftab *et al*., 2014; Ding *et al*., 2021), like the one used in the current study. The RBF kernel enabled the SVM model to construct complex, non-linear decision boundaries, thereby enhancing its effectiveness in classifying non-linearly separable Raman datasets. More precisely, SVM mapped the input data to a higher-dimensional space, enabling better separation. On the other hand, RF relied on aggregating predictions from multiple decision trees, using a single feature at each split. This could have limited RF's ability to effectively capture the complex non-linear relationship among the three sample groups, slightly lowering its classification ability relative to SVM.

Furthermore, SVM's superior performance originates from its tuning parameters, such as cost and gamma. As mentioned, the best SVM model for classifying milk had a cost of 10 and a gamma of 0.1, whereas the values for kale samples were 100 and 1, respectively. The higher cost associated with the kale class prediction allowed for stronger regularization, as misclassifications were penalized more heavily, prompting the model to create a more complex boundary that could capture more complex patterns in the spectral data. For this reason, SVM demonstrated a better classification ability than RF on both milk and kale samples. Therefore, using the RBF kernel and a suitable combination of cost and gamma enabled SVM to capture more complex patterns than RF's tree-based splitting, illustrating the superiority of the SVM approach.

The results of this study further show that the accuracy of the two models was generally lower for kale samples than for milk samples. For instance, RF's recall for the "Below"

group (94.75%) in milk dropped to 82.82% in kale. Similarly, using SVM, recall was lower for the "Below" group (84.49%) in kale than for milk (95.94%). A possible explanation for these results is the complex patterns and a higher degree of overlap among the kale sample groups, as illustrated by the PCA results in section 4.3. The complexity and class overlap among the three classification groups could have made it difficult to identify class boundaries in SVM and to make accurate splits in RF. Moreover, using aluminium as a substrate for Raman measurements of milk resulted in signal amplification, thereby enhancing classification accuracy.

Using different performance metrics such as accuracy, Kappa, precision, and recall, this study shows that RF and SVM can effectively classify milk and kale samples into groups of below MRL, MRL, and above MRL using the Raman fingerprint of chlorpyrifos (314-354 $cm^{-1}$). Both models achieved high classification accuracy across different sample types, which is essential for the rapid, on-site detection of chlorpyrifos in vegetables and milk. However, SVM consistently outperformed RF across both spectral data sets, indicating that it is a better classifier in this context.

### 4.4.6 Comparison of Model Performance in Quantifying Chlorpyrifos Residues

In the case of quantitative analysis of chlorpyrifos in milk and kale samples, comparing the RF and SVR models offers insights into their quantitative capabilities. For both samples, RF consistently exhibits lower RMSEP values of 0.0231 and 0.0182 ppm compared to SVR's 0.0897 and 0.0658 ppm. Although the two ML models achieve high $R^2$ values, RF slightly outperforms the SVR in both milk and kale samples, with the values approaching unity. The RF's lower error values on unseen data, as well as higher $R^2$ values, suggest that

the model has higher predictive accuracy in relation to individual chlorpyrifos concentrations.

Similar observations were made when considering the two models' LOD and LOQ. For milk samples, RF achieved lower LOD and LOQ values, 0.00290 ppm and 0.00966 ppm, respectively. In contrast, the minimum chlorpyrifos concentrations reliably detected and quantified using SVR in milk were 0.00307 and 0.0153 ppm. Akin to these were the values achieved on the kale samples. While RF had LOD and LOQ of 0.00212 and 0.00708 ppm, respectively, SVR had LOD and LOQ of 0.00847 and 0.0282 ppm, respectively. These results indicate that RF performed better than SVR at detecting lower chlorpyrifos concentrations. Thus, while SVM proved superior to RF in classification tasks, RF is particularly more effective in regression problems where high sensitivity in detecting substantially low chlorpyrifos levels is necessary.

Furthermore, comparing the LOD and LOQ for the two models provides insight into the methods' applicability. RF achieved values slightly lower than the chlorpyrifos MRL (0.01 ppm), suggesting it can be used to detect and quantify this pesticide in samples such as milk and kale leaves. On the other hand, while SVR's LOD values are below the MRL for both samples, the models' LOQ values are slightly above the MRL. Therefore, SVR can detect chlorpyrifos levels below the MRL but might not reliably quantify the pesticide residues.

The classification and quantification performance obtained in this study is broadly comparable to that reported for chromatographic methods such as GC-MS when used for pesticide residue analysis. The RF and SVM classifiers achieved high overall accuracies of 95% for milk and 90% for kale, with Kappa values above 0.85, indicating strong agreement between predicted and actual contamination classes. In addition, the RF and SVR regression models yielded significantly low limits of detection and quantification for chlorpyrifos: LODs of approximately 0.002-0.003 ppm for milk and 0.002-0.008 ppm for kale, with corresponding LOQs of about 0.03 ppm. These sensitivity levels are comparable to those reported in some GC-MS-based pesticide residue studies. For example, a Bangladesh vegetable study reported GC-MS LOD and LOQ values of 0.011 and 0.034 mg/kg, respectively, for chlorpyrifos in cauliflower, cabbage, and eggplant (Momtaz and Khan, 2024). In milk-powder analysis by GC-MS, chlorpyrifos was quantified with an LOD of approximately 0.284 µg/kg and an LOQ of 0.862 µg/kg (Ali and Hassan, 2023). Furthermore, multi-residue GC-MS surveys across various crops reported LOQ values around 0.01 mg/kg for some pesticides (Yun *et al*., 2024). While GC-MS remains the reference technique, the Raman-ML approach developed in the present study delivers comparable detection capability and classification performance with the additional benefits of being non-destructive, solvent-free, and more suitable for rapid screening.

The current study envisioned the development of a rapid pesticide analysis method. Consequently, the method's rapidness was evaluated to determine the time required to assess the presence and levels of chlorpyrifos in milk and kale samples. Table 4.9 summarizes the turnaround time for the developed method.

**Table 4.9: Turnaround time for the machine learning-aided Raman spectroscopy method**

| Phase | Activity | Time (minutes) |
|---|---|---|
| Sample preparation | Filtration (milk), Slicing (kale) | < 5 |
| Data acquisition | Raman measurement collection (Milk: 1 minute 43 seconds, Kale: 30 seconds) | < 2 |
| Data analysis | Preprocessing, classification, and quantification | <8 |
| Overall time per sample | | <15 |

In contrast to traditional laboratory methods, such as LC-MS and GC-MS, which require time-consuming sample pretreatment and a long analysis time of more than 60 minutes (Amirav *et al*., 2014), the developed method allows samples to be processed and results generated within 15 minutes. This feature validates the method's suitability for rapid analysis of chlorpyrifos in food samples.

# CHAPTER FIVE

# CONCLUSIONS AND RECOMMENDATIONS

## 5.1 Summary and Conclusions

This study aimed to evaluate the feasibility of using Raman spectroscopy in combination with machine learning for the detection and quantification of chlorpyrifos residues in kale and milk. By first examining the spectra of control and spiked samples using one-way ANOVA, it was possible to identify Raman bands that showed statistically significant differences between uncontaminated and contaminated samples. These bands represent spectral regions that are particularly sensitive to chlorpyrifos-induced changes in the matrices and serve as a basis for identifying a meaningful Raman fingerprint of the pesticide in kale and milk.

Using the ANOVA-informed regions, PCA revealed clear separation between contaminated and control samples within the 314-354 $cm^{-1}$ band centered at 342 $cm^{-1}$ for both matrices. This demonstrated that the selected Raman fingerprint captures sufficient information to qualitatively distinguish contamination classes, including samples at different levels relative to the maximum residue limit. The PCA results, therefore, confirmed that combining ANOVA-based band selection with PCA provides an effective way to summarize the complex Raman spectra of kale and milk into a lower-dimensional representation that preserves the contamination-related structure in the data.

Building on this fingerprint, supervised machine learning models were developed to perform both classification and quantification. In both kale and milk, classification models

based on RF and SVM achieved high overall accuracy in discriminating between control and chlorpyrifos-contaminated samples across the concentration levels investigated. Regression models constructed using RF and SVR provided reasonably accurate predictions of chlorpyrifos concentration within the studied ranges, with $R^2$ and prediction errors indicating that the models are suitable for quantitative estimation. The corresponding LOD and LOQ fell within regulatory-acceptable ranges for chlorpyrifos residues and were broadly comparable to values reported for chromatographic methods, while providing non-destructive, solvent-free measurements.

Overall, the findings support the conclusion that Raman spectroscopy, when combined with ANOVA-guided band selection, PCA, and machine learning, offers a viable and promising approach for rapid screening of chlorpyrifos residues in kale and milk. Although chromatographic techniques such as GC-MS remain essential as reference methods and for comprehensive multi-residue analysis, the results obtained here indicate that Raman-ML methods can provide screening performance that is compatible with practical monitoring needs, while offering advantages in speed, sample preservation, and reduced solvent use. At the same time, the study had two significant limitations in that it focused on a single active compound and used laboratory-spiked rather than field-collected samples. Nevertheless, the work demonstrates a clear pathway for integrating Raman spectroscopy and machine learning into more sustainable analytical approaches for pesticide-residue monitoring.

### 5.2 Recommendations and Future Prospects

Based on the findings of this study, it is recommended that Raman spectroscopy combined with machine learning be further developed and evaluated as a complementary screening tool for chlorpyrifos residue monitoring. Within practical monitoring systems, the approach is best positioned as a rapid, non-destructive frontline method for screening large numbers of samples, with only those flagged as potentially non-compliant referred to chromatographic techniques such as GC-MS or LC-MS for confirmatory analysis. Future work should focus on validating the Raman-ML approach using field-collected samples from farms and markets to assess robustness under realistic matrix variability and naturally occurring residue profiles. In addition, extending the method to other food matrices and additional pesticides or key degradation products would help establish its broader applicability. As portable Raman instruments become increasingly accessible, efforts should also be directed towards implementing and evaluating the developed models on handheld or field-deployable tools, to enable in situ screening and support more efficient and sustainable pesticide residue monitoring.

## REFERENCES


Abraham, J., & Silambarasan, S. (2016). Biodegradation of chlorpyrifos and its hydrolysis product 3,5,6-trichloro-2-pyridinol using a novel bacterium Ochrobactrum sp. JAS2: A proposal of its metabolic pathway. *Pesticide Biochemistry and Physiology*, *126*. https://doi.org/10.1016/j.pestbp.2015.07.001

Adugna, T., Xu, W., & Fan, J. (2022). Comparison of random forest and support vector machine classifiers for regional land cover mapping using Coarse Resolution FY-3C images. *Remote Sensing*, *14*(3), 574. https://doi.org/10.3390/rs14030574

Adum, A. N., Gicharu, G., Mwagandi, L. C., Sifuna, P. O., Essuman, S., & Nyabiba, M. A. (2021). Detection and Quantification of Chlorpyrifos in Soil, Milk, Dip Wash, Spray Race Residues Using High Performance Liquid Chromatography in Selected Dairy Farms in Kenya. *Science Journal of Analytical Chemistry*, *9*(4). https://doi.org/10.11648/j.sjac.20210904.12

Aftab, W., Moinuddin, M., & Shaikh, M. S. (2014). A novel kernel for RBF based neural networks. *Abstract and Applied Analysis*, *2014*. https://doi.org/10.1155/2014/176253

Ahuja, S. (2005). *Handbook of Pharmaceutical Analysis by HPLC ed. by Satinder Ahuja* . Elsevier Acad. Press, Amsterdam, 566 pp.

Alfawaz, H. A., Wani, K., Alrakayan, H., Alnaami, A. M., & Al-Daghri, N. M. (2022). Awareness, Knowledge and Attitude towards 'Superfood' Kale and Its Health Benefits among Arab Adults. *Nutrients*, *14*(2). https://doi.org/10.3390/nu14020245

Ali, A. A., & Hassan, K. I. (2023). Quantification of some pesticide residues in milk powder in Iraq by gas chromatography/ Mass Spectrometry. *Uttar Pradesh Journal of Zoology*, *44*(20), 130–137. https://doi.org/10.56557/upjoz/2023/v44i203654

Allegrini, F., & Olivieri, A. C. (2014). IUPAC-consistent approach to the limit of detection in partial least-squares calibration. *Analytical Chemistry*, *86*(15), 7858–7866. https://doi.org/10.1021/ac501786u

Alnuaimi, A. F. A. H., & Albaldawi, T. H. K. (2024). An overview of machine learning classification techniques. *BIO Web of Conferences*, *97*, 00133. https://doi.org/10.1051/bioconf/20249700133

Ambreen, S., & Yasmin, A. (2021). Novel degradation pathways for Chlorpyrifos and 3, 5, 6-Trichloro-2-pyridinol degradation by bacterial strain Bacillus thuringiensis MB497 isolated from agricultural fields of Mianwali, Pakistan. *Pesticide Biochemistry and Physiology*, *172*. https://doi.org/10.1016/j.pestbp.2020.104750

Amirav, A., Keshet, U., Alon, T., & Fialkov, A. B. (2014). Open probe fast GC–MS—real time analysis with separation. *International Journal of Mass Spectrometry*, *371*, 47–53. https://doi.org/10.1016/j.ijms.2014.08.002

Amjad, A., Ullah, R., Khan, S., Bilal, M., & Khan, A. (2018). Raman spectroscopy based analysis of milk using random forest classification. *Vibrational Spectroscopy*, *99*. https://doi.org/10.1016/j.vibspec.2018.09.003

Amores, G., & Virto, M. (2019). Total and free fatty acids analysis in milk and dairy fat. In *Separations 6*(1). https://doi.org/10.3390/separations6010014

Andreev, G. N., Schrader, B., Schulz, H., Fuchs, R., Popov, S., & Handjieva, N. (2001). Non-destructive NIR-FT-Raman analyses in practice. Part 1. Analyses of plants and historic textiles. *Analytical and Bioanalytical Chemistry*, *371*(7). https://doi.org/10.1007/s00216-001-1109-6

Asamba, M. N., Mugendi, E. N., Oshule, P. S., Essuman, S., Chimbevo, L. M., & Atego, N. A. (2022a). Molecular characterization of chlorpyrifos degrading bacteria isolated from contaminated dairy farm soils in Nakuru County, Kenya. *Heliyon*, *8*(3). https://doi.org/10.1016/j.heliyon.2022.e09176

Asamba, M. N., Oshule, P. S., Essuman, S., Atego, N., Chimbevo, L., Nderitu, J. H., & Mapesa, J. (2022b). Correlation between Chlorpyrifos Residues and Calcium Levels in Milk from Dairy Farms in Nakuru County, Kenya. *Africa Environmental Review Journal*, *5*(2), 77–88. http://ojs.uoeld.ac.ke/index.php/aerj/article/view/253

Baiz, C. R., Błasiak, B., Bredenbeck, J., Cho, M., Choi, J. H., Corcelli, S. A., Dijkstra, A. G., Feng, C. J., Garrett-Roe, S., Ge, N. H., Hanson-Heine, M. W. D., Hirst, J. D., Jansen, T. L. C., Kwac, K., Kubarych, K. J., Londergan, C. H., Maekawa, H., Reppert, M., Saito, S., … Zanni, M. T. (2020). Vibrational Spectroscopic Map, Vibrational Spectroscopy, and Intermolecular Interaction. In *Chemical Reviews 120*(15). https://doi.org/10.1021/acs.chemrev.9b00813

Baranska, M., Schulz, H., Baranski, R., Nothnagel, T., & Christensen, L. P. (2005). In situ simultaneous analysis of polyacetylenes, carotenoids and polysaccharides in carrot roots. *Journal of Agricultural and Food Chemistry*, *53*(17), 6565–6571. https://doi.org/10.1021/jf0510440

Barron, C., Robert, P., Guillon, F., Saulnier, L., & Rouau, X. (2006). Structural heterogeneity of Wheat Arabinoxylans revealed by Raman spectroscopy. *Carbohydrate Research*, *341*(9), 1186–1191. https://doi.org/10.1016/j.carres.2006.03.025

Bayona, J. M., & Pawliszyn, J. (2012). Comprehensive sampling and sample preparation. *Analytical Techniques for Scientists*, *1*. https://doi.org/10.1016/c2009-1-60918-7

Beattie, J. R., & Esmonde-White, F. W. L. (2021). Exploration of Principal Component Analysis: Deriving Principal Component Analysis Visually Using Spectra. In *Applied Spectroscopy 75*(4). https://doi.org/10.1177/0003702820987847

Bedi, J. S., Gill, J. P. S., Kaur, P., & Aulakh, R. S. (2018). Pesticide residues in milk and their relationship with pesticide contamination of feedstuffs supplied to dairy cattle in

Punjab (India). *Journal of Animal and Feed Sciences*, *27*(1). https://doi.org/10.22358/jafs/82623/2018

Bocklitz, T., Walter, A., Hartmann, K., Rösch, P., & Popp, J. (2011). How to pre-process Raman spectra for reliable and stable models? *Analytica Chimica Acta*, *704*(1–2). https://doi.org/10.1016/j.aca.2011.06.043

Boqué, R., & Heyden, Y. V. (2009). *The limit of detection*. Chromatography Online. https://www.chromatographyonline.com/view/limit-detection

Brewer, P. G., & Kirkwood, W. J. (2013). Raman spectroscopy for subsea applications. In *Subsea Optics and Imaging*. https://doi.org/10.1533/9780857093523.3.409

Bukasov, R., Sultangaziyev, A., Kunushpayeva, Z., Rapikov, A., & Dossym, D. (2023). Aluminum Foil vs. Gold Film: Cost-Effective Substrate in Sandwich SERS Immunoassays of Biomarkers Reveals Potential for Selectivity Improvement. *International Journal of Molecular Sciences*, *24*(6). https://doi.org/10.3390/ijms24065578

Bumbrah, G. S., & Sharma, R. M. (2016). Raman spectroscopy – Basic principle, instrumentation and selected applications for the characterization of drugs of abuse. In *Egyptian Journal of Forensic Sciences* *6*(3). https://doi.org/10.1016/j.ejfs.2015.06.001

Cao, Y., Shen, D., Lu, Y., & Huang, Y. (2006). A Raman-scattering study on the net orientation of biomacromolecules in the outer epidermal walls of mature wheat stems (triticum aestivum). *Annals of Botany*, *97*(6), 1091–1094. https://doi.org/10.1093/aob/mcl059

Chen, J., Dong, D., & Ye, S. (2018). Detection of pesticide residue distribution on fruit surfaces using surface-enhanced Raman spectroscopy imaging. *RSC Advances*, *8*(9). https://doi.org/10.1039/c7ra11927e

Chen, M., Zeng, H., Larkum, A. W. D., & Cai, Z. L. (2004). Raman properties of chlorophyll d, the major pigment of Acaryochloris marina: Studies using both Raman spectroscopy and density functional theory. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, *60*(3). https://doi.org/10.1016/S1386-1425(03)00258-0

Chen, S., Liu, C., Peng, C., Liu, H., Hu, M., & Zhong, G. (2012). Biodegradation of Chlorpyrifos and Its Hydrolysis Product 3,5,6-Trichloro-2-Pyridinol by a New Fungal Strain Cladosporium cladosporioides Hu-01. *PLoS ONE*, *7*(10). https://doi.org/10.1371/journal.pone.0047205

Chen, Z., Dong, X., Liu, C., Wang, S., Dong, S., & Huang, Q. (2023). Rapid detection of residual chlorpyrifos and pyrimethanil on fruit surface by surface-enhanced Raman spectroscopy integrated with deep learning approach. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-45954-y

Christensen, K., Harper, B., Luukinen, B., Buhl, K., & Stone, D. (2009). *Chlorpyrifos Technical Fact Sheet*. National Pesticide Information Center, Oregon State University Extension Services. Available from https://npic.orst.edu/factsheets/chlorpgen.html

Chu, C., Hsu, A. L., Chou, K. H., Bandettini, P., & Lin, C. P. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage*, *60*(1). https://doi.org/10.1016/j.neuroimage.2011.11.066

Contreras-Caceres, R., Sierra-Martin, B., & Fernandez-Barbero, A. (2011). Surface-Enhanced Raman Scattering Sensors based on Hybrid Nanoparticles. In *Microsensors*. https://doi.org/10.5772/18735

Dasriya, V., Joshi, R., Ranveer, S., Dhundale, V., Kumar, N., & Raghu, H. V. (2021). Rapid detection of pesticide in milk, cereal and cereal based food and fruit juices using paper strip-based sensor. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-96999-w

de Andrade, J. C., Galvan, D., Kato, L. S., & Conte-Junior, C. A. (2023). Consumption of fruits and vegetables contaminated with pesticide residues in Brazil: A systematic review with Health Risk Assessment. *Chemosphere*, *322*, 138244. https://doi.org/10.1016/j.chemosphere.2023.138244

Devitt, G., Howard, K., Mudher, A., & Mahajan, S. (2018). Raman spectroscopy: An emerging tool in neurodegenerative disease research and diagnosis. *ACS Chemical Neuroscience*, *9*(3), 404–420. https://doi.org/10.1021/acschemneuro.7b00413

Dhanani, T., Dou, T., Biradar, K., Jifon, J., Kurouski, D., & Patil, B. S. (2022). Raman Spectroscopy Detects Changes in Carotenoids on the Surface of Watermelon Fruits During Maturation. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.832522

Dhiraj, S. U. D., Kumar, J., Kaur, P., & Bansal, P. (2020). Toxicity, natural and induced degradation of chlorpyrifos. *Journal of the Chilean Chemical Society*, *65*(2). https://doi.org/10.4067/S0717-97072020000204807

Ding, X., Liu, J., Yang, F., & Cao, J. (2021). Random radial basis function kernel-based support vector machine. *Journal of the Franklin Institute*, *358*(18). https://doi.org/10.1016/j.jfranklin.2021.10.005

Dowgiallo, A. M., & Guenther, D. A. (2019). Determination of the limit of detection of multiple pesticides utilizing gold nanoparticles and surface-enhanced Raman spectroscopy. *Journal of Agricultural and Food Chemistry*, *67*(46), 12642–12651. https://doi.org/10.1021/acs.jafc.9b01544

Du, X., Wang, P., Fu, L., Liu, H., Zhang, Z., & Yao, C. (2020). Determination of Chlorpyrifos in Pears by Raman Spectroscopy with Random Forest Regression Analysis. *Analytical Letters*, *53*(6). https://doi.org/10.1080/00032719.2019.1681439

East Africa Natural History Society (EANHS). (2021, December 2). *Insecticides recommended for withdrawal in the Kenyan market. Nature Kenya - Connecting people with nature.* Available from https://naturekenya.org/2021/12/02/insecticides-recommended-for-withdrawal-in-the-kenyan-market/

Edwards, H. G. M., Farwell, D. W., & Webster, D. (1997). Ft raman microscopy of untreated natural plant fibres. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *53*(13), 2383–2392. https://doi.org/10.1016/s1386-1425(97)00178-9

Elsaadani, M., A. Abdel-Hakeem, M., Gamal, N., & Montet, D. (2025). Advancements in mycotoxin detection technologies: Safeguarding beverage quality and consumer health. *Emerging Trends in Beverage Industry [Working Title]*. https://doi.org/10.5772/intechopen.1011863

Ember, K. J., Hoeve, M. A., McAughtrie, S. L., Bergholt, M. S., Dwyer, B. J., Stevens, M. M., Faulds, K., Forbes, S. J., & Campbell, C. J. (2017). Raman spectroscopy and Regenerative Medicine: A Review. *Npj Regenerative Medicine*, *2*(1). https://doi.org/10.1038/s41536-017-0014-3

Eskenazi, B., Kogut, K., Huen, K., Harley, K. G., Bouchard, M., Bradman, A., Boyd-Barr, D., Johnson, C., & Holland, N. (2014). Organophosphate pesticide exposure, PON1, and neurodevelopment in school-age children from the CHAMACOS study. *Environmental Research*, *134*, 149–157. https://doi.org/10.1016/j.envres.2014.07.001

Esturk, O., Yakar, Y., & Ayhan, Z. (2014). Pesticide residue analysis in parsley, lettuce and spinach by LC-MS/MS. *Journal of Food Science and Technology*, *51*(3). https://doi.org/10.1007/s13197-011-0531-9

Foody, G. M. (2009). Sample size determination for image classification accuracy assessment and comparison. *International Journal of Remote Sensing*, *30*(20). https://doi.org/10.1080/01431160903130937

Foong, S. Y., Ma, N. L., Lam, S. S., Peng, W., Low, F., Lee, B. H. K., Alstrup, A. K. O., & Sonne, C. (2020). A recent global review of hazardous chlorpyrifos pesticide in fruit and vegetables: Prevalence, remediation and actions needed. *Journal of Hazardous Materials*, *400*(12300), 6. https://doi.org/10.1016/j.jhazmat.2020.123006

Frąszczak, B., Kula-Maximenko, M., Podsędek, A., Sosnowska, D., Unegbu, K. C., & Spiżewski, T. (2023). Morphological and Photosynthetic Parameters of Green and Red Kale Microgreens Cultivated under Different Light Spectra. *Plants*, *12*(22). https://doi.org/10.3390/plants12223800

Gierlinger, N., & Schwanninger, M. (2006). Chemical imaging of Poplar Wood Cell Walls by confocal Raman Microscopy. *Plant Physiology*, *140*(4), 1246–1254. https://doi.org/10.1104/pp.105.066993

Gierlinger, N., Keplinger, T., & Harrington, M. (2012). Imaging of plant cell walls by confocal Raman microscopy. *Nature Protocols*, *7*(9). https://doi.org/10.1038/nprot.2012.092

Gill, J. P. S., Bedi, J. S., Singh, R., Fairoze, M. N., Hazarika, R. A., Gaurav, A., Satpathy, S. K., Chauhan, A. S., Lindahl, J., Grace, D., Kumar, A., & Kakkar, M. (2020). Pesticide Residues in Peri-Urban Bovine Milk from India and Risk Assessment: A Multicenter Study. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-65030-z

Grewal, M. K., Huppertz, T., & Vasiljevic, T. (2018). FTIR fingerprinting of structural changes of milk proteins induced by heat treatment, deamidation and dephosphorylation. *Food Hydrocolloids*, *80*. https://doi.org/10.1016/j.foodhyd.2018.02.010

Gupta, S., Huang, C. H., Singh, G. P., Park, B. S., Chua, N. H., & Ram, R. J. (2020). Portable Raman leaf-clip sensor for rapid detection of plant stress. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-76485-5

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1). https://doi.org/10.20982/tqmp.08.1.p023

Hanson, B. (2024). *An introduction to ChemoSpec*. An introduction to ChemoSpec–CRAN. https://cran.r-project.org/web/packages/ChemoSpec/vignettes/ChemoSpec.pdf

Hayes, T., Usami, S., Jacobucci, R., & McArdle, J. J. (2015). Using classification and regression trees (CART) and random forests to analyze attrition: Results from two simulations. *Psychology and Aging*, *30*(4). https://doi.org/10.1037/pag0000046

He, W., Li, B., & Yang, S. (2020). High-Frequency Raman Analysis in Biological Tissues Using Dual-Wavelength Excitation Raman Spectroscopy. *Applied Spectroscopy*, *74*(2). https://doi.org/10.1177/0003702819881762

Hongsibsong, S., Prapamontol, T., Xu, T., Hammock, B. D., Wang, H., Chen, Z. J., & Xu, Z. L. (2020). Monitoring of the organophosphate pesticide chlorpyrifos in vegetable samples from local markets in Northern Thailand by developed immunoassay. *International Journal of Environmental Research and Public Health*, *17*(13). https://doi.org/10.3390/ijerph17134723

Hou, K., Yang, Y., Zhu, L., Wu, R., Du, Z., Li, B., Zhu, L., & Sun, S. (2022). Toxicity evaluation of chlorpyrifos and its main metabolite 3,5,6-trichloro-2-pyridinol (TCP) to Eisenia fetida in different soils. *Comparative Biochemistry and Physiology Part - C: Toxicology and Pharmacology*, *259*. https://doi.org/10.1016/j.cbpc.2022.109394

Human Rights Watch. (2023, September 14). *Kenya: Ban use of highly hazardous pesticides*. https://www.hrw.org/news/2023/09/14/kenya-ban-use-highly-hazardous-pesticides

Ikedi, R. I. O., Birech, Z., & Kaniu, M. I. (2023). Rapid Assessment of Molasses Adulterated Honey Using Laser Raman Spectroscopy and Principal Component Analysis. *Food Analytical Methods*, *16*(11–12). https://doi.org/10.1007/s12161-023-02538-w

Inonda, R., Njage, E., Ngeranwa, J., & Mutai, C. (2015). Determination of Pesticide Residues in Locally Consumed Vegetables in Kenya. *Afr. J. Pharmacol. Ther*, *4*(1), 1–6. http://journals.uonbi.ac.ke/ajpt

International Programme on Chemical Safety. (1972). *Chlorpyrifos*. IPCS INCHEM. Available from https://www.inchem.org/documents/jmpr/jmpmono/v072pr10.htm

Jacob, S. S., Lukose, J., Bankapur, A., Mithun, N., Vani Lakshmi, R., Acharya, M., Rao, P., Kamath, A., Baby, P. M., Rao, R. K., & Chidangil, S. (2022). Micro-Raman spectroscopy study of optically trapped erythrocytes in malaria, dengue and leptospirosis infections. *Frontiers in Medicine*, *9*. https://doi.org/10.3389/fmed.2022.858776

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, 426 pp.

Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PLoS ONE*, *13*(8). https://doi.org/10.1371/journal.pone.0201904

Jones, R. R., Hooper, D. C., Zhang, L., Wolverson, D., & Valev, V. K. (2019). Raman Techniques: Fundamentals and Frontiers. In *Nanoscale Research Letters 14*(1). https://doi.org/10.1186/s11671-019-3039-2

Kauffmann, T. H., Kokanyan, N., & Fontana, M. D. (2019). Use of Stokes and anti-Stokes Raman scattering for new applications. *Journal of Raman Spectroscopy*, *50*(3). https://doi.org/10.1002/jrs.5523

Keresztury, G. (2006). Raman Spectroscopy: Theory. *Handbook of Vibrational Spectroscopy*, 71–87. https://doi.org/10.1002/0470027320.s0109

Khalid, W., Iqra, Afzal, F., Rahim, M. A., Abdul Rehman, A., Faiz ul Rasul, H., Arshad, M. S., Ambreen, S., Zubair, M., Safdar, S., Al-Farga, A., & Refai, M. (2023). Industrial applications of Kale (*brassica oleracea var. Sabellica* ) as a functional ingredient: A Review. *International Journal of Food Properties*, *26*(1), 489–501. https://doi.org/10.1080/10942912.2023.2168011

Khan, H. M. H., McCarthy, U., Esmonde-White, K., Casey, I., & O'Shea, N. (2023). Potential of Raman spectroscopy for in-line measurement of raw milk composition. *Food Control*, *152*. https://doi.org/10.1016/j.foodcont.2023.109862

Kharabsheh, H. A., Han, S., Allen, S., & Chao, S. L. (2017). Metabolism of chlorpyrifos by Pseudomonas aeruginosa increases toxicity in adult zebrafish (Danio rerio).

*International Biodeterioration and Biodegradation*, *121*. https://doi.org/10.1016/j.ibiod.2017.03.024

Kherif, F., & Latypova, A. (2019). Principal component analysis. In *Machine Learning: Methods and Applications to Brain Disorders*. https://doi.org/10.1016/B978-0-12-815739-8.00012-2

Kucha, C. T., Liu, L., & Ngadi, M. O. (2018). Non-destructive spectroscopic techniques and multivariate analysis for assessment of fat quality in pork and pork products: A review. *Sensors (Switzerland)*, *18*(2), 377. https://doi.org/10.3390/s18020377

Kuhar, N., Sil, S., Verma, T., & Umapathy, S. (2018). Challenges in application of Raman spectroscopy to biology and materials. In *RSC Advances 8*(46). https://doi.org/10.1039/c8ra04491k

Landi, N., Ragucci, S., & Di Maro, A. (2021). Amino acid composition of milk from cow, sheep and goat raised in ailano and valle agricola, two localities of 'alto casertano' (Campania region). *Foods*, *10*(10). https://doi.org/10.3390/foods10102431

Langer, J., de Aberasturi, D. J., Aizpurua, J., Alvarez-Puebla, R. A., Auguié, B., Baumberg, J. J., Bazan, G. C., Bell, S. E. J., Boisen, A., Brolo, A. G., Choo, J., Cialla-May, D., Deckert, V., Fabris, L., Faulds, K., Javier García de Abajo, F., Goodacre, R., Graham, D., Haes, A. J., … Liz-Marzán, L. M. (2020). Present and future of surface-enhanced Raman scattering. In *ACS Nano 14*(1). https://doi.org/10.1021/acsnano.9b04224

Lee, K., Yarbrough, D., Kozman, M., Herrman, T., Park, J., Wang, R., & Kurouski, D. (2020). Sensitive SERS Characterization and Analysis of Chlorpyrifos and Aldicarb Residues in Animal Feed using Gold Nanoparticles. *Journal of Regulatory Science*, *8*. https://doi.org/10.21423/jrs-v08lee

Leskovac, A., & Petrović, S. (2023). Pesticide Use and Degradation Strategies: Food Safety, Challenges and Perspectives. *Foods*, *12*(14), 2709. https://doi.org/10.3390/foods12142709

Li, J., Liu, J., Shen, W., Zhao, X., Hou, Y., Cao, H., & Cui, Z. (2010). Isolation and characterization of 3,5,6-trichloro-2-pyridinol-degrading Ralstonia sp. strain T6. *Bioresource Technology*, *101*(19). https://doi.org/10.1016/j.biortech.2010.04.030

Li, J.-X., Qing, C.-C., Wang, X.-Q., Zhu, M.-J., Zhang, B.-Y., & Zhang, Z.-Y. (2024). Discriminative feature analysis of dairy products based on machine learning algorithms and Raman spectroscopy. *Current Research in Food Science*, *8*, 100782. https://doi.org/10.1016/j.crfs.2024.100782

Li, N., Hussain, N., Ding, Z., Qu, C., Li, Y., Chu, L., & Liu, H. (2024). Guidelines for Raman spectroscopy and imaging techniques in food safety analysis. *Food Safety and Health*, *2*(2), 221–237. https://doi.org/10.1002/fsh3.12040

Li, X., Zhang, S., Yu, Z., & Yang, T. (2014). Surface-Enhanced Raman Spectroscopic Analysis of Phorate and Fenthion Pesticide in Apple Skin Using Silver Nanoparticles. *Applied Spectroscopy*, *68*(4). https://doi.org/10.1366/13-07080

Ling, Y., Wang, H., Yong, W., Zhang, F., Sun, L., Yang, M. L., Wu, Y. N., & Chu, X. G. (2011). The effects of washing and cooking on chlorpyrifos and its toxic metabolites in vegetables. *Food Control*, *22*(1). https://doi.org/10.1016/j.foodcont.2010.06.009

Liu, Y. De, Zhang, Y. X., Wang, H. Y., & Ye, B. (2016). Detection of pesticides on navel orange skin by surface-enhanced Raman spectroscopy coupled with Ag nanostructures. *International Journal of Agricultural and Biological Engineering*, *9*(2). https://doi.org/10.3965/j.ijabe.20160902.1960

Liu, Y., He, B., Zhang, Y., Wang, H., & Ye, B. (2015). Detection of Phosmet Residues on Navel Orange Skin by Surface-enhanced Raman Spectroscopy. *Intelligent Automation and Soft Computing*, *21*(3). https://doi.org/10.1080/10798587.2015.1015770

Lorena, A. C., & de Carvalho, A. C. P. L. F. (2008). Evolutionary tuning of SVM parameter values in multiclass problems. *Neurocomputing*, *71*(16–18), 3326–3334. https://doi.org/10.1016/j.neucom.2008.01.031

Lussier, F., Thibault, V., Charron, B., Wallace, G. Q., & Masson, J. F. (2020). Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering. *TrAC - Trends in Analytical Chemistry*, *124*(11579), 6. https://doi.org/10.1016/j.trac.2019.115796

Ma, P., Wang, L., Xu, L., Li, J., Zhang, X., & Chen, H. (2020). Rapid quantitative determination of chlorpyrifos pesticide residues in tomatoes by surface-enhanced Raman spectroscopy. *European Food Research and Technology*, *246*(1). https://doi.org/10.1007/s00217-019-03408-8

Ma, Y., & Guo, G. (2014). *Support Vector Machines Applications*. https://doi.org/10.1007/978-3-319-02300-7

Mali, H., Shah, C., Raghunandan, B. H., Prajapati, A. S., Patel, D. H., Trivedi, U., & Subramanian, R. B. (2023). Organophosphate pesticides an emerging environmental contaminant: Pollution, toxicity, bioremediation progress, and remaining challenges. In *Journal of Environmental Sciences (China)* 127. https://doi.org/10.1016/j.jes.2022.04.023

Marchevsky, A. M., Walts, A. E., Lissenberg-Witte, B. I., & Thunnissen, E. (2020). Pathologists should probably forget about kappa. Percent agreement, diagnostic specificity and related metrics provide more clinically applicable measures of interobserver variability. *Annals of Diagnostic Pathology*, *47*. https://doi.org/10.1016/j.anndiagpath.2020.151561

Mayorga, C., Athalye, S. M., Boodaghidizaji, M., Sarathy, N., Hosseini, M., Ardekani, A., & Verma, M. S. (2025). Limit of detection of Raman spectroscopy using polystyrene

particles from 25 to 1000 nm in aqueous suspensions. *Analytical Chemistry*, *97*(16), 8908–8914. https://doi.org/10.1021/acs.analchem.5c00182

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3). https://doi.org/10.11613/bm.2012.031

Mebdoua, S. (2019). Pesticide Residues in Fruits and Vegetables. *Reference Series in Phytochemistry*, 1715–1753. https://doi.org/10.1007/978-3-319-78030-6_76

Menges, F. (2022). *SpectraGryph: optical spectroscopy software* (Version 1.2.16.1). http://www.effemm2.de/spectragryph/

Mikac, L., Kovačević, E., Ukić, Raić, M., Jurkin, T., Marić, I., Gotić, M., & Ivanda, M. (2021). Detection of multi-class pesticide residues with surface-enhanced Raman spectroscopy. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, *252*. https://doi.org/10.1016/j.saa.2021.119478

Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., Panda, S., & Laishram, M. (2017). Multivariate Statistical Data Analysis-Principal Component Analysis (PCA). *International Journal of Livestock Research*, *7*(5), 1. https://doi.org/10.5455/ijlr.20170415115235

Momtaz, M., & Khan, M. S. (2024). Analysis of chlorpyrifos pesticide residue in locally grown cauliflower, cabbage, and eggplant using gas chromatography–mass spectrometry (GC-MS) technique: A bangladesh perspective. *Foods*, *13*(11), 1780. https://doi.org/10.3390/foods13111780

Momtaz, M., & Khan, M. S. (2024). Analysis of chlorpyrifos pesticide residue in locally grown cauliflower, cabbage, and eggplant using gas chromatography–mass spectrometry (GC-MS) technique: A bangladesh perspective. *Foods*, *13*(11), 1780. https://doi.org/10.3390/foods13111780

Mwendwa, M. (2023). *Why banned toxic pesticides from EU markets are a concern for Cameroon and Kenya*. Journalismfund Europe. https://www.journalismfund.eu/supported-projects/why-banned-toxic-pesticides-eu-markets-are-concern-cameroon-and-kenya

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons.B*, *4*, 51–62. https://doi.org/10.20544/horizons.b.04.1.17.p05

Oleneva, E., Khaydukova, M., Ashina, J., Yaroshenko, I., Jahatspanian, I., Legin, A., & Kirsanov, D. (2019). A simple procedure to assess limit of detection for Multisensor Systems. *Sensors*, *19*(6), 1359. https://doi.org/10.3390/s19061359

Olusanya, M. O., Ogunsakin, R. E., Ghai, M., & Adeleke, M. A. (2022). Accuracy of Machine Learning Classification Models for the Prediction of Type 2 Diabetes Mellitus: A Systematic Survey and Meta-Analysis Approach. In *International Journal*

*of Environmental Research and Public Health 19*(21). https://doi.org/10.3390/ijerph192114280

Opazo-Navarrete, M., Burgos-Díaz, C., Soto-Cerda, B., Barahona, T., Anguita-Barrales, F., & Mosi-Roa, Y. (2021). Assessment of the Nutritional Value of Traditional Vegetables from Southern Chile as Potential Sources of Natural Ingredients. *Plant Foods for Human Nutrition*, *76*(4). https://doi.org/10.1007/s11130-021-00935-2

OriginPro. (2024). *OriginPro* (Version 2024b) [Computer software]. https://www.originlab.com

Peris-Vicente, J., Peris-García, E., Albiol-Chiva, J., Durgbanshi, A., Ochoa-Aranda, E., Carda-Broch, S., Bose, D., & Esteve-Romero, J. (2022). Liquid chromatography, a valuable tool in the determination of antibiotics in biological, food and environmental samples. *Microchemical Journal, 177,* 107309. https://doi.org/10.1016/j.microc.2022.107309

Pham, U. T., Phan, Q. H. T., Nguyen, L. P., Luu, P. D., Doan, T. D., Trinh, H. T., Dinh, C. T., Van Nguyen, T., Tran, T. Q., Le, D. X., Pham, T. N., Le, T. D., & Nguyen, D. T. (2022). Rapid Quantitative Determination of Multiple Pesticide Residues in Mango Fruits by Surface-Enhanced Raman Spectroscopy. *Processes*, *10*(3). https://doi.org/10.3390/pr10030442

Pilot, R., Signorini, R., Durante, C., Orian, L., Bhamidipati, M., & Fabris, L. (2019). A review on surface-enhanced Raman scattering. In *Biosensors 9*(2). https://doi.org/10.3390/bios9020057

Pimenta, S., & Correia, J. H. (2025). Biomedical applications of raman spectroscopy: A Review. *Photochem, 5*(4), 29. https://doi.org/10.3390/photochem5040029

R Core Team. (2024). *A language and environment for statistical computing. R Foundation for Statistical Computing*. https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf

Raj, A., & Kumar, A. (2022). Recent advances in assessment methods and mechanism of microbe-mediated chlorpyrifos remediation. *Environmental Research*, *214*. https://doi.org/10.1016/j.envres.2022.114011

Rekha, P. R. (2005). *Dissipation of Chlorpyrifos in Red Loam Soil and its Effect on Soil Organisms*. *Journal of Environmental Science and Health, Part B*, *39*(4), 517–531. https://doi.org/10.1081/pfc-200026697

Rodriguez-Saona, L., Ayvaz, H., & Wehling, R. L. (2017). Infrared and Raman Spectroscopy. In *Food analysis* (pp. 107–127). https://doi.org/10.1201/b15995-48

Rusu, E. A., & Baia, M. (2023). Moving from Raman Spectroscopy Lab towards Analytical Applications: A Review of Interlaboratory Studies. In *Instruments 7*(4). https://doi.org/10.3390/instruments7040030

Šamec, D., Urlić, B., & Salopek-Sondi, B. (2019). Kale (Brassica oleracea var. acephala) as a superfood: Review of the scientific evidence behind the statement. In *Critical Reviews in Food Science and Nutrition 59*(15). https://doi.org/10.1080/10408398.2018.1454400

Sanchez, L., Pant, S., Xing, Z., Mandadi, K., & Kurouski, D. (2019). Rapid and noninvasive diagnostics of Huanglongbing and nutrient deficits on citrus trees with a handheld Raman spectrometer. *Analytical and Bioanalytical Chemistry*, *411*(14). https://doi.org/10.1007/s00216-019-01776-4

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Sass, J. (2022, April 20). *EPA bans chlorpyrifos on food crops.* NRDC. https://www.nrdc.org/bio/jennifer-sass/epa-bans-chlorpyrifos-food-crops

Satheesh, N., & Fanta, S. W. (2020). Kale: Review on nutritional composition, bio-active compounds, anti-nutritional factors, health beneficial properties and value-added products. *Cogent Food & Agriculture*, *6*(1), 1811048. https://doi.org/10.1080/23311932.2020.1811048

Schulte, C. (2021, May 20). *Canada bans use of toxic pesticide*. Human Rights Watch. https://www.hrw.org/news/2021/05/20/canada-bans-use-toxic-pesticide

Schulz, H., & Baranska, M. (2007). Identification and quantification of valuable plant substances by IR and Raman spectroscopy. *Vibrational Spectroscopy*, *43*(1), 13–25. https://doi.org/10.1016/j.vibspec.2006.06.001

Schulze, H. G., Rangan, S., Vardaki, M. Z., Blades, M. W., Turner, R. F. B., & Piret, J. M. (2022). Critical Evaluation of Spectral Resolution Enhancement Methods for Raman Hyperspectra. *Applied Spectroscopy*, *76*(1). https://doi.org/10.1177/00037028211061174

Seki, T., Chiang, K. Y., Yu, C. C., Yu, X., Okuno, M., Hunger, J., Nagata, Y., & Bonn, M. (2020). The bending mode of water: A powerful probe for hydrogen bond structure of aqueous systems. In *Journal of Physical Chemistry Letters 11*(19). https://doi.org/10.1021/acs.jpclett.0c01259

Sen, R., & Das, S. (2023). Unsupervised Learning. In *Indian Statistical Institute Series*. https://doi.org/10.1007/978-981-19-2008-0_21

Shaker, E. M., & Elsharkawy, E. E. (2015). Organochlorine and organophosphorus pesticide residues in raw buffalo milk from agroindustrial areas in Assiut, Egypt. *Environmental Toxicology and Pharmacology*, *39*(1). https://doi.org/10.1016/j.etap.2014.12.005

Sheats, J. L., & Middlestadt, S. E. (2013). Salient beliefs about eating and buying dark green vegetables as told by Mid-western African-American women. *Appetite*, *65*. https://doi.org/10.1016/j.appet.2013.02.001

Sheridan, R. S., & Meola, J. R. (1999). Analysis of pesticide residues in fruits, vegetables, and milk by gas chromatography/tandem mass spectrometry. *Journal of AOAC International*, *82*(4). https://doi.org/10.1093/jaoac/82.4.982

Shi, G., Shen, X., Ren, H., Rao, Y., Weng, S., & Tang, X. (2022). Kernel principal component analysis and differential non-linear feature extraction of pesticide residues on fruit surface based on surface-enhanced Raman spectroscopy. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.956778

Shipp, D. W., Sinjab, F., & Notingher, I. (2017). Raman spectroscopy: techniques and applications in the life sciences. *Advances in Optics and Photonics*, *9*(2). https://doi.org/10.1364/aop.9.000315

Silva, M. G., de Paula, I. L., Stephani, R., Edwards, H. G. M., & de Oliveira, L. F. C. (2021). Raman spectroscopy in the quality analysis of dairy products: A literature review. In *Journal of Raman Spectroscopy 52*(12). https://doi.org/10.1002/jrs.6214

Singh, K. S., Majik, M. S., & Tilvi, S. (2014). *Vibrational Spectroscopy for Structural Characterization of Bioactive Compounds* (pp. 115–148). https://doi.org/10.1016/B978-0-444-63359-0.00006-9

Stocka, J., Biziuk, M., & Namiešnik, J. (2016). Analysis of pesticide residue in fruits and vegetables using analytical protocol based on application of the QuEChERS technique and GC-ECD system. *International Journal of Global Environmental Issues*, *15*(1–2). https://doi.org/10.1504/IJGENVI.2016.074361

Sultangaziyev, A., Akhmetova, A., Kunushpayeva, Z., Rapikov, A., Filchakova, O., & Bukasov, R. (2020). Aluminum foil as a substrate for metal enhanced fluorescence of bacteria labelled with quantum dots, shows very large enhancement and high contrast. *Sensing and Bio-Sensing Research*, *28*. https://doi.org/10.1016/j.sbsr.2020.100332

Sun, L., Yu, Z., Alsammarraie, F. K., Lin, M. H., Kong, F., Huang, M., & Lin, M. (2021). Development of cellulose Nanofiber-based substrates for rapid detection of ferbam in kale by Surface-enhanced Raman spectroscopy. *Food Chemistry*, *347*. https://doi.org/10.1016/j.foodchem.2021.129023

Surzhykov, A., Yerokhin, V. A., Stöhlker, T., & Fritzsche, S. (2015). Rayleigh x-ray scattering from many-electron atoms and ions. *Journal of Physics B: Atomic, Molecular and Optical Physics*, *48*(14). https://doi.org/10.1088/0953-4075/48/14/144015

Synytsya, A., Čopíková, J., Matějka, P., & Machovič, V. (2003). Fourier transform Raman and infrared spectroscopy of pectins. *Carbohydrate Polymers*, *54*(1), 97–106. https://doi.org/10.1016/s0144-8617(03)00158-9

Tandon, H., Chakraborty, T., & Suhag, V. (2019). A new scale of atomic static dipole polarizability invoking other periodic descriptors. *Journal of Mathematical Chemistry*, *57*(9). https://doi.org/10.1007/s10910-019-01055-8

Tao, M., Fang, H., Feng, X., He, Y., Liu, X., Shi, Y., Wei, Y., & Hong, Z. (2022). Rapid Trace Detection of Pesticide Residues on Tomato by Surface-Enhanced Raman Spectroscopy and Flexible Tapes. *Journal of Food Quality*, *2022*. https://doi.org/10.1155/2022/6947775

Terrones, O., Olazar-Intxausti, J., Anso, I., Lorizate, M., Nieto-Garai, J. A., & Contreras, F. X. (2023). Raman Spectroscopy as a Tool to Study the Pathophysiology of Brain Diseases. In *International Journal of Molecular Sciences 24*(3). https://doi.org/10.3390/ijms24032384

The United Nations Environment Programme (UNEP). (2022). *Chlorpyrifos: Draft risk profile*. https://assets.publishing.service.gov.uk/media/626a62908fa8f57a3cdbbe6d/Chlorpyrifos_draft_risk_profile.pdf

Thuku, J. M., Kaniu, M. I., Ndung'u, C. N., Kiruri, L. W., & Kaduki, K. A. (2025a). Rapid trace detection of chlorpyrifos in vegetables using 2d raman correlation spectroscopy and machine learning. *LWT, 237,* 118742. https://doi.org/10.1016/j.lwt.2025.118742

Thuku, J. M., Kaniu, M. I., Ndung'u, C. N., Kiruri, L. W., & Kaduki, K. A. (2025b). Targeted detection of low-level chlorpyrifos residues in milk using chemometrics-aided Raman spectroscopy. *European Journal of Advanced Chemistry Research, 6*(3), 1–12. https://doi.org/10.24018/ejchem.2025.6.3.165

Thyr, J., & Edvinsson, T. (2023). Evading the Illusions: Identification of False Peaks in Micro-Raman Spectroscopy and Guidelines for Scientific Best Practice. *Angewandte Chemie*, *135*(43). https://doi.org/10.1002/ange.202219047

U.S. Environmental Protection Agency. (2020). *Chlorpyrifos proposed interim registration review decision: Case number 0100.* U.S. Environmental Protection Agency. https://www.epa.gov/sites/default/files/2020-12/documents/chlorpyrifos_pid_signed_120320.pdf

Ubaid ur Rahman, H., Asghar, W., Nazir, W., Sandhu, M. A., Ahmed, A., & Khalid, N. (2021). A comprehensive review on chlorpyrifos toxicity with special reference to endocrine disruption: Evidence of mechanisms, exposures and mitigation strategies. In *Science of the Total Environment* 755. https://doi.org/10.1016/j.scitotenv.2020.142649

Vandenabeele, P., Jehlička, J., Vítek, P., & Edwards, H. G. M. (2012). On the definition of Raman spectroscopic detection limits for the analysis of biomarkers in solid matrices. *Planetary and Space Science, 62*(1), 48–54. https://doi.org/10.1016/j.pss.2011.12.006

Vemuri, S. (2016). Dissipation pattern of chlorpyriphos, cypermethrin, Ethion, Profenophos and Triazophos in Curry Leaf. *Indian Journal of Applied Science*, *6*(11), 1–5. https://doi.org/10.15436/2377-0619.16.1127

Vettorazzi, G. (1977). Pesticide Residues in Food in the Context of Present and Future International Pesticide Managerial Approaches. In *Pesticide Management and Insecticide Resistance*. https://doi.org/10.1016/b978-0-12-738650-8.50010-2

Vujović, Ž. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6). https://doi.org/10.14569/IJACSA.2021.0120670

Walse, K. H., Dharaskar, R. V., & Thakare, V. M. (2016). PCA based optimal ANN classifiers for human activity recognition using mobile sensors data. *Smart Innovation, Systems and Technologies*, *50*. https://doi.org/10.1007/978-3-319-30933-0_43

Wang, K., Li, Z., Li, J., & Lin, H. (2021). Raman spectroscopic techniques for nondestructive analysis of agri-foods: A state-of-the-art review. In *Trends in Food Science and Technology* 118. https://doi.org/10.1016/j.tifs.2021.10.010

Wang, M., Niu, Y., Peng, H., Zhang, P., Bu, Q., Song, X., & Yuan, S. (2025). Nanomaterial-enabled Spectroscopic Sensing: Building a new paradigm for precision detection of pesticide residues. *Nanomaterials*, *15*(21), 1634. https://doi.org/10.3390/nano15211634

Wang, X., Jiang, S., Liu, Z., Sun, X., Zhang, Z., Quan, X., Zhang, T., Kong, W., Yang, X., & Li, Y. (2024). Integrated surface-enhanced Raman spectroscopy and convolutional neural network for quantitative and qualitative analysis of pesticide residues on pericarp. *Food Chemistry*, *440*. https://doi.org/10.1016/j.foodchem.2023.138214

Waras, M K., Ismail, B.I., & Lenehan, C. E. (2020). *Law and regulations to control pesticide exposure among the general population: Comparing the Australian and the European Union pesticide regulatory system*. Preprints.org. Available from https://www.preprints.org/manuscript/202012.0117

Warrens, M. J. (2011). Cohen's kappa is a weighted average. *Statistical Methodology*, *8*(6). https://doi.org/10.1016/j.stamet.2011.06.002

Weng, S., Zhu, W., Dong, R., Zheng, L., & Wang, F. (2019). Rapid detection of pesticide residues in paddy water using surface-enhanced raman spectroscopy. *Sensors (Switzerland)*, *19*(3). https://doi.org/10.3390/s19030506

Wołejko, E., Łozowicka, B., Jabłońska-Trypuć, A., Pietruszyńska, M., & Wydro, U. (2022). Chlorpyrifos Occurrence and Toxicological Risk Assessment: A Review. In *International Journal of Environmental Research and Public Health 19*(19). https://doi.org/10.3390/ijerph191912209

Xu, Q., Guo, X., Xu, L., Ying, Y., Wu, Y., Wen, Y., & Yang, H. (2017). Template-free synthesis of SERS-active gold nanopopcorn for rapid detection of chlorpyrifos residues. *Sensors and Actuators, B: Chemical*, *241*. https://doi.org/10.1016/j.snb.2016.11.021

Xue, Y., & Jiang, H. (2023). Monitoring of Chlorpyrifos Residues in Corn Oil Based on Raman Spectral Deep-Learning Model. *Foods*, *12*(12). https://doi.org/10.3390/foods12122402

Yang, A., Bai, Y., Liu, H., Jin, K., Xue, T., & Ma, W. (2022). Application of SVM and its Improved Model in Image Segmentation. *Mobile Networks and Applications*, *27*(3). https://doi.org/10.1007/s11036-021-01817-2

Yang, S. (2022). High-Wavenumber Raman Analysis. In *Recent Developments in Atomic Force Microscopy and Raman Spectroscopy for Materials Characterization*. https://doi.org/10.5772/intechopen.100474

Yazdanpanah, A., Revilla, R. I., Franceschi, M., Fabrizi, A., Khademzadeh, S., Khodabakhshi, M., De Graeve, I., & Dabalà, M. (2024). Unveiling the impact of laser power variations on microstructure, corrosion, and stress-assisted surface crack initiation in laser powder bed fusion-processed Ni-Fe-Cr alloy 718. *Electrochimica Acta*, *476*. https://doi.org/10.1016/j.electacta.2023.143723

Yazgan, N. N., Genis, H. E., Bulat, T., Topcu, A., Durna, S., Yetisemiyen, A., & Boyaci, I. H. (2020). Discrimination of milk species using Raman spectroscopy coupled with partial least squares discriminant analysis in raw and pasteurized milk. *Journal of the Science of Food and Agriculture*, *100*(13). https://doi.org/10.1002/jsfa.10534

Ye, W., Yan, T., Zhang, C., Duan, L., Chen, W., Song, H., Zhang, Y., Xu, W., & Gao, P. (2022). Detection of Pesticide Residue Level in Grape Using Hyperspectral Imaging with Machine Learning. *Foods*, *11*(11), 1609. https://doi.org/10.3390/foods11111609

Yu, M. M., Schulze, H. G., Jetter, R., Blades, M. W., & Turner, R. F. (2007). Raman microspectroscopic analysis of triterpenoids found in plant cuticles. *Applied Spectroscopy*, *61*(1), 32–37. https://doi.org/10.1366/000370207779701352

Yun, D.-Y., Bae, J.-Y., Kang, Y.-J., Lim, C.-U., Jang, G.-H., Eom, M.-O., & Choe, W.-J. (2024). Simultaneous analysis of 272 pesticides in agricultural products by the Quechers method and gas chromatography with Tandem Mass Spectrometry. *Molecules*, *29*(9), 2114. https://doi.org/10.3390/molecules29092114

Zedler, L., Hager, M. D., Schubert, U. S., Harrington, M. J., Schmitt, M., Popp, J., & Dietzek, B. (2014). Monitoring the chemistry of self-healing by vibrational spectroscopy - Current state and perspectives. In *Materials Today 17*(2). https://doi.org/10.1016/j.mattod.2014.01.020

Zhang, X., Zhou, Q., Huang, Y., Li, Z., & Zhang, Z. (2011). Contrastive analysis of the Raman spectra of polychlorinated benzene: Hexachlorobenzene and benzene. *Sensors*, *11*(12). https://doi.org/10.3390/s111211510

Zheng, G., Han, C., Liu, Y., Wang, J., Zhu, M., Wang, C., & Shen, Y. (2014). Multiresidue analysis of 30 organochlorine pesticides in milk and milk powder by gel permeation chromatography-solid phase extraction-gas chromatography-tandem mass spectrometry. *Journal of Dairy Science*, *97*(10). https://doi.org/10.3168/jds.2014-8192

Zhu, J., Agyekum, A. A., Kutsanedzie, F. Y. H., Li, H., Chen, Q., Ouyang, Q., & Jiang, H. (2018). Qualitative and quantitative analysis of chlorpyrifos residues in tea by surface-enhanced Raman spectroscopy (SERS) combined with chemometric models. *LWT*, *97*. https://doi.org/10.1016/j.lwt.2018.07.055

Zhu, X., Li, W., Wu, R., Liu, P., Hu, X., Xu, L., Xiong, Z., Wen, Y., & Ai, S. (2021). Rapid detection of chlorpyrifos pesticide residue in tea using surface-enhanced Raman spectroscopy combined with chemometrics. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, *250*. https://doi.org/10.1016/j.saa.2020.119366

Zou, M., Huang, M., Zhang, J., & Chen, R. (2022). Exploring the effects and mechanisms of organophosphorus pesticide exposure and hearing loss. In *Frontiers in Public Health* 10. https://doi.org/10.3389/fpubh.2022.1001760

## LIST OF APPENDICES

### Appendix I: Codes Used for Data Analysis

### PCA Code in R for Exploratory Analysis

```
# Load Required Libraries
library(ChemoSpecUtils)
library(ChemoSpec)
library(chemometrics)
library(knitr)
library(R.utils)
library(utils)
library(kableExtra)
library(ggplot2)
library(dplyr)
# Importing and Preparing the Spectral Data
rawspec <- matrix2SpectraObject(gr.crit = c("C.", "T."),
                                gr.cols = c("green", "red"),
                                freq.unit = "Raman shift (cm-1)",
                                int.unit = "Intensity",
                                descrip = "Milk Score plots for PA",
                                in.file = "Kale_PA_1000c.csv",
                                out.file = "Kale_PCA_data_PB",
                                chk = TRUE,
                                sep = ",",
                                dec = ".")
# Summarizing the Spectral Data
sumSpectra(rawspec)
# Select Regions of Interest (ROI)
newspec1 <- rawspec
# Normalization of Spectra
spec1 <- normSpectra(newspec1)
```

```
# Principal Component Analysis (PCA)
pca1 <- c_pcaSpectra(spec1, choice = "noscale", cent = TRUE)
# Plotting PCA Score Plots
p <- plotScores(spec1, pca1, pcs = c(1, 2), ellipse = "none", tol = 0)
p <- p + geom_vline(xintercept = 0) + geom_hline(yintercept = 0)
p
# Plot PCA Loadings for the full spectrum
plotLoadings(spec1, pca1, loads = c(1, 2), ref = 1)
# Outlier Detection in PCA
p <- diagnostics <- pcaDiag(spec1, pca1, quantile = 0.90, pcs = 2, plot = "OD")
p <- diagnostics <- pcaDiag(spec1, pca1, quantile = 0.90, pcs = 2, plot = "SD")
# Scree Plot
plotScree(pca2, style = "alt", main = "Scree Plot: 314-354 cm-1 Spectra")
```

**RF Code in R Used for Classification**

```
# Load Libraries
library(randomForest)
library(caret)
library(dplyr)
library(doSNOW)
# Load Data
df <- read.csv("pcs_data_all_PA_Milk_fingerprint.csv", stringsAsFactors = FALSE)
df <- df %>%
  mutate(Group = factor(Group, levels = c("Below", "MRL", "Above")))
df[, paste0("PC", 1:4)] <- lapply(df[, paste0("PC", 1:4)], as.numeric)
# Split Data
set.seed(1234)
data_split <- createDataPartition(df$Group, times = 1, p = 0.7, list = FALSE)
train_set <- df[data_split, ]
test_set <- df[-data_split, ]
# Train Random Forest Model
```

```
train.control <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
tune.grid <- expand.grid(.mtry = c(1:4))
cl <- makeCluster(2, type = "SOCK")
registerDoSNOW(cl)
RFmodel <- train(Group ~ .,
                 data = train_set[, c("Group", paste0("PC", 14))],
                 method = "rf",
                 tuneGrid = tune.grid,
                 trControl = train.control,
                 ntree = 500)
stopCluster(cl)
# Performance on Training Data
preds_train <- predict(RFmodel, train_set[, paste0("PC", 1:4)])
cm_train <- confusionMatrix(preds_train, train_set$Group, positive = "Above")
print(cm_train)
# Performance on Test Data
preds_test <- predict(RFmodel, test_set[, paste0("PC", 1:4)])
cm_test <- confusionMatrix(preds_test, test_set$Group, positive = "Above")
print(cm_test)
```

**RF Code in R Used for Regression**

```
library(randomForest)
library(mlbench)
library(caret)
library(dplyr)
library(e1071)
library(parallel)
library(doSNOW)
# Load data (replace with your dataset)
df <- read.csv("2pcs_data_milk_PA_quant_finger.csv", stringsAsFactors = FALSE)
```

```
# Ensure PC1, PC2, and Target (assuming 'Target' is the regression variable) are numeric
df <- df %>%
  mutate(
    Target = as.numeric(Level),
    PC1 = as.numeric(PC1),
    PC2 = as.numeric(PC2)
  )
# Handle missing values by removing them
df <- na.omit(df)  # Remove rows with any missing values
# Split Data
set.seed(1234)
data_split <- createDataPartition(df$Target, times = 1, p = 0.7, list = FALSE)
train_set <- df[data_split, ]
test_set <- df[-data_split, ]
# Set up caret to perform 10-fold cross-validation repeated 3 times for regression
train.control <- trainControl(method = "repeatedcv",
                              number = 10,
                              repeats = 3,
                              search = "grid",
                              verboseIter = TRUE)
# Define the grid of hyperparameters for Random Forest
tune.grid <- expand.grid(.mtry = c(1:2))
# Use the doSNOW package to enable caret to train in parallel
cl <- makeCluster(detectCores() - 1, type = "SOCK")
registerDoSNOW(cl)
# Train the Random Forest model using caret for regression
RFmodel <- train(Target ~ .,
                 data = train_set,
```

```
            method = "rf",
            tuneGrid = tune.grid,
            trControl = train.control,
            metric = "RMSE")
# Stop the cluster
stopCluster(cl)
# Make predictions on the test set
test_preds <- predict(RFmodel, test_set)
# Calculate standard deviation of predictions for each instance in test set for error bars
rf_model <- RFmodel$finalModel
test_pred_all <- predict(rf_model, test_set, predict.all = TRUE)$individual
test_errors <- apply(test_pred_all, 1, sd)
# Compute RMSEP and R² for the test set
test_rmsep <- sqrt(mean((test_preds - test_set$Target)^2))
test_r2 <- cor(test_preds, test_set$Target)^2
# Select a subset of points for displaying error bars
set.seed(123)  # For reproducibility
error_indices <- sample(1:nrow(test_set), 10)
# Filter for points with non-zero error bars
non_zero_error_indices <- error_indices[test_errors[error_indices] != 0]
# Plot Actual vs. Predicted for the Test set with full-length error bars
plot(test_set$Target, test_preds, col = 'black', pch = 16,
    main = "RF on Milk Test Set",
    xlab = "Actual Concentration", ylab = "Predicted Concentration")
abline(a = 0, b = 1, col = "red", lwd = 2)  # Add 45-degree reference line
# Add error bars only for points with non-zero error lengths
arrows(
  x0 = test_set$Target[non_zero_error_indices],
```

```
  y0 = test_preds[non_zero_error_indices] - test_errors[non_zero_error_indices],
  x1 = test_set$Target[non_zero_error_indices],
  y1 = test_preds[non_zero_error_indices] + test_errors[non_zero_error_indices],
  code = 3, angle = 90, length = 0.1, col = "black"
)
# Add text for RMSEP and R-squared on the plot
text(x = min(test_set$Target), y = max(test_preds),
     labels = paste("RMSEP =", format(test_rmsep, digits = 5), "ppm\nR² =",
format(test_r2, digits = 5)),
     pos = 4, col = "black", cex = 0.8)
```

**SVM Code in R Used for Classification**

```
# Load Libraries
library(caret)
library(dplyr)
library(doSNOW)
library(parallel)
# Load Data
df <- read.csv("pcs_data_all_PA_Milk_fingerprint.csv", stringsAsFactors = FALSE)
df <- df %>%
  mutate(Group = factor(Group, levels = c("Below", "MRL", "Above")),
         across(starts_with("PC"), as.numeric))
# Split Data
set.seed(1234)
data_split <- createDataPartition(df$Group, times = 1, p = 0.7, list = FALSE)
train_set <- df[data_split, ]
test_set <- df[-data_split, ]
# Train SVM Model
train.control <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
radial_grid <- expand.grid(C = 10^(-2:3), sigma = c(0.001, 0.01, 0.1, 1, 5, 10))
```

```
cl <- makeCluster(detectCores() - 1, type = "SOCK")
registerDoSNOW(cl)
SVMmodel <- train(Group ~ .,
                  data = train_set[, c(paste0("PC", 1:4), "Group")],
                  method = "svmRadial",
                  tuneGrid = radial_grid,
                  trControl = train.control)
stopCluster(cl)
# Performance on Training Data
preds_train <- predict(SVMmodel, train_set[, paste0("PC", 1:4)])
confusionMatrix(preds_train, train_set$Group)
# Performance on Test Data
preds_test <- predict(SVMmodel, test_set[, paste0("PC", 1:4)])
confusionMatrix(preds_test, test_set$Group)
```

**SVR Code in R Used for Regression**

```
library(caret)
library(dplyr)
library(e1071)
library(doSNOW)
library(parallel)
# Load data
df <- read.csv("2pcs_data_milk_PA_quant_finger.csv", stringsAsFactors = FALSE)
# Ensure PC1, PC2, and Target are numeric
df <- df %>%
  mutate(
    Target = as.numeric(Level),
    PC1 = as.numeric(PC1),
    PC2 = as.numeric(PC2)
  )
# Handle missing values by removing them
```

```
df <- na.omit(df)
# Split Data
set.seed(1234)
data_split <- createDataPartition(df$Target, times = 1, p = 0.7, list = FALSE)
train_set <- df[data_split, ]
test_set <- df[-data_split, ]
# Set up caret to perform 10-fold cross-validation repeated 3 times
train.control <- trainControl(method = "repeatedcv",
                              number = 10,
                              repeats = 3,
                              search = "grid")
# Define the grid of hyperparameters for SVM Radial
radial_grid <- expand.grid(C = 10^(-2:2), sigma = c(0.01, 0.1, 1))
# Use the doSNOW package to enable caret to train in parallel
cl <- makeCluster(detectCores() - 1, type = "SOCK")  # Use all but one core
registerDoSNOW(cl)
# Train the SVM Radial model using caret for regression
SVMmodel <- train(Target ~ .,
                  data = train_set,
                  method = "svmRadial",
                  tuneGrid = radial_grid,
                  trControl = train.control)
# Stop the cluster
stopCluster(cl)
# Make predictions on the test set
test_preds <- predict(SVMmodel, test_set)
# Calculate residuals as error bars (difference between actual and predicted)
residuals <- test_set$Target - test_preds
# Compute RMSEP and R² for the test set
test_rmsep <- sqrt(mean(residuals^2))
test_r2 <- cor(test_preds, test_set$Target)^2
```

```
# Select indices for points to display error bars
set.seed(123)
test_error_indices <- sample(1:nrow(test_set), 10)
# Plot Actual vs. Predicted for the Test set with error bars
plot(test_set$Target, test_preds, col = 'black', pch = 16,
    main = "SVR on Milk Test Set",
    xlab = "Actual Concentration", ylab = "Predicted Concentration")
# Add error bars for selected points
arrows(test_set$Target[test_error_indices],
       test_preds[test_error_indices] - residuals[test_error_indices],
       test_set$Target[test_error_indices],
       test_preds[test_error_indices] + residuals[test_error_indices],
       angle = 90, code = 3, length = 0.1, col = "black")
# Add 45-degree reference line
abline(a = 0, b = 1, col = "red", lwd = 2)
# Display RMSEP and R² on the plot
text(x = min(test_set$Target), y = max(test_preds),
    labels = paste("RMSEP =", format(test_rmsep, digits = 5), "\nR² =", format(test_r2,
digits = 5)),
    pos = 4, col = "black", cex = 0.8)
```

## Appendix II: Additional Photos Taken During the Research

Back to Nature Organic Farm

**Appendix III: Pesticide Information**



| **Pesticide** | DuoDip | Ranger |
|---|---|---|
| **Target (Animal/Crop)** | Animals | Crops |
| **Target (Pests/Disease)** | Ticks, fleas, mange, tsetse flies, lice | Aphids, Berry Borer, Termites, Thrips, Whiteflies |
| **Active ingredients** | • 50% Chlorpyrifos<br>• 5% Cypermethrin | • 48% Chlorpyrifos |
| **Application** | • Hand Spraying<br>• Dipping | • Spraying |