MULTIVARIATE CHEMOMETRICS USING R IN FORENSIC CLASSIFICATION OF CERTAIN ANIMAL HAIRS.

MUHAMMAD ZULKHAIRIE BIN KAMISAN

UNIVERSITI SAINS MALAYSIA

2025

MULTIVARIATE CHEMOMETRICS USING R IN FORENSIC CLASSIFICATION OF CERTAIN ANIMAL HAIRS.

by

MUHAMMAD ZULKHAIRIE BIN KAMISAN

Thesis submitted in fulfilment of the requirements
for the Bachelors
Degree of Science (Honours) (Forensic Science)

FEBRUARY 2025

CERTIFICATE

This is to certify that the dissertation entitled Multivariate Chemometrics Using R in Forensic Classification of Certain Animal Hairs is the bona fide record of research work done by Muhammad Zulkhairie bin Kamisan during the period from November 2024 to January 2025 under my supervision. I have read this dissertation, and that in my opinion, it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation to be submitted in partial fulfilment for the degree of Bachelor of Science (Forensic Science).

Supervisor,	Co-Supervisor,
…… ……..	……..
(Dr Dzulkiflee Ismail)	(Dr Wan Nur Syuhaila Mat Desa)
Date: 18th February 2025	Date:

DECLARATION

I hereby declare that this dissertation is the result of my own investigations, except where otherwise stated and duly acknowledged. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at Universiti Sains Malaysia or other institutions. I grant Universiti Sains Malaysia the right to use the dissertation for teaching, research, and promotional purposes.

…… ……

(Muhammad Zulkhairie bin Kamisan)

Date: 18th Februari 2025

ACKNOWLEDGEMENTS

I am grateful to have Dr Dzulkiflee Ismail as my supervisor and Dr Wan Nur Syuhaila as my co-supervisor. They are the two people that exposed R to me, and the one that I look upon. I am also grateful to have the support of my family members throughout the duration of my study in USM Kubang Kerian, and I am proud to say that I have the best parents on this very Earth. Things happen sometimes, but I do believe in miracles, and I would say the completion of this research project is one of them. Alhamdulillah, thank you Allah, for giving me mercy this time!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

HCA	Hierarchical Cluster Analysis
LDA	Linear Discriminant Analysis
PCA	Principal Component Analysis
t-SNE	t-Distributed Stochastic Neighbor Embedding
ATR-FTIR	Attenuated Total Reflectance-Fourier Transform Infrared

KEMOMETRIK PELBAGAI MENGGUNAKAN R DALAM PENGELASAN FORENSIK BULU HAIWAN TERTENTU.

ABSTRAK

Bulu haiwan adalah salah satu bukti surih yang boleh ditemui di tempat kejadian, walaupun disebabkan saiznya yang kecil, ia boleh diabaikan oleh penyiasat forensik. Metodologi konvensional untuk analisis sampel rambut haiwan biasanya lebih tertumpu pada analisis DNA dan mikroskopi rambut haiwan, yang memerlukan kepakaran penyiasat forensik dan mengambil masa yang agak lama untuk disiapkan. Kajian ini menggunakan teknik analisis pantas dan tanpa musnah, iaitu spektroskopi Attenuated Total Reflectance Fourier Transform-Infrared (ATR-FTIR) disertai dengan kemometrik pelbagai menggunakan R dalam RStudio untuk pengelasan sampel rambut haiwan bagi tujuh haiwan berbeza. Kemometrik pelbagai yang digunakan dalam kajian ini ialah Analisis Kluster Hierarki (HCA), Analisis Komponen Utama (PCA), Pembenaman Jiran Stokastik Taburan-t (t-SNE), dan Analisis Diskriminasi Linear (LDA), yang digunakan untuk menjana tafsiran visual keputusan. Pemplotan 3D telah dilaksanakan pada PCA, t-SNE dan LDA untuk pemisahan kelompok yang lebih baik, yang menggambarkan perbezaan ketara kelompok untuk sampel rambut babi daripada jenis haiwan lain. Kerja semasa menunjukkan kaedah yang pantas dan tidak merosakkan untuk klasifikasi sampel rambut haiwan dalam penyiasatan forensik dengan pelaksanaan kemometrik menggunakan perisian statistik RStudio sumber terbuka.

MULTIVARIATE CHEMOMETRICS USING R IN FORENSIC CLASSIFICATION OF CERTAIN ANIMAL HAIRS.

ABSTRACT

Animal hairs are one of the trace evidence that could be encountered at a crime scene, though due to their minute size, they could be overlooked by forensic investigators. Conventional methodologies for analysis of animal hair samples usually are more focused on DNA and microscopy analysis of the animal hairs, which would require expertise of the forensic investigators and consume a good fraction of time to be completed. This study utilizes rapid and non-destructive analytical technique, namely Attenuated Total Reflectance Fourier Transform-Infrared (ATR-FTIR) spectroscopy accompanied by multivariate chemometrics using R in RStudio for classification of animal hair samples of seven different animals. The multivariate chemometrics used in this study are Hierarchical Cluster Analysis (HCA), Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Linear Discriminant Analysis (LDA), which were employed to generate a visual interpretation of the results. 3D plotting was implemented on PCA, t-SNE and LDA for better separation of the clusters, which illustrates a significant difference of the clusters for pig hair samples from the other type of animals. The current work demonstrated a rapid and non-destructive method for classification of animal hair samples in forensic investigation with the implementation of chemometrics using an open-source RStudio statistical software.

CHAPTER 1 INTRODUCTION

1.1 Background of Study

Referring to Tridico in 2005, the earliest known publication that describes the significance and evidentiary value of animal hairs in the medico-legal study date back to the 19th century, in which Alfred Taylor in 1894 documented a case that successfully proved the innocence of the suspect by concluding that the hair on the alleged murder weapon was from animal origin. From the early 1900s onwards, microscopic examination of hair was well established and by the late 1930, hair structure studies were conducted in detail by the Wool Industries Research Association (WIRA). Throughout the 19th century, the study of hair samples in general was more focused on the microscopic methods, with notable descriptive studies by Mathiak (1938), Mayer (1952), and Moore (1978), was transfer of interest occurs with the technological advancement in DNA analysis that allows individualization of animal hair that was not possible via microscopy in the earlier days.

In recent years, different approaches in the descriptive study about animal hair have been done through chemical characterization means, which was possible with the emergence of spectroscopic techniques. However, such studies are often limited to a specified sample of population and animal species, with regard to the lack of reference library to be used in assisting and confirming those analysis. Furthermore, spectroscopic analysis of animal hair samples would often require the assignment of each band in the spectra, and due to the natural composition of hair being keratin, those spectra of different animal samples would display similar characteristics or peaks. This is due to the high degree of variables and dimensionality of the spectral data, which almost certainly results in misclassification and failure to generalize the animal hair samples. Therefore, the implementation of chemometrics to reduce the dimensionality of those spectral datasets would aid in establishing patterns and determination of the animal type and hence, this summarizes the background of this study.

1.2 Problem Statement

The study of hair is often recognized due to its importance as trace evidence in criminal investigation, which dates to the 1800s and is typically associated with the case of the murder of the Duchesse de Praeslin in Paris in 1847 (Bertino, 2012). However, the use of hair to solve crimes is often limited to the use of human hair as a preferrable samples, with most perspectives would narrate several methods of analysis for hair evidence, such as side-by-side analysis using comparison microscope. Recent advancements in hair analysis continue throughout the 20th century, with the introduction of more sophisticated instruments and analytical methods that enable a wide range of analysis to be performed onto hair evidence and samples. Bertino and Bertino in 2012, however, emphasized that even though the hairs from a crime scene exhibit almost identical characteristics, those evidence might not be from the same source.

Referring to Jose et al. (2024), the application of vibrational spectroscopy in forensic discrimination of certain animals (mongoose, cattle and human hairs) has gained popularity over the recent years, noting that the infrared spectra of those animal hair samples would display the vibrational characteristics of a sample based on the different absorption frequencies of each functional groups in the samples. Several similar studies were conducted using similar methods of FTIR (Fourier Transform Infrared) which focus on discrimination and identification of the source correspondence of the hair samples (Espinoza et al., 2008, Bhatia et al., 2024, Jose et al., 2024, and Manheim, 2015). Though, most of those studies relied on chemometric techniques to distinguish the hair samples, as there is little visual difference in the spectra, mainly due to all hair is made of keratin proteins. In addition, studies that focus on non-destructive approach for analysis of animal hairs are hard to come by, with note that a study by Mohamad Zharif et al. in 2021, uses human hairs, pig bristles and avian feathers to determine source correspondence of L-cysteine with the utilization of chemometrics. Though several similar studies were evident in manifesting the applicability of chemometrics in addition to visual inspection of analytical data

(spectra and chromatograms), most of the applied chemometrics uses proprietary software which proved to be burdensome and requires additional resources to be replicated.

Therefore, this study aims to explore the variations in the IR spectra of certain animal hairs, which would notably decide the suitability of such data to be used in forensic discrimination of animal hairs. Furthermore, the applicability of RStudio as an open-source data and statistical analytics tool is to be assessed in conjunction with the ATR-FTIR dataset in this study.

1.3 Objectives

1.3.1 General Objectives

This research aims to study the significance and evidentiary value of certain animal hairs in forensic investigation using ATR-FTIR spectroscopy and chemometrics techniques.

1.3.2 Specific Objectives

To classify certain animal hairs using ATR-FTIR spectroscopy and chemometrics techniques.
To demonstrate the use of R in RStudio IDE for chemometrics techniques such as HCA, PCA, t-SNE and LDA on FTIR spectral dataset.

1.3.3 Significance of Study

The study of animal hairs as trace evidence is often treated as a branch is the scope of the study of hairs in general, by which these types of evidence are often analyzed using methods of microscopy and DNA analysis. However, reliance on such methods would prove to be a disadvantage, mainly due to the nature of such methods which typically consume a lot of time, and particularly for microscopic evaluation, it requires extensive expertise of the forensic examiner. Currently, the chemical characteristics of animal hair samples have gained popularity in recent years, with the implementation of chemometrics techniques to assist in the

discrimination of animal hairs, however, noting that most of publications related to this approach would prefer destructive methods on the samples.

Therefore, this study aims to explore the chemical characteristics of different animal hair samples which include the use of ATR-FTIR spectroscopy to investigate the variations in the spectra of the animal hair samples. This study also incorporates the method of chemometrics; the application of mathematical and statistical methods, to assist in the data processing of the animal hair spectra. In the light of this, this research aims to provide a new insight into the potential of open-source software, RStudio, to be used as a free alternative for statistical software. In addition, specific uses of hair samples as evidence are also related to wildlife forensic, by which it can be useful in animal identification, and therefore, methodological approaches in classification of animal hair samples using FTIR and chemometrics are to be discussed in this study.

CHAPTER 2 LITERATURE REVIEW

2.1 Animal Hair

Referring to Ruben and Jones in 2000, hair, fur and feathers are the most obvious anatomical differences between mammals and birds. Despite the significant differences between the two, they also share a key characteristic that sets them apart from all other vertebrates: true endothermic homeothermy at rest, or "warm-bloodedness." This distinct metabolic trait has sometimes resulted in misperceived claims that birds and mammals are closely related in evolutionary terms. Though, selection for an enhanced thermoregulatory capacity is often assumed to have resulted in the simultaneous evolution of elevated metabolic rates and insulator fur or hair. With the evidence of both fossil and physiological in recent discoveries, the enhanced thermoregulatory capacity is unlikely to be associated with the elevation of metabolic rates in mammals or birds. Ruben and Jones noted that the maintenance of similar temperature-corrected metabolic rates in most extant mammals is highly similar in a wider scope of other endothermal processes and anatomical structure such as hairs and sweat glands.

In addition, the authors suggest that a complete insulator covering of hair or fur might not exist until the appearance of the earliest mammals, by which the evidence from fossils showed that Harderian glands, structures for grooming and maintenance of insulator pelage in extant mammals were absent in therapsids (an advanced synapsid vertebrates and common ancestor of mammals). However, such generalizations are deemed speculative, and instead the authors suggested that insulative fur covering was necessary only when taxa (characteristic-based grouping of animal) in the therapsid-mammal lineage becomes extremely small. Hairs or furs for insulative property in mammals might evolved with respect of endothermic homeothermy, but only after the evolution of mammal-like metabolic rates in therapsids. Hence, in accordance to the authors, the initial appearance of an insulator pelage occurred after the “arrival” or advent of Mammalia, and it is likely that the variations in the development of hairs

and furs for insulation and thermoregulation are due to the decrease in size and nocturnal habits of mammals.

Concisely, hair is a characteristic of mammals, with the main purpose of thermoregulation function and others such as camouflage, sensory function, signaling or communication and sometimes for defensive purposes. Hair can change color over time in both animals and humans. For instance, babies typically have soft, uncolored vellus hair, which gradually transforms into thicker, colored hair as they grow, such as in the beards of adult males. As people age, hair loses its pigment and eventually turns white. Hair originates from the ectoderm of the skin and is an accessory structure of the integument, alongside sebaceous glands, sweat glands, and nails (Orasan etal., 2016).

The authors also mentioned that the term "hair" refers to two distinct parts: the hair follicle beneath the skin and the visible hair shaft that extends above the surface. The shaft consists of three layers: the outer cuticle, which is made up of overlapping, flat cells; the cortex, which contains keratin bundles in rod-like structures; and the medulla, a disorganized area in the center. The bulb in the dermis houses stem cells that not only regenerate hair after it falls out but also assist in skin repair following injury. The pigment in the hair shaft is produced by the hair follicle pigmentary unit, which involves interactions between melanocytes, keratinocytes, and dermal papilla fibroblasts. The dermal papilla is crucial for hair growth, formation, and cycling, while its blood vessels supply essential nutrients and oxygen to hair follicles and epidermal cells.

As described by Robertson in 1999, in animal hairs, different types of hair can be present in fur or pelage. Visual classification for hairs of different animals is typically based on their degree of coarseness, with the most prominent difference seen in guard- and under-hairs. The main difference between guard-hairs and under-hairs is that the former is generally longer and coarser than the latter, with guard-hairs often having the widest range of microscopic features, and therefore useful for identification. The author also mentioned some examples of common animal hairs to be of wool-type from different breeds of sheep and cat and dog hairs.

2.2 Animal Hair as Evidence in Forensic Science

Hair evidence has been in use to solve many types of criminal cases such as murder, sexual assault, burglaries, abuse cases, arson and terrorist incidents (Bailey, 2016). In 2016, Bailey mentioned that animal hairs as evidence have their own significance in solving human crimes, with the example of domestic dog hair that had been transferred from the dog's owner to a victim of an assault. Animal hair analysis in criminal cases, however, mainly focuses on the species identification of the animal, though identification of the individual animal has been done. This in turn is highly associated with animal-related crimes, including wildlife crimes such as poaching, hit and runs, import/export of endangered animals, and intentional poisoning. Though, animal hairs are often used as corroborative or supporting evidence, with the purpose of justifying the conclusions further. The author also noted the systematic searching of objects and animals for hair evidence that should be done in areas where hair transfer is most likely to occur, and by simple means, hair found in a bite wound of a dog suspected to involve in dog fighting is more evidentially valuable than if just found on the collar of the dog. This justification of evidentiary value of said animal hair could be strengthened with additional information such as the knowledge of animal access since the crime.

In accordance with the evidentiary value of animal hair, Boehme et al. in 2009 proposed a study of persistence of animal hairs in the context of forensic science. In their study, the persistence of cat and dog hair on different garments and fabrics were investigated, and the researchers found three factors that are associated with the differences in persistence namely: 1) fibre type, 2) yarn or knit tightness, and 3) the presence or absence of surface fibres on the recipient fabric. However, this study is case-specific, by which the researchers only focus on a certain probable scenario where animal hair might be found. Robertson in 1999 described the occurrence of hair as physical evidence by adhering to Exchange Principle by Edmond Locard, such that the contact between one surface and another may transfer some of the hair onto the surface. The latter is it by means of direct or indirect transfer. Direct transfer, which would

always be primary transfer, occurs when hairs directly from original source was transferred to another surface. Indirect transfer, on the other hand, can be either primary or secondary, that could involve one or more intermediaries such as clothing and bedding.

Referring to Bailey in 2016, the principle of *GIFT (Get It First Time)* can be used by anyone attempting to retrieve questioned hairs, which emphasis that the retrieval method must allow all target hairs to be recovered without any loss or contamination. The method of recovery would depend on factors such as surface type, surface area, presence of debris and whether the hair was embedded on surface or not. Bailey suggested a few retrieval methods such as tape lifting, tweezering, vacuuming, shaking, scrapping, combing and filtering. It is important to note that tape lifting and combing are used as the retrieval method of this research due to the ability to capture multiple hair strands in quick succession of time. Similar approaches were proposed by Robertson in 1999 whereby he described that the recovery of hair evidence can be done in two locations: at the crime scene or in the laboratory. Recovery at the crime scene should be done with extra caution, by which each exhibit items that are suspected to attain the hairs shall be packaged immediately in its own separate container. In addition, to avoid accidental contamination, exhibits in their unpackaged state shall be handled by only one person or examiner. Robertson suggested the simplest method of recovering hair s in the laboratory by searching the items visually under oblique lighting at various angles and using tweezers to pick of any readily visible hairs. The method of lifting by tape was also mentioned, with a major advantage of being able to retrieve many hairs which are not readily visible.

2.3 Forensic Analysis of Animal Hair

In forensic science, the analysis of animal hair is often done by the same procedure and approaches for human hair. Bertino in 2012, gave a brief description of hair analysis as to be done by both macroscopically and microscopically. For macroscopic characteristics, hair evidence was inspected for their lengths, colours, and curliness. In the microscopic analysis of hairs, characteristics such as the pattern of the medulla, pigmentation of the cortex, and types of scales on the cuticle were often considered. Both macroscopic and microscopic examination of hairs are often done in non-destructive manner, such that morphological and physical characteristics of the samples are the most significant, and by retaining the original state of the hairs, an accurate description of the sample can be completed. According to Ahmed et al. in 2018, by which the researchers used both methods of macroscopy and microscopy, they were able to identify and compare the hair cuticle scales and hair medulla and pigmentation in domestic animals namely large ruminants, equine, small ruminants and canines. The researchers concluded on several prominent variations in terms of the characteristics mentioned previously, though some animals show similarities in the characteristics. The similarities, however, would likely result in false perception and generalization on animal hairs in, whereby referring to a study by Tridico et al. in 2014, the authors mentioned some misconception regarding morphological identification of animal hairs. The misconceptions are as follows: 1) cat and dog hairs can be reliably identified solely on root shapes, 2) pig hairs may be mistaken for human hairs, and 3) polar bear hairs are hollow, and inevitably, those misconceptions should motivate the development of another reliable method for forensic analysis of animal hairs.

Analytical approaches with the application of chemometrics were also able to demonstrate meaningful interpretation for different purposes such as characterization of amino acid sources in hair samples. Referring to Mohamad Zharif et al., in 2021, the researchers were able to differentiate five distinct clusters in PCA of FTIR data, by which they refer to as a fast and environmentally friendly approach to distinguish the primary sources of L-cysteine. The

other purpose of analytical methods that use hair samples is to determine effects of photodegradation and thermal degradation, as demonstrated by Carr and Lewist in 2008. In the study, the authors explained that FTIR analysis of wool samples after thermal and photodegradation were able to point to some differences in certain amino acids after degradation, by looking into the intensity of the amino acids' absorbance at their specific wavelength.

In recent years, the development of analytical methods for the discrimination of animal hairs has been a popular topic among researchers, and light has been shed leading to a new method of analysis. Some of the most notable and related research regarding this issue is the study by Jose et al. in 2024, by which they showed that FTIR analysis assisted by chemometrics are feasible. In their study, Jose and other corresponding authors proved that the FTIR analysis of the samples, combined with chemometrics, namely PCA and PLS-DA, was a useful tool in the discrimination of animal hairs and synthetic fibres in the sample. From Bala, Sharma and Sharma in 2024, differentiation of various animal families based on machine learning (ML) classifier and spectral data of ATR-FTIR were found to be a remarkable improvement of the chemometrics method in the discrimination of animal families. Using hair samples of certain animals from different families, Bala and his peers mentioned that random forest (RF) classifier as the most effective for family discrimination model, and helpful when the visual examination of the spectra was considered non-applicable. However, the approaches in the studies mentioned previously were to be considered for different scenarios and circumstances, due to some limitations mentioned by the authors, such as from Jose et al. in 2024, where they suggest a higher volume of samples would be needed for validation of the results.

2.4 ATR-FTIR Spectroscopy

Infrared (IR) spectroscopy is a technique based on the vibrations of the atoms of a molecule, and commonly obtained by passing infrared radiation through a sample and

determining what fraction of the incident radiation is absorbed at a particular energy. The frequency of vibration of a part of a sample molecule is then associated with this corresponding energy at which any peak in an absorption spectrum appears (Stuart, 2004). According to Stuart in 2004, understanding the molecular symmetry and group theory is important when assigning infrared bands of the sample. The interactions of infrared radiation with matter can be simplified in terms of changes in molecular dipoles associated with vibrations and rotations. The author described the properties of a molecule to be able to absorb radiation when the incoming IR radiation is of the same frequency as one of the fundamental modes of vibration of the molecules, thus these changes in radiation are detected by the instrument.

Referring to Singh, Pradhan and Materny in 2021, infrared (IR) spectroscopy is a popular fundamental spectroscopic tool that involves the analytical processing of atomic and molecular vibrations of the molecule upon resonantly associating with an incident IR radiation source. The main purpose of IR spectroscopy is to selectively probe the molecular functional groups through their unique IR absorption bands, by which these absorption bands constitutionally dependent on the molecular composition, conformation, and the condition of the surrounding medium. Additionally, major advancements and revamps of IR spectroscopy have led to the evolution of sensitivity and efficiency of probing IR spectroscopic techniques, which includes the implementation of Fourier transform infrared (FT-IR) spectrometer in 1980 (Singh et al., 2021). This application of Fourier transform has led to an increasing spectrometer performance which enables higher signal-to-noise (S/N) ratio, along with the introduction of diamond as the ATR material for IR spectroscopy (Milosevic, 2012).

According to Lindon et al. in 2010, FT-IR spectrometers have almost entirely replaced all other dispersive instruments, with note that the most significant feature of Fourier transform (FT) spectrometers is that radiation from all wavelengths is measured simultaneously. Being more efficient than the dispersive instruments, by which most spectroscopists refer to as the Fellgett advantage, faster and more sensitive measurement could be achieved using FT-IR.

With regards to the dispersive instruments, these IR instruments typically employ prisms made of materials such as sodium chloride (NaCl). The schematic in **Figure 2.4.1** depicts the optical path of an infrared spectrometer which uses a grafting monochromator. According to Stuart in 2004, dispersion occurs when the energy falling to the entrance slit is collimated (radiation or ray of lights made accurately parallel) onto the dispersive element and the dispersed radiation is then reflected to the exit slit to be detected by the detector.

However, Stuart noted a few limitations of this dispersive mode of IR instrument, that includes the detector must have an adequate sensitivity to the incident radiation from the sample over the entire spectral region, and the radiation source must be sufficiently intense over the wavenumber and transmittance range. Given that dispersive spectrometer is generally constricted with the use of monochromator, the narrow entrance and exit slits would limit the wavenumber range of the radiation reaching the detector to one resolution width (Stuart, 2004).

The emergence of ATR-FTIR spectroscopy is often associated with the understanding and application of reflectance methods in the IR spectroscopy field. Stuart in 2004 described reflectance techniques that are best used for samples that are difficult to analyze by the conventional transmittance methods, and this method is subdivided into two main categories, which are internal and external reflectance. With respect to internal reflectance, attenuated total reflectance (ATR) spectroscopy is a method that uses the phenomenon of total internal reflection, by which a beam of radiation would undergo total internal reflection upon entering the crystal, given that the angle of incidence at the interface of the sample and crystal is greater than the critical angle. When a material that selectively absorbs radiation is within proximity of the reflecting surface, the beam loses energy at the wavelength where the material is absorbed. The resultant attenuated (reduced in value) radiation is measured and plotted as a function of wavelength by the spectrometer and hence gives rise to the absorption spectral characteristics of the sample. **Figure 2.4.2** depicts the schematic of a typical attenuated total reflectance cell found in ATR-FTIR spectrometer.

The electromagnetic spectrum of the infrared region is situated at 12,500 to 500 cm^{-1}. Nevertheless, mid-IR (region 4000 to 400 cm^{-1}) and near-IR (region 12,500 to 4000 cm^{-1}) are the regions most analysed in FTIR spectroscopy. During the sample analysis in the mid-IR region, the molecules will absorb mid-IR energy, enabling the fundamental vibration of specific functional groups to be observed. The mid-IR fingerprint region (1800 to 600 cm^{-1}) also displays well-defined spectra for detecting lipids, polysaccharides, proteins, and carotenoid molecules (Mendes & Duarte, 2021). **Table 2.4** summarizes the associated bond stretching with their respective frequency ranges.

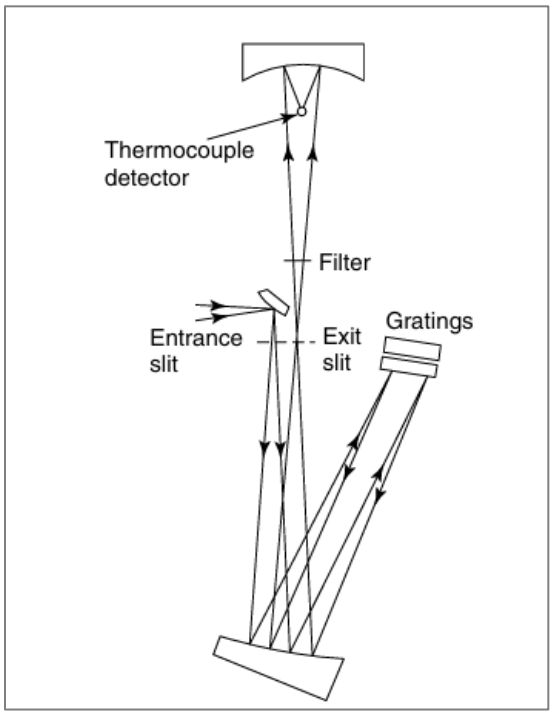

Figure 2.4.1 Schematic of the optical path of a double-beam infrared spectrometer with a grafting monochromator (Source: Stuart, 2004)

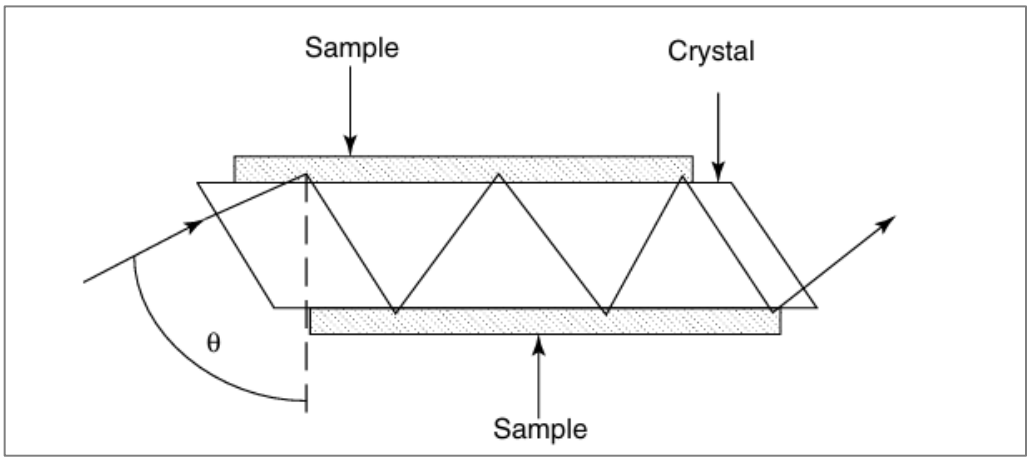

Figure 2.4.2 Schematic of a typical attenuated total reflectance cell (Source: Stuart, 2004)

Table 2.4 Summary of bond association with their respective frequency ranges (Source: Mendes & Duarte, 2021)

Bond	**Type of Compound**	**Frequency Range (cm^{-1})**
C — H	Alkanes	2850 – 5970 1340 – 1470
C — H	Alkenes	3010 – 3095 675 – 995
C — H	Alkynes	3300
C — H	Aromatic rings	3010 – 3100 690 – 900
O — H	Monomeric alcohols; Phenols	3590 – 3650 3200 – 3600 3500 – 3650 2500 – 2700
N — H	Amines, Amides	3300 – 3500

C ═ C	Alkenes	1610 – 1680
C ═ C	Aromatic rings	1500 – 1600
C ≡ C	Alkynes	2100 – 2260
C — N	Amines, Amides	1180 – 1360
C ≡ N	Nitriles	2210 – 2280
C — O	Alcohols, Ethers, Carboxylic acids, Esters	1050 – 1500
C ═ O	Aldehydes, Ketones, Carboxylic acids, Esters	1690 – 1760
NO_2	Nitro compounds	1500 – 1570 1300 – 1370

2.5 Chemometrics

According to Chu et al. in 2022, chemometrics was born in the early 1970s, and is generally defined as a branch of chemistry, which utilizes mathematical and statistical methods with computer technology, designs and selects the best measurement procedure and experimental methods to obtain the maximum information by interpreting chemical data. While the main goal of chemometrics is to extract the most useful information from the measured data, the most unique feature of chemometrics is to construct the chemical measurement as a mathematical model that can be expressed and processed through mathematical formulas. The prominent difference of chemometrics with other branches of theoretical mathematic, according to the authors, is that chemometrics is a discipline of all theories and methods based on chemical experimental data. The application of chemometrics has been integrated into the current interdisciplinary analytical techniques, by which it has been useful method for spectrum discrimination and simultaneous identification of multiple components in a complex system.

The concept of principal component analysis (PCA) was first proposed by K. Pearson in 1901, and over the decades, specifically until 1972, PCA was majorly used to deconvolute overlapping peaks in chromatograms. In the 1971, a Swedish chemist named S. Wold proposed the study of main concepts namely: chemical data analysis, computer in chemistry, and chemometrics, whereby the progression in chemometrics started three years later with the establishment of the International Chemometrics Society (ICS). Though, thriving period or the "golden age" of chemometrics was in the 1980s, in which due to advancements and availability of more sophisticated computer technology enables the methods of unique multivariate calibration, discrimination, and chemical pattern recognition such as partial least squares and rank annihilation factor analysis. The authors summarized the contents of chemometrics as shown in **Figure 2.5.**

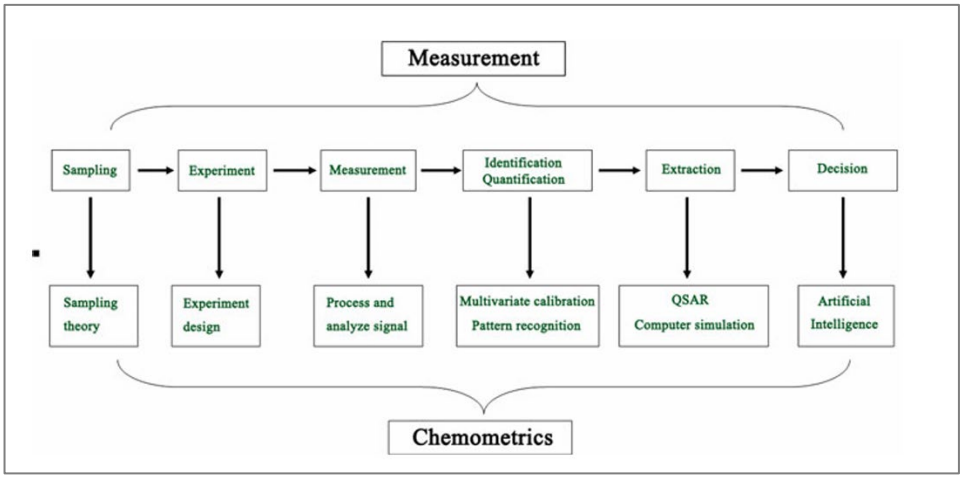

Figure 2.5 Summary of the contents in chemometrics (Chu et al., 2022)

Chemometrics has been recognized to be a corroborative necessity in quantitative and qualitative analyses of spectroscopy that serves the benefits such as: 1) in multivariate calibration to improve accuracy and precision of analysis, 2) in signal process technology to improve instrument's signal-to-ratio, and 3) in pattern recognition of chemical data.

The results of chemometrics, especially of multivariate datasets, are often represented in the form of graphics, by which according to Brereton in 2018, the ability to convert numbers into a simple diagram has allowed a large community to better understand the relationship between components in their datasets. Principal components for PCA, as well as the key elements of other chemometrics are often treated as much as objects in an imaginary multivariate space than mathematical entities. One of the simplest graphical representations is with score plots, usually in the form of the scores of one computed variable with the other (biplot). The author suggests that interpretation of score plots can be done visually, whereby in the scenario of which distinct clusters are observed, the clusters can be associated with the corresponding parameters. In addition, spatial representation of the samples is also possible, given that this addition of dimension would result in the creation of 3D plots, and it is highly plausible if the clustering in the biplot is convoluted. Spatial mapping, in general, would give a much more meaningful representation of the computed variables by separating the clusters into a three-dimensional space (Chu et al., 2022)

2.5.1 Hierarchical Cluster Analysis (HCA)

Cluster analysis refers to a group of algorithms in which the samples are grouped according to their relative similarity. According to Sauzier and others in 2021, Hierarchical Cluster Analysis (HCA) is a widely used method in which the samples are connected to form clusters based on separation distance. This approach is typically agglomerative, which started with a single object (sample) and progressively to obtain larger clusters. The formed clusters are dependent on the

user's selection on an appropriate measure of distance between objects, such that those clusters closest together are merged to form a new cluster, and repetition of this process occurs until a single cluster is obtained. The most common distance metrics and linkage criterion is Euclidean mode and single linkage respectively. In Euclidean distancing, an imaginary straight line to represent the metric distance is created between objects in *n*-dimensional space, and then the shortest distance between the objects of two adjacent clusters are linked, resulting in a 'single linkage' that forms long clusters. The hierarchical relationship between clusters is typically illustrated using a dendrogram, by which the samples are plotted according to their determined clustering pattern (Sauzier et al., 2021). Referring to Miller and others in 2018, the interpretation of a dendrogram could be done visually, such that the example of HCA dendrogram is given in **Figure 2.5.1**. According to the authors, by 'cutting the tree', or stop the grouping at the dotted line, the compounds A-L would fall into two distinct groups. This simple interpretation is made possible by the mode of distancing and linkage, though for other approaches, the interpretation may vary (Miller et al.,2018).

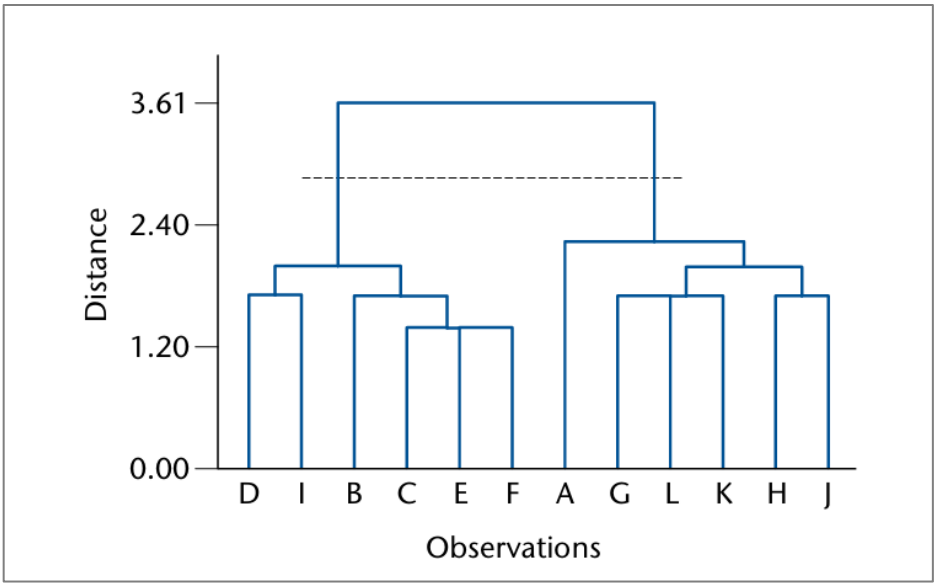

Figure 2.5.1 Example of dendrogram used in HCA (Miller et al., 2018).

2.5.2 Principal Component Analysis (PCA)

According to Greenacre et al. in 2023, principal component analysis is a multivariate statistical method for reducing a cases-by variables data table to its essential features which are called principal components. Regarding principal components, they are defined as a few linear combinations of the original variables that maximally explain the variance of all the variables in a specific sample dataset

One of the simplest graphical representations is with score plots, usually in the form of the scores of one principal components (PC) against other PCs. As a result, the resultant graphs can be interpreted visually, usually by associating clustered data points into their respective groups. For example, in resolving convoluted peaks in a chromatogram, the clustered data points of the first two PCs can display regions of the chromatogram where there are pure compounds in the sample. Other than that, the closer the points to the origin, the lower the intensity. The example of a PCA visualization was illustrated in **Figure 2.5.2** (Brereton, 2018).

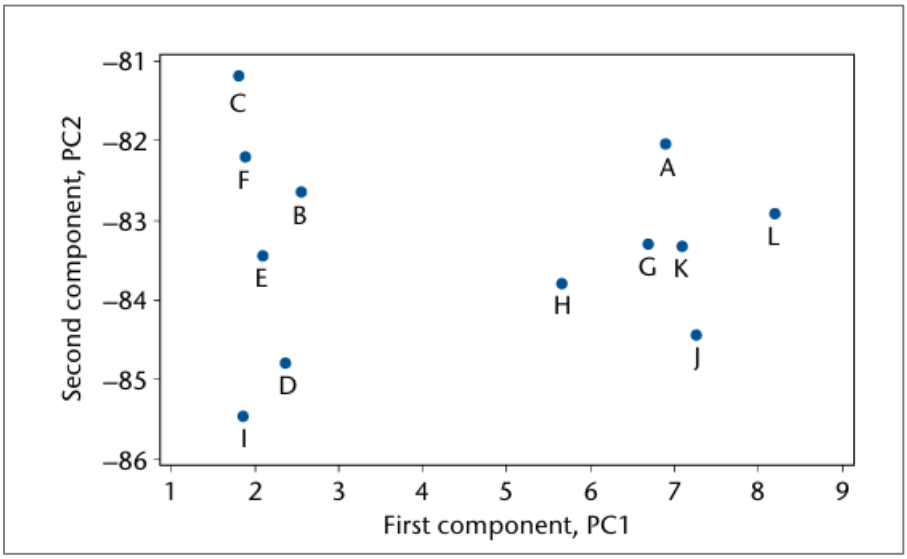

Figure 2.1.2 Example of PCA score plot (Brereton, 2018).

2.5.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)

Referring to Maaten and Hinton in 2008, the visualization of high-dimensional data is an important problem in many different domains and often requires researchers to deal with data of widely varying dimensionality. The authors suggested an improvised version of Stochastic Neighbor Embedding (SNE), which is t-Distributed Stochastic Neighbor Embedding (t-SNE), by which this optimized technique managed to solve the "crowding problem" that mainly occurs in normal SNE approaches. In summary, SNE is an unsupervised chemometrics that starts with the conversion of a high-dimensional Euclidean distance between datapoints into conditional probabilities that represent similarities. Utilization of SNE mainly aims to find a low-dimensional data representation that minimizes the mismatch between the conditional probability of the points (p) and the low-dimensional probability counterparts of the said datapoints (q). This technique, however, has a relatively large effect on the representation of the nearby datapoints, as the usage of a widely separated map points would lead to waste of some of the probability mass in the relevant Q distribution. The author described SNE as an inferior method that requires regular optimization, and hence they suggest an optimized t-SNE for better data visualization. The cost function used by t-SNE is generally different from normal SNE, by which to solve the "crowding problem", the author proposed t-SNE as it uses a symmetrized version of the SNE cost function with simpler gradients that were newly introduced during their time of research. T-SNE is also favorable due to the utilization of Student t-distribution rather than Gaussian in computing the similarity between two points in the *low-dimensional space* (Maaten & Hinton, 2008).

Fundamentally, t-SNE can be used for FTIR spectral dataset due to its prowess in dimensionality reduction and visualization. Chaber and his peers in 2018 employed the dimensionality reduction method using t-SNE in distinguishing Ewing sarcoma and osteomyelitis results in the clustering of the datapoints into two apparent subgroups (Chaber et al. 2018). However, the performance of t-SNE for spectral dataset is best represented by

Mohamad Asri et al. in 2022 with the discrimination and source correspondence of black inks using Raman spectroscopy. Although the approach of obtaining data is different, their implementation of t-SNE on the similarly structured dataset would suggest that t-SNE could be used on FTIR spectral dataset.

2.5.4 Linear Discriminant Analysis (LDA)

According to Gambella and others in 2021, linear discriminant analysis (LDA) is an approach for classification and dimensionality reduction and often applied to data that contains many features or variables, in which reducing the number of features is fundamentally necessary to obtain robust classification. The authors mentioned that LDA tends to be more robust than PCA since LDA takes the data labels into account in the process of computing the optimal projection matrix or the graphical representation of the reduced features. However, Sauzier and his peers in 2021 emphasize that users should be aware of accepting the results of LDA without further critique. This is because LDA is considered as a 'hard' classification method; in a way that this chemometric technique will forcefully assign the objects (samples) into a single class to which it demonstrates the closest similarity. Though, with the proper labelling prior to LDA, this 'brute' force classification might come in handy, by which the clustering might depict favorable representation of the samples. The difference between PCA and LDA can be displayed in **Figure 2.5.4**, which the implementation of 'force' classification might result in poor clustering of certain groups of samples in the plot.

It is important to note that LDA is classified as a supervised chemometric method with the main purpose of supervised method is to use the objects in the dataset to find a rule for allocations of any new object of unknown origin. However, due to the nature of this method, the new object might in some circumstances be classified into a specific group despite it coming

from another group of objects, which would likely be considered as a misclassification (Sauzier et al., 2021).

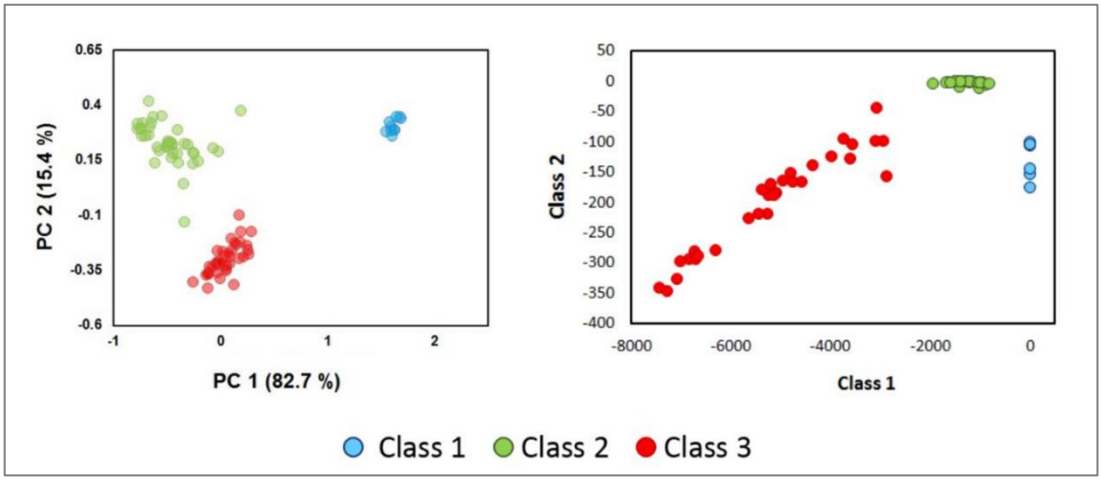

Figure 2.5.4 Difference between PCA (left) and LDA (right). (Gambella et al., 2021).

2.6 R within RStudio

The application of R for chemometrics techniques are mostly demonstrated and utilized for the purpose of dimensionality reduction of the spectral dataset in this study. Referring to Sardareh and Brown in 2021, R through RStudio is syntax driven, by which the use of analysis procedures in RStudio requires the users to write text-based syntax on command line in an identical fashion to working in native R. In RStudio software, the user can find by default the console in which the users can directly type in commands and view their results. Through the RStudio interface, R provides a comprehensive library of packages to support those analysis, ranging from basic to extremely complex procedures. However, the users must ensure that the required packages are installed before running the desired procedures and executing the commands or syntax.

In addition, RStudio offers a wide variety of graphs for data visualization, in which these excellent properties of visualization support became available through a wide range of modules,

especially "*ggplot2*" package, which supports detailed customization for graphs of various data type. Next, various data files can be imported into RStudio, and the users can view the imported data by clicking on the dataset name or spreadsheet icon in the "environment". In addition, importing the data files can also be done by users' syntax inputs, such as using "read.csv()" command.

However, some challenges that make the utilization of R and RStudio relatively difficult for novices are that, firstly, multiple separate packages were developed independently for the open-source environment, and those packages can often implement the same statistical procedures. For example, the packages "*stats*" and "*MASS*" includes useful functions for regression analysis, and the end-user must evaluate which package is better. Next, the use of "ggplot()" function requires the user to master its syntax, which can be time consuming and challenging for novices, and often requires multiple trial-and-error for determining the optimal syntax in the environment.

CHAPTER 3 METHODOLOGY

3.1 Materials and Methods

A few sheets of white A4 paper were purchased and selected to be the stage for the animal hair samples. A black duct tape and a masking tape were purchased to assist in fixing the hair sample onto the makeshift paper stage and onto ATR-FTIR instrument respectively. Liquid nitrogen was supplied by Gas Pantai Timur Sdn. Bhd. (Kota Bharu, Kelantan).

3.2 Sample Collection

The animal hair samples (cow, goat, sheep, horse, rabbit and pig) were obtained from farmhouses near Kelantan around October 2017. The hair samples of cats were collected by taping the body of cats owned by Dr Dzulkiflee bin Ismail. The hair samples were kept in sealed plastic bags inside a drawer of the office room and used in the FTIR spectroscopic analysis on 24 January 2025. Three individual animals from each type were chosen, and three strands of hair from the body region of the animal were chosen to be studied. The specific animal types used were tabulated in **Table 3.2.**

Table 3.2 Seven different animal types used in this study with labels.

Animal Type	**Label**
Cat	A1 – A3
Cow	B1 – B3
Goat	C1 – C3
Horse	D1 – D3
Rabbit	E1 – E3
Sheep	F1 – F3
Pig	X1 – X3