AUTHENTICATION OF EDIBLE OILS USING FOURIER TRANSFORM INFRARED SPECTROSCOPY AND PATTERN RECOGNITION METHODS

By

ISIO SOTA-UBA

Bachelor of Science in Chemistry
University of Ibadan
Ibadan, Oyo
2007

Master of Science in Analytical Chemistry
University of Ibadan
Ibadan, Oyo
2011

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2022

AUTHENTICATION OF EDIBLE OILS USING
FOURIER TRANSFORM INFRARED SPECTROSCOPY
AND PATTERN RECOGNITION METHODS

Dissertation Approved:

Dr. Barry K. Lavine
Dissertation Adviser

Dr. Ziad El-Rassi

Dr. Nicholas Materer

Dr. Richard Bunce

Dr. Albert Rosenberger

ACKNOWLEDGEMENTS

My sincere gratitude goes to my advisor Dr. Barry Lavine for his support and expert guidance throughout my PhD work. I would like to thank my advisory committee members for constructive criticisms and pointers to make me a better scientist.

Special thanks go to my current lab mates -Dr. Collin White for useful insights and to George, Tom, Haoran and Elizabeth for interesting discussions and making the lab fun. Also, I have had the pleasure to work with other lab mates who have moved on to bigger things including Dr. Nuwan Don Perera, Dr. Kaushalya Sharma Dahal, Dr. Francis Kwofie, Matthew Bamidele and Tanner Hagerman (undergraduate).

I want to thank my parents Macauley and Zeniatu Sota for helping me understand the value of acquiring knowledge from a very young age. My deep gratitude goes to my husband Franklin for supporting my decision to apply for a PhD and his tremendous encouragement during this work. To Nova, my daughter, thanks for helping to improve my multitasking skills. To my cheerleaders, my siblings, Efe, Vwairhe, Ochuko and Ese, thank you for always hoping for the best.

Name: ISIO SOTA-UBA

Date of Degree: JULY, 2022

Title of Study: AUTHENTICATION OF EDIBLE OILS USING FOURIER TRANSFORM INFRARED SPECTROSCOPY AND PATTERN RECOGNITION METHODS

Major Field: CHEMISTRY

Abstract: A potential method to determine whether two edible oils can be differentiated by infrared spectroscopy is proposed. IR spectra of the pure edible oils and mixtures of these edible oils in known amounts are compared using pattern recognition techniques to solve a ternary classification problem. The edible oil mixtures span a large concentration range. If the IR spectra of the two edible oils and their binary mixtures are differentiable then differences between the IR spectra of the two edible oils are of sufficient magnitude to ensure that a reliable classification of these two edible oils can be obtained by infrared spectroscopy. The mixtures were prepared gravimetrically or from digitally blended data of the pure edible oils using an edible oil spectral library. The feasibility of authenticating edible oils such as extra virgin olive oil was demonstrated using this approach. For these studies, both digital and experimental data were combined to generate training and validation data sets to assess detection limits for adulterants.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER I

Introduction

1.1. Adulteration of Edible Oils

Adulteration of food is a recurring problem. In recent years, there has been a rise in cases of food adulteration with its attendant economic and health consequences. One contributing factor is the steady increase in the cost of food products.[1-1] To keep the cost of food at a manageable level for the consumer, some manufacturers have replaced a component or ingredient in food with one that is less expensive to maintain the same price range and profit margin. Another factor that has contributed to food adulteration is inadequate legislation to assure fair trade of food and food products. A well-known example is the adulteration of plant-based edible oils by mixing a less expensive edible oil with a more expensive one.[1-2] Clearly, the adulteration of edible oils raises questions about food safety and quality. This dissertation describes the development of an analytical methodology to detect the presence of adulterants in plant-based edible oils.

Adulteration of plant-based edible oils can be incidental or deliberate. It is incidental when foreign substances are added to food as a result of ignorance, negligence, or improper maintenance of facilities. Several manufacturers may use the same production facility to manufacture other products. Improper cleaning and inspection of the facility

before other products are produced can lead to contamination. On the other hand, deliberate adulteration is the intentional addition of foreign substances or the substitution of other components for economic gain. In a report commissioned by the National Center for Food Fraud and Defense in collaboration with the Department of Homeland Security, so-called "food fraud" was defined as "a collective term that encompasses the deliberate substitution, addition, tampering or misrepresentation of food, food ingredients or false misleading statements made about a product for economic gain."[1-3] Adulteration of edible oils may present some public health implications or hazards for the consumer. An example of this is 'toxic oil syndrome' in Spain where consumers were adversely impacted by adulteration. [1-4 – 1-6] In this incident, poorly refined rapeseed oil mixed with aniline-based compounds marketed as olive oil were consumed by approximately 20,000 people and resulted in approximately 300 deaths. Similarly, in India, the mixing of an edible oil with argemone or mustard oil led to an epidemic in dropsy and glaucoma.[1-7 – 1-9] Adulteration of edible oils has become more sophisticated as less expensive edible oils with similar fatty acid and sterol profiles are mixed with more expensive edible oils which make it far more difficult for them to be detected by unsuspecting consumers.

1.2. Components of Edible Oils

Edible oils are food substances obtained from plants or animal sources. They are usually liquids at room temperature and consist mainly of triglycerides. Tropical oils such as palm oil, palm kernel oil and coconut oil may be solid at room temperature because they contain a high amount of short chain triglycerides and saturated fatty acids.[1-10, 1-11]

Triglycerides are compounds formed from the condensation reaction between glycerol and three molecules of fatty acids (see Figure 1.1). These fatty acids can be saturated (SFA), mono-unsaturated (MUFA), or poly-unsaturated (PUFA) see Figure 1.2. The chemical structure of a saturated fatty acid does not contain C=C bonds. MUFA contains only one C=C bond while polyunsaturated fatty acid contains more than one C=C bond in its structure. The human body can synthesize SFAs and MUFAs, but the simplest PUFA such

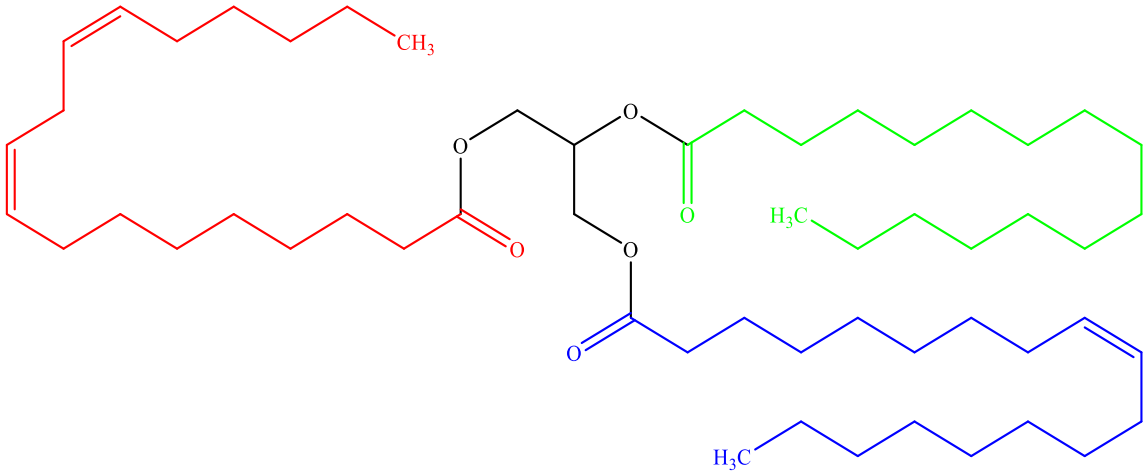

Figure 1.1. A triglyceride showing linoleic acid in the SN 1 position (red), palmitic acid (green) in the SN 2 position and oleic acid (blue) in the SN 3 position of the glycerol backbone.

as linoleic and α-linolenic acids can only be synthesized in plants.[1-12] Fatty acids that cannot be synthesized by the body are called 'essential' fatty acids because they must be provided by foods. Fatty acids are important as they offer both nutritional and health benefits. PUFA has been shown to reduce the risk of heart problems [1-13] while cis-MUFA may contribute to high density lipoprotein (HDL), which reduce the risk of atherosclerosis and cardiovascular diseases.[1-14, 1-15] Fatty acids are a source of energy for metabolic processes and are important for stabilizing biologic membranes by creating physical properties that are optimal for the transport of substances across the membrane.[1-16, 1-17]

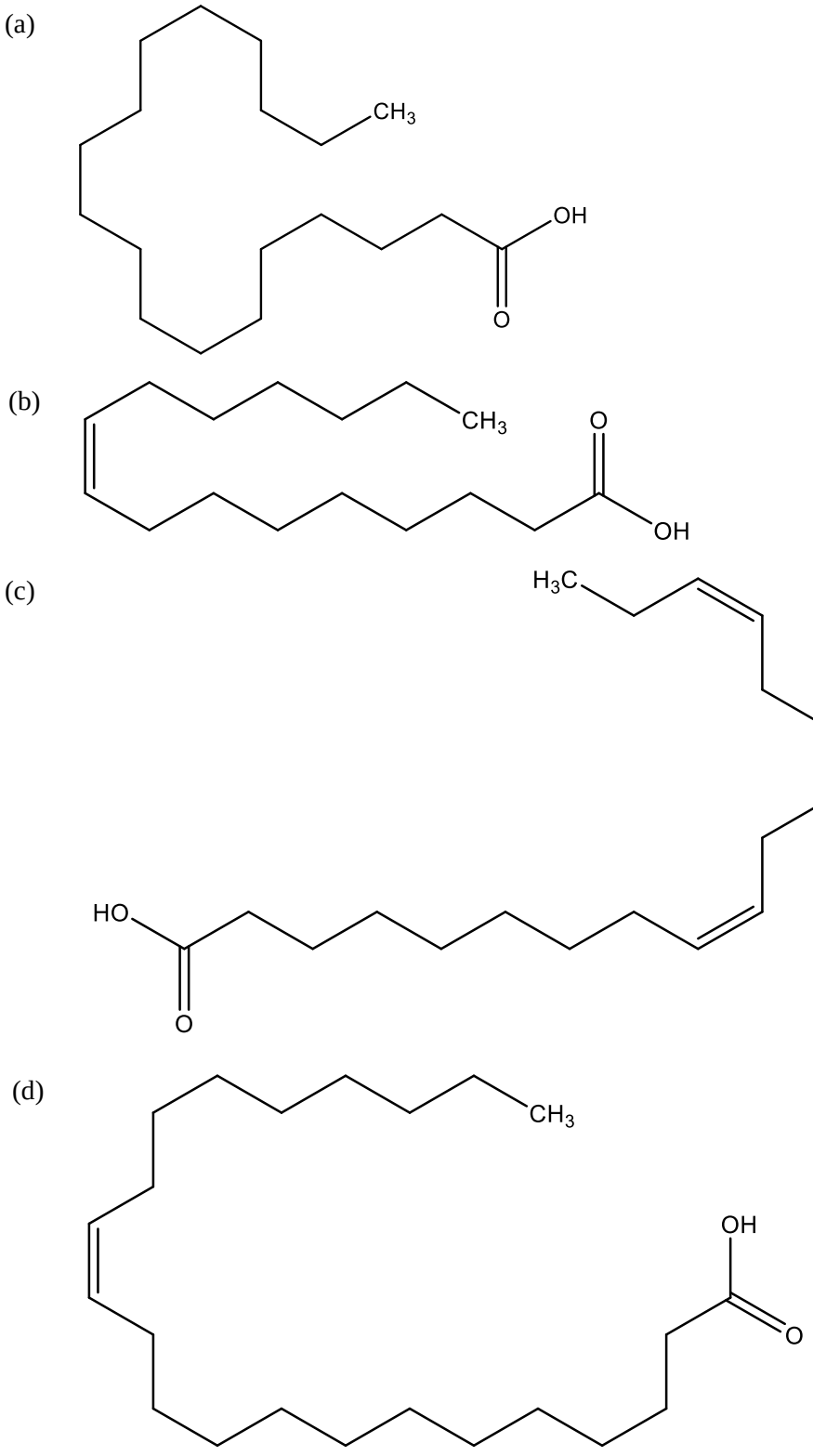

Figure 1.2. Structure of (a) stearic acid 18:0, a saturated fatty acid (b) palmitoleic acid 16:1(9), a mono-unsaturated fatty acid (c) α-linolenic acid 18:3(9,12,15), a polyunsaturated fatty acid and (d) erucic acid 22:1(13), another monounsaturated fatty acid.

Unsaturated fatty acids can take the cis- or trans- configuration. Trans fatty acids are geometric isomers of cis fatty acids. Trans fatty acids are considered unhealthy as they contribute to low density lipoprotein (LDL) and total cholesterol which can raise the risk of heart disease and stroke. Apart from classification based on the degree of unsaturation, fatty acids can also be classified based on chain length as either short chain, medium chain, or long chain fatty acids. A short-chain fatty acid is between 2 and 6 carbon atoms long. Medium chain fatty acids are 8 to 12 carbons long while long-chain fatty acids are 14–24 carbons long. Examples of long chain fatty acids are oleic acid 18:1(9), linoleic acid 18:2(9,12) and α-linolenic acid 18:3(9,12,15). The number(s) in parenthesis represent the position(s) of unsaturation in the chain.

Although, edible oils are composed primarily of triglycerides (approximately 96%), other components are also present in the oil. The composition of a particular edible oil can be influenced by several factors such as the botanical origin and species, climatic conditions, soil conditions, harvesting conditions, refining and storage conditions.[1-10, 1-18] For example, linoleic acid concentration is significantly reduced in canola cultivars grown during drought.[1-19] Most crude plant-based edible oils are obtained from the seeds and fruits of plants. Oil extraction from seeds is achieved by pressing and/or by solvent extraction. Oil seeds include canola, corn, peanut and sunflower. Oils such as olive and palm, are pressed out of the soft fruit otherwise known as the endosperm. The mechanical method utilized in extraction of the oil can affect certain parameters of the oil. Virgin olive oil obtained by centrifugation method has been found to contain a lower polyphenol content compared to the one obtained by pressing and percolation methods.[1-18, 1-20] This is because

the centrifugation method requires lukewarm water in dilution of the paste which partially dissolves the polyphenol.

Most oils are refined, some more than others, before they are sold commercially. Typically, commercially sold edible oils go through the refining process to remove undesirable components and to ensure suitability for human consumption. Undesirable components can affect the taste, aesthetic value or induce oxidation or hydrolysis in oil. The refining process of oils is designed to remove undesirable materials such as monoacylglycerols, diacylglycerols, phospholipids, trace metals, oxidized materials and pigments. Refining is also intended to maximize desirable components like antioxidants and vitamins. Hydrogenation of the oil may also take place during refining. Brush hydrogenation, a very light form of hydrogenation, is applied to rapeseed oil and soybean oil to reduce the linolenic acid content. Although linolenic acid is an essential fatty acid, it is a cause of oxidative instability of these oils. Therefore, reducing the linolenic acid content through hydrogenation extends the shelf life of an edible oil.

As mentioned earlier, the major dietary fat from plants and animals takes the form of triacylglycerol. Triacylglycerols (TAGs) are commonly called triglycerides and are formed when three fatty acids react with 1, 2, 3-propanetriol (glycerol) to form an ester. The fatty acids in this reaction are almost invariably long-chain fatty acids (C_{14} to C_{22}). The type and position of the fatty acids attached to the glycerol backbone play an important role in determining the physical properties of a triglyceride. The length of the chain as well as the conformation and position of the double bond(s) can influence the physical properties such as melting point and flash point. Short chain fatty acids will melt at a lower temperature than those with a longer chain length. Melting point is also lowered with

increasing unsaturation of the oil. The stereochemical position of each fatty acid on the triglyceride can be a factor in determining the physical properties of the oil.[1-20] For oils that are considered edible, the unsaturated fatty acid of the C_{18} form are predominantly oleic and linoleic acids. Typically for these oils, odd numbered or branched-chain fatty acids are rarely observed. The chain length is usually between C_{16} and C_{24}. For instance, safflower and olive oil contain a high amount of oleic acid, whereas linseed oil (which is not considered edible) contains about 60% α-linolenic acid.

1.2.1. Minor Components of Edible Oils

1.2.1.1. Free Fatty Acids (FFAs)

Edible oils also contain minor components that contribute to their chemical profile or fingerprint. These include free fatty acids (FFAs), monoacylglycerols (MAGs) diacylglycerols (DAGs), phospholipids, tocopherols, pigments, water, vitamins, and trace metals. They comprise about 0.2 to 1.3% of edible oils. When TAGs react with water, FFAs and DAGs are formed. This hydrolysis reaction can be catalyzed by lipases at extreme pH and heat during the refining and storage of the oil. This hydrolysis reaction can also occur during the refining of the oil or when it is in storage. Short chain FFAs can also be formed from the secondary oxidation of unsaturated aldehydes which are produced from the cleavage of lipid hydroperoxides. Although FFAs are undesirable and are usually removed during refining, they cannot be eliminated because their concentration continues to increase as the oil ages.[1-21] The presence of FFAs in oil even at very low concentrations further catalyze the breakdown of TAGs, thereby causing a continuous increase in the level of FFA in the oil in storage and on the shelf. The amount of FFA in an edible oil is an indication of the degree of degradation of the oil and can impact flavor and taste and

decrease the smoke point of the oil. Such a reduction in temperature limits the degree to which the oil can be heated. FFAs also increase the acidity and lower the oxidative stability of the oil. Therefore, the amount of FFA is an important parameter of quality.

1.2.1.2. Mono- and Di-acylglycerols

MAGs and DAGs are formed from the hydrolysis of triglycerides during refining and storage. They may also be found naturally in some seed oils as they are precursors to the formation of TAGs. Therefore, they can be formed due to incomplete biosynthesis of TAGs.[1-22, 1-23] MAGs are monoesters of glycerol in which only one of the hydroxyl groups is esterified with a long-chain fatty acid. MAG has a glycerol backbone with two free hydroxyl groups while DAG has only one free hydroxyl group. For DAG, two of the hydroxyl groups have been esterified with fatty acids. MAG can exist as 1-MAG or 2-MAG based on the position of the ester bond on the glycerol. In 2014, Chen *et al*., determined that MAG at levels of 0.5 wt.% suppresses the effectiveness of α-tocopherol in soybean oil.[1-24] Alpha-tocopherol is considered an antioxidant and is a desirable constituent of edible oil. DAG has been found to increase the oxidation in edible oils leading to rancidity.[1-25] When compounds like FFAs, MAGs and DAGs are formed in edible oils, the surface tension of edible oils decreases with a corresponding increase in the diffusion rate of oxygen from the headspace of the container to the oil. This increase in oxygen diffusion leads to faster oxidation of the oil and a reduced shelf life.[1-26]

1.2.1.3. Phospholipids

Phospholipid typically refers to lipids that contain phosphoric acid and/or other phosphorus containing acids in ester form. They are made up of a diacylglycerol, a phosphate group, and a simple organic molecule, such as ethanolamine or choline.[1-27]

Phospholipids contain both hydrophilic and lipophilic groups and can make up 0.1 to 1.8% of the total lipids extracted from crude edible oils. In freshly refined oils, they are usually found in small amounts, about 40 to 135 mg/Kg. Phosphatidic choline is the most common phospholipid in seed oils. Phospholipids include phosphatidic acid, phosphatidylcholine and phosphatidylethanolamine.

Phospholipids are beneficial to human health as they exhibit antioxidative properties and are essential to the efficiency of cell membranes in living organisms. They are also easily oxidized by unsaturated fatty acids and have a gummy consistency, which gives the oil a cloudy appearance and causes foaming when frying. During storage, they can precipitate and form an unsightly residue. Phospholipids are usually removed during refining through the process of degumming.

1.2.1.4. Fat Soluble Vitamins

Fat soluble vitamins in edible oils are present as carotenes and tocopherols, which is also known as vitamin E. The amount of vitamin E present in edible oils varies among oils. Tocopherol can be present in two forms: alpha, beta, gamma and delta- tocopherol or tocotrienols. They are regarded as a class of phenolic antioxidants. Tocopherol is an antioxidant that protects the oil from auto-oxidation. They are considered desirable constituents because they are natural antioxidants. The concentration of tocopherols in edible oils can range from 200 to 1000 ppm. The antioxidant activity of tocopherols increases from the α- through the δ-isomer, while the reverse is observed for the vitamin activity which decreases from the α through the δ-isomer.[1-28] Tocopherols scavenge radicals and react with singlet oxygen. Therefore, they inhibit the rate of photooxidation and autooxidation in edible oils.

1.2.1.5. Beta-carotene, Sterols and Metals

Beta-carotene is another antioxidant that is present in edible oils. It is a lipid soluble natural pigment that can prevent Vitamin A deficiency. However, deodorizing edible oils will result in complete removal of the beta-carotene content. For this reason, beta-carotene is usually added to edible oils after deodorization. Sterols obtained from plants are usually referred to as phytosterols. Phytosterols are naturally occurring steroid alcohols and make up about 0.1 to 1% of seed oils. Phytosterols may compete with cholesterol at absorption sites in fat tissues. Therefore, consumption of phytosterols may effectively lower blood cholesterol levels. Metals such as iron and copper may also be present in edible oils. The presence of these transition metals can facilitate the decomposition of lipid hydroperoxides and increase local radical concentration leading to faster oil degradation. These metals also contribute to the decomposition of antioxidants such as tocopherols and phenolic compounds during frying.[1-29]

1.2.1.6. Other Minor Constituents

Other minor constituents of edible oil include oxygen, water, and hydrocarbons. Hydrocarbons include compounds such as alkanes, alkenes and squalene. The hydrocarbons present in olive oil consist mainly of squalene, which makes up about 90% of the total hydrocarbon. Squalene ($C_{30}H_{50}$) is an open chain triterpene which is highly unsaturated and an intermediate compound in the biosynthesis of sterols. Under extreme heat and light conditions, oxidative degradation of squalene can occur, leading to the formation of aldehydes and ketones which may lead to an off-flavor in the oil.[1-30, 1-31]

1.3. Differences between Plant-based and Animal-based Oils.

The edible oils investigated in the research described in this dissertation are plant-based edible oils. Plant-based edible oils differ from oils obtained from animal sources in a variety of ways. Cholesterol, the most abundant sterol in animal fats and oils is only present in plant-based edible oils in negligible amounts.[1-32] In plant-based edible oils, β-Sitosterol is found to be predominant in the sterol fractions except in pumpkin seed oil.[1-33] Animal oils contain a large amount of SFAs compared to vegetable oils.[1-34] Oils from animal sources tend to have more free fatty acids and are composed mainly of MUFAs while oils obtained from plant sources are composed mainly of PUFA.[1-35]

1.4. Detecting Adulteration in Edible Oils

1.4.1. Gas Chromatography and Liquid Chromatography

Techniques such as capillary column gas chromatography (CGC), gas chromatography/mass spectrometry (GC/MS), gas chromatography-mass spectrometry/mass spectrometry (GC-MS/MS), gas chromatography equipped with flame ionization detection (GC-FID) and high-performance liquid chromatography (HPLC) have been used to successfully authenticate edible oils. These techniques are used for both qualitative and quantitative analyses of compounds. Many established international regulations have defined the quality of edible oils based on these techniques. It is also not uncommon for researchers to combine these methods with chemometric techniques in authentication studies. Chromatographic techniques often require isolation and/or derivatization of compounds using procedures that require an appreciable amount of manual work and time. The instrumentation for these methods is expensive compared to

spectroscopic procedures. Moreover, gas chromatography and liquid chromatography are difficult to use for field work and on-line monitoring.

An American Oil Chemist's Society (AOCS) method has been developed using GC employs standards containing different alkyl chain lengths which must be injected under the selected conditions of the experiment in order to establish the retention indices of the metabolite. In analyzing a sample, the triglycerides are broken down into fatty acids using organic solvent(s). The lipids are then methyl esterified to form fatty acid methyl esters (FAMEs). In this method, the edible oil is first heated with methanol and sodium methylate under reflux. Extraction is then carried out using diethyl ether and the ester is analyzed using GC-FID.[1-36] This method, though sensitive, is laborious and time consuming.

Al-Ismail *et al*., 2010 investigated adulteration in olive oil by corn, soybean, sunflower and cotton seed oils.[1-37] CGC with a polar column was used to monitor the four different types of sterols in olive oil and in prepared adulterated samples. This method took advantage of the difference in the campesterol and stigmasterol content of olive oil compared to other plant-based oils as the campesterol content in olive oil must be less than or equal to 4% according to EC regulation 2568, (enacted in 1991). The workers found that the amount of sterols in olive oils increased as the amount of adulteration increased. However, this method cannot be used to identify the type of oil used as the adulterant. Also, it is problematic to use this method for the detection of olive oil adulteration by hazelnut oil because of the low campesterol and stigmasterol content of hazelnut oil.

In another study carried out by Mariani *et al.* in 2006, the free (polar fraction) and esterified (nonpolar fraction) of sterols from refined olive oils and their adulterated mixtures were separated using silica gel chromatography.[1-38] This method involved the

preparation of a silica gel column which was laborious. The workers were able to achieve limits of detection ranging from 5% to 10% for the presence of refined hazelnut oils in olive oils depending upon the variety of olive oil. Li *et al.*, 2016, used GC-MS for the classification of six different types of vegetables. [1-39]. The mass selective detector allowed for the identification of the compounds comprising the oils. The vegetable oils investigated were soybean, camellia, peanut, corn germ, sesame and rapeseed oils. In their study, the Kennard-Stone algorithm was used to select the samples that comprised the training set. A genetic algorithm optimized support vector machine was used to classify the samples with misclassification rates of 8.48% and 3.03% for the training and test sets respectively.

Methyl trans-esterified fractions of pure olive oil (e.g., extra virgin and virgin olive oils) and adulterated olive oils were assessed using HPLC.[1-40] The oils used for the preparation of the adulterated blends included soya, sunflower, canola, corn, peanut, sesame and grapeseed oils. The trans-esterified fraction was prepared using a mixture of sodium methoxide in methanol in methyl *tert*-butyl ether along with centrifuging with water and *n*-hexane. The data obtained was taken through several preprocessing steps including filtering to eliminate noise, baseline correction, peak alignment and mean centering. Partial least squares regression (PLS-R) and support vector-regression (SV-R) were performed on the preprocessed data obtained from the chromatograms. Two forms of SV-R were tested without X-block compression, followed by testing with X-block compression using principal component analysis (PCA) and PLS. In comparing the results obtained from PLS-R with SV-R, certain quality metrics were used. These included R^2, root mean square error of validation (RMSEV), mean absolute error of validation (MAEV) and median absolute error of validation (MdAEV). PLS-R was found to produce

comparable R^2 values with SV-R while the errors (RMSEV, MAEV and MdAEV), were generally lower for PLS-R than for SV-R.

Characterization and classification of several edible oils were performed through polyphenolic fingerprints acquired at three wavelengths (257 nm, 280 nm and 316 nm) using HPLC.[1-41] The edible oils employed in this study included olive, sunflower, corn and soy oils. Extraction was achieved using ethanol and water along with 24 hours freezing at -18 ^{o}C. The extract was defatted using hexane and the aqueous ethanolic extract was analyzed by HPLC. Principal component analysis was performed on data obtained from the raw chromatograms. While there was a high degree of discrimination between olive oils and the other edible oils, discrimination among the other edible oils was not achieved. Separation was better with the data acquired at 257 nm compared to the other two wavelengths. There was no clear separation between olive oils and the other oils with the other two wavelengths. Mixtures were prepared consisting of Arbequina extra virgin olive oil (EVOO) adulterated with Picual EVOO, refined olive oil, or sunflower oil. PLS-R was applied to the data obtained from these mixtures with the overall errors in the quantitation of adulteration in the Arbequina EVOO (minimum of 2.5% adulteration) below 2.9%.

Data from different analytical sources have been combined using a technique known as data fusion. It is expected that fusion of analytical data should provide broader and more accurate information about the sample. In a method proposed by Vera et al., data from reverse and normal phase HPLC using a charged aerosol detector (HPLC-CAD) was combined with data obtained from a high temperature GC-FID using low- and high-level fusion.[1-42] Denoising and smoothing was applied to the data followed by baseline correction, chromatogram profile alignment and mean centering. SIMCA, along with low-

and high-level fusion techniques, were applied to authenticate olive oil samples from a particular cultivar based on its geographical region. The results from SIMCA were compared to those obtained with partial least squares-discriminant analysis (PLS-DA). The best classification result obtained from this study was from PLS-DA high-level fusion. This result was superior in terms of sensitivity, specificity and inconclusive classifications than SIMCA and superior to the individual fingerprint of each technique. Approaches using data fusion are often expensive and time consuming because data from multiple techniques are combined leading to increased cost.

1.4.2. Spectroscopy

Matrix-assisted laser desorption/ionization-imaging mass spectroscopy **(**MALDI-MSI) has been used to investigate the adulteration of edible oils that often occur in used cooking oils.[1-43] Principal component analysis (PCA) score plots showed that fresh edible oils could be readily differentiated from deep-fried and gutter oils (reprocessed cooking oils). Discrimination of deep-fried cooking oils from recycled cooling oils was more challenging. The workers postulated that molecular ions with m/z of 1752.5 and 795.6 are crucial for achieving this discrimination. The second part of this study entailed the discrimination of nine varieties of edible oils including canola, corn, grapeseed, olive, palm, peanut sesame, soybean and sunflower oils. Each edible oil variety could be differentiated from the others using PCA.

A methodology using direct electron impact ionization -mass spectrometry (EI-MS) was employed to assess the authenticity of edible oils.[1-44] This technique authenticates edible oils based on their m/z fragmentation pattern. No chromatographic separation is performed with this technique. The oils used in this study were virgin olive, hazelnut,

sunflower, soybean, cotton seed and black cumin oils. Esterification of the triglycerides to fatty acid methyl ester was performed and the product was stored at -30 ^{o}C. The fatty acid profiles of the edible oils were determined by GC-FID. Several preprocessing methods were applied to the data obtained from EI-MS, and it was determined that standard normal variate coupled to Savitzky-Golay smoothing gave the best results. After preprocessing, different regions of the recorded m/z values between 90 and 400 were examined. The spectrogram was divided into two regions and divided again in increments of ten before PCA was performed. With PCA, some of the plots showed hazelnut and virgin olive oil clustering together while the other oils remained in a different cluster, whereas other plots showed hazelnut oil well discriminated from virgin olive oil. It was difficult to obtain good separation within the cluster of the other oils using PCA score plots. Linear discriminant analysis applied to the same data showed good discrimination among the oils except for soybean oil, which was projected in the sunflower oil cluster and the black cumin seed oil cluster.

1.4.3. Nuclear Magnetic Resonance Spectroscopy

^{1}H, ^{13}C and ^{31}P NMR have also been used to study the authenticity of edible oils. Although the NMR technique is easy to use and may not require derivatization or elaborate extraction procedure as in chromatography or mass spectrometry, NMR instruments as with MALDI are large and very costly compared to optical methods and cannot be employed for infield operation. The fatty acid that are esterified to a glycerol backbone give specific NMR signals, which on visual inspection can allow direct identification and quantification. High resolution ^{1}H NMR spectroscopy has been applied to the determination of the fatty acid composition of edible oils. The use of an ultrafast low field

2D benchtop NMR was reported by Gouilleux and co-workers in the discrimination of six varieties of edible oils.[1-45] Although, the use of 2D high field NMR requires more time, the data obtained minimizes ambiguity in peak assignments and generally yields a more comprehensive profile of the sample compared to 1D ^{1}H NMR. The NMR used in the report by Gouilleux is a low field instrument with 2D capabilities. It acquires the 2D spectra in the same time frame as 1D NMR. This low field 2D NMR was used for authentication studies of edible oils because edible oils are considered concentrated (96% triglyceride). PCA was able to classify the 23 edible oil samples into 6 groups. Apart from this classification study, the adulteration of olive oil by hazelnut oil was also investigated. The workers modelled the percent adulteration of hazelnut oil in each sample of olive oil using a two-component PLS regression model, where $R^2 = 0.9801$ and RMSEP = 6.27% w/w.

In another study, ^{13}C NMR was used to characterize olive oils from different cultivars and geographical regions.[1-46] The application of stepwise discriminant analysis allowed the classification of virgin olive oil and oils with a higher amount of oleic acid from those with a lower amount of linoleic acid. Although discrimination between oils based on high oleic acid and high linoleic acid content is straight-forward, it is problematic to discriminate virgin olive oil from oils with high oleic acid content. It was also found that virgin olive oil could be discriminated from virgin olive oil adulterated by hazelnut oil (5% to 20%).

The total amount of sterols *in* edible oils was also determined using ^{31}P.[1-47] In this study, the olive oils were phosphitylated by the derivatization of the labile hydrogens of the hydroxyl groups of the diglycerides (1,2- and 1,3-) with 2-chloro-4,4,5,5-

tetramethyldioxaphospholanels. This method involved an elaborate preparation of the stock solution which comprised pyridine, $CDCl_3$, chromium acetylacetonate and cyclohexanol (internal standard). The stock solution was mixed with the edible oil sample before phosphitylation was performed. The duration of the measurements was reduced by the presence of the paramagnetic metal center of $Cr(acac)_3$ which helped to lower the relaxation times of the phosphorus nuclei. The diglyceride content was used to differentiate the olive oils from the other vegetable oils.

1.4.4. UV/Visible Spectroscopy

Characterization and authentication of edible oils have also been performed using UV absorbance spectroscopy. In one report, cold pressed oils were mixed in their refined version and UV spectroscopy was used to discriminate the pure cold pressed oil from the adulterated oils.[1-48] Four types of oils were studied: canola, coconut, sunflower and grapeseed oils. Visual inspection, after the use of illuminant D65, showed that refined oils were lighter in color compared with the corresponding cold pressed samples. This trend was observed for all the oils except for coconut oil. The absorbance spectra of the pure and adulterated mixtures showed a triplet between 400 and 500 nm and a small peak at 650 nm. These peaks were monitored to establish the percent of adulteration present in the cold pressed oils. The refined oils were found to be missing these two peaks. However, the triplet peak was not observed for cold pressed coconut oil. This peak diminished as the percent of refined oils increased. According to this report, this triplet peak may be associated with concentrations of oleic acid, linoleic acid and linolenic acid that were found to be significantly lower in the refined oils employed in this study compared to cold pressed oils.

Another study used UV-Visible absorbance spectroscopy to quantify the degree of adulteration of EVOO with refined olive oil and with refined olive-pomace oil.[1-49] Some parameters were calculated using data obtained from the UV-Vis absorbance spectra of the mixtures of EVOO and refined olive oil, and EVOO and refined olive-pomace oils. These parameters showed a linear correlation with the amount of EVOO adulteration. Upon validation of the model, a mean square error of 1% and a mean correlation coefficient of 0.97 were obtained.

In one of the studies performed by Didham *et al*., the UV-Vis absorbance spectra of pure and adulterated olive oils were obtained.[1-50] These spectra showed absorption at specific wavelengths associated with polyphenols, carotenoids and chlorophyll. The data obtained from these spectra were subjected to PCA and a high degree of discrimination was obtained between pure EVOO, canola and sunflower oils. Analysis of the loadings showed that the largest eigenvectors obtained for the first three principal components correspond to the wavelengths (400, 500, and 650 nm) mentioned earlier in this section. However, this analysis appears to be limited to only one sample of sunflower oil and one sample of canola oil.

1.4.5. Vibrational Spectroscopy

Vibrational spectroscopy has proved to be a very useful technique in the authentication of edible oils because of the detailed information on chemical composition and existing functional groups provided by this technique. In contrast to chromatography and mass spectrometry-based techniques and others mentioned previously discussed in this chapter, vibrational spectroscopic based methods are rapid, non-destructive and produce

less chemical waste. Vibrational spectroscopy, therefore, provides a fast, simple and reproducible means for evaluating edible oils. Vibrational spectroscopy can also be used for in-line and in-field measurements and can be easily automated. Vibrational spectroscopy in combination with chemometrics is an even more powerful tool in the authentication of edible oils where both quantitative and qualitative information about the composition of edible oils can be obtained from the spectroscopic fingerprint of each edible oil. Analysis of these fingerprints by pattern recognition methods is crucial to ensure the effective extraction of qualitative and quantitative information necessary to verify the authenticity of an edible oil and to detect adulteration.

Multivariate classification and calibration methods have been applied to spectra obtained from infrared and Raman spectroscopy of edible oils to improve classification success rates and to obtain lower detection limits for adulterants.[1-51 – 1-58] Generally, classification success rates of around 90% for edible oils have been reported in the literature with detection limits of adulterants at 10% using IR spectroscopy. However, these studies were generally limited to approximately ten samples spanning five or six edible oils using either PLS or linear discriminant analysis to perform a horizontal (i.e., flat) classification of the data. Furthermore, the edible oils investigated in almost all of these studies were represented by a single source (often a single bottle of an edible oil from a single brand) or a few brands. Thus, edible oils from a particular batch are "nicely" clustered and differentiated from other varieties of edible oils that are also obtained from a single source (specific brand and batch). For the studies carried out in this dissertation, a variety of samples were systematically collected over three years to account for seasonal variation in

the composition of the edible oils as well as manufacturer to manufacturer variation. This dissertation addresses the issue of sample variability in the authentication of edible oils.

1.4.5.1. Near infrared spectroscopy (NIR)

In NIR absorbance spectroscopy, the absorption of electronic radiation by a sample is measured in the range of 750 to 2500 nm. NIR bands are typically broad as they are combinations and overtones of vibrational modes of the C-H, O-H and N-H bonds. NIR spectroscopy in combination with pattern recognition techniques has been applied to the discrimination of edible oils such as coconut oil, olive oil, rice bran oil, sesame oil, soybean oil and sunflower oil.[1-59] Coconut oil was well separated from the other oils in the PCA score plot. The other edible oil varieties could not be discriminated from each other in the score plot. Pattern recognition techniques including SIMCA, PLS-DA, k-nearest neighbor (k-NN), support vector machines (SVM), and neural networks were then applied to this data set to identify the different varieties of edible oils. The results from these five pattern recognition techniques were compared. PLS-DA and SVM performed best in terms of the precision and accuracy of the classification models developed.

The application of 2D NIR spectroscopy for discrimination of edible oils has also been investigated using four different varieties of edible oils (sesame, peanut, palm and soybean oils).[1-60] The NIR spectra for each variety of the edible oils were similar with only slight differences due to significant peak overlap which is typical of NIR spectra. Clearer peak information was obtained in the range of 8750 cm^{-1} to 4500 cm^{-1} when 'auto-power spectra' were used as dynamic NIR spectra and were collected at temperatures between 50 to 160 °C at intervals of 10 °C. This allowed for the discrimination of the oils by variety and by manufacturer. The contour level was set at 16, and it was observed that all the four

edible oils exhibited an obvious auto-peak at 4663 cm^{-1}. Many differences were observed in other areas of the dynamic spectra. Therefore, these four varieties of edible oils could be discriminated from each other. The study also assessed the discriminating power of 2D NIR spectroscopy using peanut oils from different manufacturers. Peanut oils from two holding companies were used in this study. Differences in these peanut oils were attributable to the effects of process technology on the quality of the oils. This study would be more impactful if edible oils from a larger number of manufacturers were used. Unfortunately, this technique destroys the sample due to heating.

1.4.5.2. Mid Infrared (MIR) Spectroscopy

Mid IR spectroscopy has been widely used as a technique for authentication of edible oils.[1-61 - 164] Qualitative and quantitative investigations can be performed using this method. It requires very little sample preparation and no hazardous chemical reagents or solvents are needed. MIR spectroscopy performed using a Fourier transform infrared (FTIR) spectrometer uses an interferogram to adaptively recalculate the spectrum. The application of FTIR spectroscopy can help with improved accuracy of detection for adulterants. In recent years, there has been an increase in the use of FTIR spectroscopy because of the ease of combining it with chemometrics. FTIR gives the chemical fingerprint of a sample which can be exploited using multivariate techniques to obtain valuable information about the composition and structure present in a sample. The challenge with visual assessment of edible oil spectra for the purpose of discrimination and detection of adulterants is the similarity of the spectra of the oils. Individual inspections of spectra can be tedious and time consuming especially when comparing many spectra.

A study discussed earlier in this chapter that employed UV absorbance spectroscopy also utilized MIR spectroscopy as part of the study design.[1-50] This study utilized an FTIR spectrometer equipped with a ZnSe window upon which samples were dropped. Savitzky-Golay filtering was performed using the second derivative with a 20-point window for smoothing and a second order polynomial. The data set was divided into a high range from 10% to 50% adulteration, in increments of 10% w/w and a low range from 0.2% to 10% in increments of 0.2% w/w. Principal component analysis (PCA) and partial least squares discriminant analysis (PLSDA) were used to identify trends present in the spectra and to classify samples according to whether they were pure or adulterated. The PLS-DA model for MIR data gave an R^2 and SECV of 0.98 and 3.05 for the entire range (0.2% to 50%) and 0.91 and 1.01 for the low range. The authors concluded that this technique can reliably detect adulteration above 10% because of the low error in prediction. They also concluded that adulteration levels below 10% may be problematic to detect. The authors used one or two samples each of olive, sunflower and canola in their study.

MIR has also been used to study the adulteration of olive oil with refined and lampante virgin olive oils.[1-65] This study also investigated the adulteration of olive oil by hazelnut oil with percent adulteration ranging from 2% to 20%. An FTIR spectrometer fitted with a horizontal ATR ZnSe crystal with six internal reflections was employed for the absorption measurements. Multivariate analyses using stepwise linear discriminant analysis (SLDA), was performed on the edible oils using the full spectral range and on the unsaponifiable matter extracted from the oils. The SLDA models selected ten wavelengths for the oils and eight wavelengths for the unsaponifiable matter. The percentage of correct classification for olive oil adulterated by hazelnut oil were higher when the models were

constructed using the extracted unsaponifiable matter than the original edible oils. This same trend was also observed for refined and lampante virgin olive oils. For the test set, the unsaponifiable matter samples also gave the best results. However, two false positives were reported. Furthermore, the SLDA plot only showed clusters representative of olive oil, hazelnut oil, or adulterated olive oil. It is difficult to discern whether the spectra of the mixtures (olive oil adulterated with hazelnut oil or refined and lampante virgin olive oil) were linearly additive.

1.4.5.3. Raman Spectroscopy

Raman spectroscopy has been applied to classification and authentication of edible oils.[1-66 – 1-69] A previous study carried out in our research group entailed the use of Raman spectroscopy for the classification of edible oils.[1-70] This study incorporated brand variability through a systematic collection of samples from several manufacturers over a three-year period. This study used 15 varieties of edible oils/blends spanning 53 samples from which 215 Raman spectra were collected. Prior to pattern recognition analysis, the spectra were filtered, smoothed and normalized. The spectra were then truncated to five bands ranging from 1750 cm^{-1} to 1270 cm^{-1}. These bands were found to be informative in accounting for the differences between the classes in the training set. The 15 varieties of edible oils were partitioned into five distinct groups based on their degree of saturation and the ratio of polyunsaturated fatty acids to monounsaturated fatty acids. Edible oils assigned to one group could be readily differentiated from those assigned to other groups, whereas Raman spectra within the same group more closely resembled each other and were more difficult to classify by variety.

Raman spectroscopy has also been used to distinguish pure edible oils from used cooking oils.[1-71] In order to discriminate used cooking oils from pure edible oils such as olive, rapeseed, soybean, corn or peanut oils, spectral bands at 869, 969, 1302 and 1080 cm^{-1} were found to be crucial. Heated oil in this study is defined as edible oil heated at 200 ^{o}C for one hour, whereas waste cooking oil is the oil obtained from a designated company for the disposal of kitchen waste. The authors prepared adulterated mixtures by mixing waste cooking oils with soybean, peanut and olive oils in different proportions. PCA score plots of the adulterated mixtures indicated a linear additive model for the components comprising the adulterated mixtures. However, the original waste and pure oils were not included in these score plots. The authors also noted two characteristic peaks, one at 1183 and the other at 1554 cm^{-1}, for the waste cooking oils. These same peaks were also observed for heated oils. The authors opined that the vibrational modes corresponding to these two peaks have not been reported in the literature and attributed them to repeated heating. Furthermore, the peak at 1441 cm^{-1} was unaffected by the proportions of waste cooking oils added to the pure oils. The authors were able to identify five peaks related to the degree of adulteration and plotted the ratio of each of these peaks and the peak at 1441 cm^{-1} against percent adulteration of the mixtures to obtain R^2 values in the range of 0.97. As in the case of other studies reported in the published literature, only one sample was used for each of the five pure oils, the waste cooking oil and the heated oil.

1.5. Organization of this Dissertation

The focus of the research described in this dissertation is the application of FTIR spectroscopy and pattern recognition methods to the problem of authenticating plant-based

edible oils. FTIR spectra of ninety-seven edible oil samples from twenty plant-based varieties collected over a three-year period were analyzed using the four major types of pattern recognition methods: mapping and display, cluster analysis, variable selection, and classification. The ninety-seven edible oil samples selected for this study encompass multiple brands representing both supplier to supplier variation and seasonal and batch variation within a supplier. Using a hierarchical classification scheme, the twenty plant-based varieties of edible oils can be divided into four distinct edible oil groups. Edible oils from different oil groups can be reliably discriminated, whereas the discrimination of edible oils within the same group is problematic. Adulteration of plant based edible oils by other edible oils in the same group (e.g., extra virgin olive oil by almond oil) cannot be reliably detected using FTIR spectroscopy, whereas adulteration of edible oils by other oils not in the same group (e.g., EVOO adulterated by corn or canola oil) can be detected at concentration levels as low as 10% (v/v) which is consistent with the results reported in previously published studies **using partial least squares regression.** A unique aspect of this work is the incorporation of edible oils collected systematically over three years, which introduces a heretofore unseen variability in the chemical composition of the edible oils. This work also demonstrates that previously published studies (which have relied on a single source to represent each type of edible oil) provide an overly optimistic estimate of the capability of FTIR spectroscopy to discriminate plant based edible oils by type and to detect the presence of adulterants in edible oils.

This dissertation is divided into five chapters. The first chapter is entitled, "Introduction," and provides a brief survey of the literature on edible oils and analytical methods used to discriminate the different varieties of plant-based edible oils and detect

the presence of adulterants in these edible oils. The second chapter focuses on the methodology used in the dissertation research, whereas the third and fourth chapters highlight the challenges encountered when using FTIR spectroscopy and chemometrics to authenticate edible oils. The fifth and final chapter summarizes the research reported in this dissertation with suggestions about potential future directions.

R**eferences**

1-1. Holbrook, E., Dining on deception: the rising risk of food fraud and what is being done about it. *Risk Management* **2013**, p 28+.

1-2. Rifna, E. J.; Pandiselvam, R.; Kothakota, A.; Subba Rao, K. V.; Dwivedi, M.; Kumar, M.; Thirumdas, R.; Ramesh, S. V., Advanced process analytical tools for identification of adulterants in edible oils – A review. *Food Chemistry* **2022,** *369*, 130898.

1-3. Spink, J.; Moyer, D. C., Defining the public health threat of food fraud. *Journal of Food Science* **2011,** *76* (9), R157-R163.

1-4. Gelpí, E.; de la Paz, M. P.; Terracini, B.; Abaitua, I.; de la Cámara, A. G.; Kilbourne, E. M.; Lahoz, C. et al. "The Spanish toxic oil syndrome 20 years after its onset: a multidisciplinary review of scientific knowledge." *Environmental Health Perspectives* **2002,** *110*(5), 457-464.5.

1-5. Ruiz-Gutiérrez, V.; Maestro-Durán, R., Lymphatic absorption of 3-phenylamino-1,2-propanediol and its esters. *Exp. Toxicol. Pathol.* **1992,** *44* (1), 29-33.

1-6. World Health Organization, E., Toxic Oil Syndrome - 10 Years of Progress. **2004**, 5-27.

1-7. Babu, C. K.; Khanna, S. K.; Das, M., Adulteration of mustard cooking oil with argemone oil: do Indian food regulatory policies and antioxidant therapy both need revisitation? *Antioxid. Redox Signal* **2007,** *9* (4), 515-25.

1-8. Sood, N. N.; Sachdev, M. S.; Mohan, M.; Gupta, S. K.; Sachdev, H. P., Epidemic dropsy following transcutaneous absorption of Argemone mexicana oil. *Trans R Soc. Trop. Med. Hyg.* **1985,** *79* (4), 510-2.

1-9. Gomber, S.; Daral, T. S.; Sharma, P. P.; Faridi, M. M., Epidemic dropsy in Trans Yamuma areas of Delhi and U.P. *Indian Pediatr.* **1994,** *31* (6), 671-4.

1-10. Timms, R. E., Physical properties of oils and mixtures of oils. *Journal of the American Oil Chemists' Society* **1985,** *62* (2), 241-249.

1-11. Lichtenstein, A. H., Fats and Oils. In *Encyclopedia of Human Nutrition (Third Edition)*, Caballero, B., Ed. Academic Press: Waltham, 2013; pp 201-208.

1-12. Patel, A.; Rova, U.; Christakopoulos, P.; Matsakas, L., Introduction to Essential Fatty Acids. John Wiley & Sons, Inc: Hoboken, NJ, USA, 2020; pp 1-22.

1-13. Mozaffarian, D.; Micha, R.; Wallace, S., Effects on coronary heart disease of increasing polyunsaturated fat in place of saturated fat: a systematic review and meta-analysis of randomized controlled trials. *PLoS Medicine* **2010,** *7* (3), e1000252.

1-14. Kris-Etherton, P. M., Monounsaturated fatty acids and risk of cardiovascular disease. *Circulation* **1999**; *100* (11), 1253-1258.

1-15. FAO/WHO In *Fats and fatty acids in human nutrition. Report of an expert consultation, 10-14 November **2008**; Geneva*, FAO/WHO, Geneva, Switzerland, Geneva, Switzerland, 2008.

1-16. Nagy, K.; Tiuca, I.-D., Importance of fatty acids in physiopathology of human body. In *Fatty acids*, IntechOpen: 2017.

1-17. Valentine, R. C.; Valentine, D. L., Omega-3 fatty acids in cellular membranes: a unified concept. *Progress in Lipid Research* **2004**; *43* (5), 383-402.

1-18. Di Giovacchino, L.; Solinas, M.; Miccoli, M., Effect of extraction systems on the quality of virgin olive oil. *Journal of the American Oil Chemists' Society* **1994**; *71* (11), 1189-1194.

1-19. Folayan, A. J.; Anawe, P. A. L.; Aladejare, A. E.; Ayeni, A. O., Experimental investigation of the effect of fatty acids configuration, chain length, branching and degree of unsaturation on biodiesel fuel properties obtained from lauric oils, high-oleic and high-linoleic vegetable oil biomass. *Energy Reports* **2019**; *5*, 793-806.

1-20. Di Giovacchino, L.; Sestili, S.; Di Vincenzo, D., Influence of olive processing on virgin olive oil quality. *European Journal of Lipid Science and Technology* **2002**; *104* (9-10), 587-601.

1-21. Lanser, A. C.; List, G. R.; Holloway, R. K.; Mounts, T. L., FTIR estimation of free fatty acid content in crude oils extracted from damaged soybeans. *Journal of the American Oil Chemists' Society* **1991**; *68* (6), 448-449.

1-22. Angerosa, F.; Campestre, C.; Giansante, L., 7 - Analysis and authentication. In *Olive Oil (Second Edition)*, Boskou, D., Ed. AOCS Press: **2006**; pp 113-172.

1-23. Kahveci, D.; Zhong, N.; Xu, X., Chapter 8 - Ionic liquids in Acylglycerol Synthesis and Modification. In *Ionic Liquids in Lipid Processing and Analysis*, Xu, X.; Guo, Z.; Cheong, L.-Z., Eds. AOCS Press: **2016**; pp 251-278.

1-24. Chen, B.; McClements, D. J.; Decker, E. A., Impact of diacylglycerol and monoacylglycerol on the physical and chemical properties of stripped soybean oil. *Food Chemistry* **2014**; *142*, 365-372.

1-25. Wang, Y.; Zhao, M.; Tang, S.; Song, K.; Han, X.; Ou, S., Evaluation of the oxidative stability of diacylglycerol-enriched soybean oil and palm olein under rancimat-accelerated oxidation conditions. *Journal of the American Oil Chemists' Society* **2010**, *87* (5), 483-491.

1-26. Choe, E.; Min, D. B., Mechanisms and factors for edible oil oxidation. *Comprehensive Reviews in Food Science And Food Safety* **2006**, *5* (4), 169-186.

1-27. Bowen-Forbes, C. S.; Goldson-Barnaby, A., Chapter 21 - Fats. In *Pharmacognosy*, Badal, S.; Delgoda, R., Eds. Academic Press: Boston, **2017**; pp 425-441.

1-28. Kamal-Eldin, A.; Appelqvist, L. A., The chemistry and antioxidant properties of tocopherols and tocotrienols. *Lipids* **1996**, *31* (7), 671-701.

1-29. Frankel, E., Lipid oxidation. *Lipid oxidation.* **2005**, (Ed. 2).

1-30. Budge, S. M.; Barry, C., Determination of squalene in edible oils by transmethylation and GC analysis. *MethodsX* **2019**, *6*, 15-21.

1-31. Shimizu, N.; Ito, J.; Kato, S.; Eitsuka, T.; Miyazawa, T.; Nakagawa, K., Significance of Squalene in Rice Bran Oil and Perspectives on Squalene Oxidation. *J Nutr Sci Vitaminol (Tokyo)* **2019**, *65* (Supplement), S62-s66.

1-32. Sheppard, A. J.; O'Dell, R. G.; Pennington, J. A. T., Cholesterol | Properties and Determination. In *Encyclopedia of Food Sciences and Nutrition (Second Edition)*, Caballero, B., Ed. Academic Press: Oxford, **2003**; pp 1220-1226.

1-33. Jeong, T. M.; Itoh, T.; Tamura, T.; Matsumoto, T., Analysis of sterol fractions from twenty vegetable oils. *Lipids* **1974**, *9* (11), 921-927.

1-34. Mattson, F.; Lutton, E., The specific distribution of fatty acids in the glycerides of animal and vegetable fats. *J Biol Chem* **1958**, *233* (4), 868-871.

1-35. Canakci, M.; Van Gerpen, J., Biodiesel production from oils and fats with high free fatty acids. *Transactions of the ASAE* **2001**, *44* (6), 1429.

1-36. AOCS, Fatty Acid Composition by Gas Chromatography. In *AOCS Method*, Society, A. O. C., Ed. AOCS Official Methods **2005**; Vol. AOCS Method Ce 1-62.

1-37. Al-Ismail, K. M.; Alsaed, A. K.; Ahmad, R.; Al-Dabbas, M., Detection of olive oil adulteration with some plant oils by GLC analysis of sterols using polar column. *Food Chemistry* **2010**, *121* (4), 1255-1259.

1-38. Mariani, C.; Bellan, G.; Lestini, E.; Aparicio, R., The detection of the presence of hazelnut oil in olive oil by free and esterified sterols. *European Food Research and Technology* **2006**, *223* (5), 655-661.

1-39. Li, X.; Kong, W.; Shi, W.; Shen, Q., A combination of chemometrics methods and GC–MS for the classification of edible vegetable oils. *Chemometrics and Intelligent Laboratory Systems* **2016**, *155*, 145-150.

1-40. Jiménez-Carvelo, A. M.; González-Casado, A.; Cuadros-Rodríguez, L., A new analytical method for quantification of olive and palm oil in blends with other vegetable edible oils based on the chromatographic fingerprints from the methyl-transesterified fraction. *Talanta* **2017**, *164*, 540-547.

1-41. Carranco, N.; Farrés-Cebrián, M.; Saurina, J.; Núñez, O., Authentication and quantitation of fraud in extra virgin olive oils based on HPLC-UV fingerprinting and multivariate calibration. *Foods* **2018**, *7* (4), 44.

1-42. Vera, D. N.; Jiménez-Carvelo, A. M.; Cuadros-Rodríguez, L.; Ruisánchez, I.; Callao, M. P., Authentication of the geographical origin of extra-virgin olive oil of the Arbequina cultivar by chromatographic fingerprinting and chemometrics. *Talanta* **2019**, *203*, 194-202.

1-43. Cao, G.; Hong, Y.; Wu, H.; Chen, Z.; Lu, M.; Cai, Z., Visual authentication of edible vegetable oil and used cooking oil using MALDI imaging mass spectrometry. *Food Control* **2021**, *125*, 107966.

1-44. Kenar, A.; Çiçek, B.; Arslan, F. N.; Akin, G.; Elmas, Ş. N. K.; Yilmaz, I., Electron Impact–Mass Spectrometry Fingerprinting and Chemometrics for Rapid Assessment of Authenticity of Edible Oils Based on Fatty Acid Profiling. *Food Analytical Methods* **2019**, *12* (6), 1369-1381.

1-45. Gouilleux, B.; Marchand, J.; Charrier, B.; Remaud, G. S.; Giraudeau, P., High-throughput authentication of edible oils with benchtop Ultrafast 2D NMR. *Food Chemistry* **2018**, *244*, 153-158.

1-46. Zamora, R.; Alba, V.; Hidalgo, F. J., Use of high-resolution 13C nuclear magnetic resonance spectroscopy for the screening of virgin olive oils. *Journal of the American Oil Chemists' Society* **2001**, *78* (1), 89-94.

1-47. Vigli, G.; Philippidis, A.; Spyros, A.; Dais, P., Classification of edible oils by employing 31p and 1h nmr spectroscopy in combination with multivariate statistical analysis. A Proposal for the Detection of Seed Oil Adulteration in Virgin Olive Oils. *Journal of Agricultural and Food Chemistry* **2003**, *51* (19), 5715-5722.

1-48. Popa, S.; Milea, M. S.; Boran, S.; Nițu, S. V.; Moșoarcă, G. E.; Vancea, C.; Lazău, R. I., Rapid adulteration detection of cold pressed oils with their refined versions by UV–Vis spectroscopy. *Scientific Reports* **2020,** *10* (1), 16100.

1-49. Torrecilla, J. S.; Rojo, E.; Domínguez, J. C.; Rodríguez, F., A novel method to quantify the adulteration of extra virgin olive oil with low-grade olive oils by UV−Vis. *Journal of Agricultural and Food Chemistry* **2010,** *58* (3), 1679-1684.

1-50. Didham, M.; Truong, V. K.; Chapman, J.; Cozzolino, D., Sensing the addition of vegetable oils to olive oil: The ability of UV–VIS and MIR spectroscopy coupled with chemometric analysis. *Food Analytical Methods* **2020,** *13* (3), 601-607.

1-51. El-Abassy, R.; Donfack, P.; Materny, A., Visible Raman spectroscopy for the discrimination of olive oils from different vegetable oils and the detection of adulteration. *Journal of Raman Spectroscopy: [An International Journal for Original Work in all Aspects of Raman Spectroscopy, Including Higher Order Processes, and also Brillouin and Rayleigh Scattering]* **2009,** *40* (9), 1284-1289.

1-52. Gurdeniz, G.; Ozen, B., Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data. *Food Chemistry* **2009**, *116* (2), 519-525.

1-53. López-Díez, E. C.; Bianchi, G.; Goodacre, R., Rapid quantitative assessment of the adulteration of virgin olive oils with hazelnut oils using Raman spectroscopy and chemometrics. *Journal of agricultural and food chemistry* **2003**, *51* (21), 6145-6150.

1-54. Philippidis, A.; Poulakis, E.; Papadaki, A.; Velegrakis, M., Comparative study using Raman and visible spectroscopy of cretan extra virgin olive oil adulteration with sunflower oil. *Analytical Letters* **2017**, *50* (7), 1182-1195.

1-55. Rohman, A.; Che Man, Y.; Yusof, F. M., The use of FTIR spectroscopy and chemometrics for rapid authentication of extra virgin olive oil. *Journal of the American Oil Chemists' Society* **2014**, *91* (2), 207-213.

1-56. Vlachos, N.; Skopelitis, Y.; Psaroudaki, M.; Konstantinidou, V.; Chatzilazarou, A.; Tegou, E., Applications of Fourier transform-infrared spectroscopy to edible oils. *Analytica chimica acta* **2006**, *573*, 459-465.

1-57. Oussama, A.; Elabadi, F.; Platikanov, S.; Kzaiber, F.; Tauler, R., Detection of olive oil adulteration using FT-IR spectroscopy and PLS with variable importance of projection (VIP) Scores. *Journal of the American Oil Chemists' Society* **2012**, *89* (10), 1807-1812.

1-58. Yang, H.; Irudayaraj, J., Comparison of near-infrared, fourier transform-infrared, and fourier transform-raman methods for determining olive pomace

oil adulteration in extra virgin olive oil. *Journal of the American Oil Chemists' Society* **2001**, *78* (9), 889.

1-59. Lapcharoensuk, R.; Malithong, A.; Thappho, D.; Phonpho, P., Discrimination of vegetable oil types using Fourier transforms near infrared spectroscopy coupled with pattern recognition techniques. *IOP Conference Series: Earth and Environmental Science* **2019**, *301* (1), 012067.

1-60. Chen, B.; Tian, P.; Lu, D.-L.; Zhou, Z.-Q.; Shao, M.-L., Feasibility study of discriminating edible vegetable oils by 2D-NIR. *Analytical Methods* **2012**, *4* (12), 4310-4315.

1-61. Marigheto, N.; Kemsley, E.; Defernez, M.; Wilson, R., A comparison of mid-infrared and Raman spectroscopies for the authentication of edible oils. *Journal of the American Oil Chemists' Society* **1998**, *75* (8), 987-992.

1-62. Galtier, O.; Abbas, O.; Le Dréau, Y.; Rebufa, C.; Kister, J.; Artaud, J.; Dupuy, N., Comparison of PLS1-DA, PLS2-DA and SIMCA for classification by origin of crude petroleum oils by MIR and virgin olive oils by NIR for different spectral regions. *Vibrational Spectroscopy* **2011**, *55* (1), 132-140.

1-63. Samyn, P.; Van Nieuwkerke, D.; Schoukens, G.; Vonck, L.; Stanssens, D.; Van Den Aabbeele, H., Quality and statistical classification of Brazilian vegetable oils using mid-infrared and Raman spectroscopy. *Applied Spectroscopy* **2012**, *66* (5), 552-565.

1-64. Gurdeniz, G.; Ozen, B., Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data. *Food Chemistry* **2009**, *116* (2), 519-525.

1-65. Baeten, V.; Fernández Pierna, J. A.; Dardenne, P.; Meurens, M.; García-González, D. L.; Aparicio-Ruiz, R., Detection of the Presence of hazelnut oil in olive oil by FT-Raman and FT-MIR Spectroscopy. *Journal of Agricultural and Food Chemistry* **2005**, *53* (16), 6201-6206.

1-66. Baeten, V.; Hourant, P.; Morales, M.; Aparicio, R., Oil and fat classification by FT-Raman spectroscopy. *Journal of Agricultural and Food Chemistry (USA)* **1998**.

1-67. Guzmán, E.; Baeten, V.; Pierna, J. A. F.; García-Mesa, J. A., Application of low-resolution Raman spectroscopy for the analysis of oxidized olive oil. *Food Control* **2011**, *22* (12), 2036-2040.

1-68. Korifi, R.; Le Dréau, Y.; Molinet, J.; Artaud, J.; Dupuy, N., Composition and authentication of virgin olive oil from French PDO regions by chemometric treatment of Raman spectra. *Journal of Raman Spectroscopy* **2011**, *42* (7), 1540-1547.

1-69. Zhang, X.; Qi, X.; Zou, M.; Liu, F., Rapid authentication of olive oil by Raman spectroscopy using principal component analysis. *Analytical Letters* **2011**, *44* (12), 2209-2220.

1-70. Kwofie, F.; Lavine, B. K.; Ottaway, J.; Booksh, K., Incorporating brand variability into classification of edible oils by Raman spectroscopy. *Journal of Chemometrics* **2020**, *34* (7), e3173.

1-71. Jin, H.; Li, H.; Yin, Z.; Zhu, Y.; Lu, A.; Zhao, D.; Li, C., Application of Raman spectroscopy in the rapid detection of waste cooking oil. *Food Chemistry* **2021**, *362*, 130191.

CHAPTER II

Materials and Methods

2.1. Collection of Samples

Ninety-seven edible oil samples (see Table 2.1) from multiple brands spanning twenty distinct varieties of plant based edible oils were purchased over a three-year period from supermarkets in the Newark, DE metropolitan area to account for seasonal and batch variations within each supplier as well as the variation between suppliers. All edible oil samples used in this study were stored in glass containers with plastic caps at room temperature prior to infrared analysis**.** Binary mixtures of the edible oils were prepared to simulate adulterated edible oils by mixing a more expensive edible oil (e.g., Extra Virgin olive oil) with a less expensive oil (e.g., corn oil) by mixing the appropriate volumes using a digital pipette (Eppendorf). For example, an eighty-five percent/fifteen percent mixture of extra virgin olive oil and corn oil was prepared by mixing 850 μL of extra virgin olive oil and 150 μL of corn oil in a 1.5 mL sterile falcon tube using a MaxiMixPlus vortex mixer (Thermolyne, East Lyme, CT).

The peroxide value of each edible oil sample was measured using a Milwaukee Lab Mi490 Photometer (Milwaukee Instruments, Rocky Mount, NC). Prior to photometric analysis, each sample was diluted with acetic acid and n-hexane (3:2). For this measurement, 0.2 mL of the edible oil sample was mixed with 0.8 mL of acetic acid/*n*-hexane mixture and then added to a vial containing the Mi490A-0 reagent. The photometer was standardized (zeroed) using this mixture. A packet of reagent Mi490B-0 (white powder) was then added to this vial followed by vigorous shaking for one minute. The vial after shaking was inserted into the instrument. After five minutes, the peroxide value of the sample was obtained in milli-equivalents of O_2/Kg.

Table 2.1. Edible Oil Samples

Edible Oil Type	p	Number of Samples
EVOO	A	26
ELOO	A	7
Olive	A	8
Avocado	A	2
Peanut	A	4
Sweet Almond	A	2
Almond	A	4
Safflower	A	2
Hazelnut	A	2
Avocado-Olive-Flaxseed	A	1
Sunflower	A	1
Canola	B	9
Canola-Vegetable	B	1
Extra Virgin Sesame	B	3
Toasted Sesame	B	1
Canola-Sunflower-Soybean	B	1
Corn	C	9
Grapeseed	C	8
Vegetable	C	4
Walnut	D	2
Total		97

2.2. Fourier Transform Infrared Spectroscopy

FTIR absorbance spectra (4000 cm^{-1} to 400 cm^{-1}) of the commercial edible oils (see Table 2.1) and their binary mixtures were measured in triplicate or quadruplicate (64 scans each) using an iS50 FTIR spectrometer (Thermo-Nicolet**,** Madison, WI) equipped with a deuterated triglyceride sulfate (DTGS) detector. Although most IR spectral libraries utilize transmission spectra, attenuated total reflection (ATR) [2-1] was the technique used to collect IR spectra of the edible oils for the two studies that are highlighted in this dissertation. ATR requires little or no sample preparation and consistent results (as in the case of transmission) are usually obtained. For liquid samples, the technique requires minimal training. Furthermore, samples can be measured in their neat state and dilution is not required in order to obtain a spectrum. As some samples may change their spectral characteristics after grinding, melting or pressing, ATR is an excellent alternative to the use of KBr disks or mineral oil mulls. For these reasons ATR has become the most widely used sampling method in IR spectroscopy.

In ATR, the IR beam traverses the internal reflecting element (IRE) (which is a diamond crystal in the case of the iS50 FTIR spectrometer) at an incident angle greater than the critical angle. This causes internal reflection to occur, with an evanescent wave generated that penetrates into the sample (see Figure 2.1). In the mid-IR region where the edible oil absorbs IR radiation, the evanescent wave from the IRE is attenuated, and the reflected beam is the basis of the IR spectrum of the sample. The depth of penetration (d_p) of the evanescent wave[2-2] is given by Equation 2.1 where λ is the wavelength of light, θ is the angle of incidence of the light, n_1 is the refractive index of the IRE and n_2 is the refractive index of the sample. For internal reflection to occur, the refractive index of the

IRE must be greater than the refractive index of the sample. (Materials used as internal reflection elements in ATR accessories such as diamond (n = 2.41) or germanium (n = 4.00) have a higher refractive index that an edible oil whose refractive index can vary from 1.4 to 1.5.)

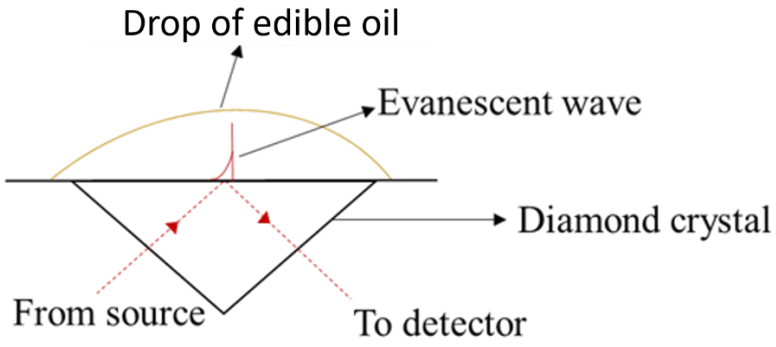

Figure 2.1. Internal reflection at a diamond crystal in an ATR accessory. The evanescent wave penetrates into a drop of edible oil placed on the crystal.

$$d_p = \frac{\lambda}{2\pi(n_1^2\, sin^2\theta - n_2^2)^{0.5}} \qquad 2.1$$

The penetration depth of the evanescent wave is directly proportional to the wavelength of the IR radiation and is independent of sample thickness. Good and uniform contact between the sample and crystal is crucial as the value of the absorbance is affected by non-homogenized contact. For edible oils, good contact typically occurred with the crystal except in only a few cases where bubbles were observed after dropping a sample of the edible oil onto the ATR crystal. It was a simple matter to eliminate the bubbles, thereby ensuring good contact between the sample and the crystal.

To collect an IR spectrum, the crystal was first cleaned with a cotton swab doused in isopropanol. The spectral background was obtained after the isopropanol on the crystal evaporated to dryness. A drop of the oil was then placed on the ATR crystal and 64 scans were collected for each sample at 4 cm^{-1} resolution. The crystal was always cleaned with a cotton swab and isopropanol after each sample measurement to avoid cross contamination by the previous sample during collection of the spectra. The spectral region corresponding to the absorption of the diamond crystal (~2300 cm^{-1}- 1900 cm^{-1}) was zeroed out for all IR spectra collected for this project.

2.3. Pattern Recognition Analysis of IR Spectra

FTIR and Raman spectroscopy differ from chromatographic techniques in their approach for authenticating edible oils. In contrast to chromatography which tends to isolate the components of an edible oil prior to analysis, vibrational spectroscopic techniques treat the IR or Raman spectrum as a chemical "fingerprint" of the sample. Objective analysis of these spectral profiles depends upon the use of multivariate statistical methods. Pattern recognition methods [2-3 - 2-5] were selected for analyzing the IR spectra of the edible oils because of the attributes of the procedures. First, there are methods available that assume no mathematical model but rather seek relationships that provide definitions of similarity between groups of data. Second, pattern recognition methods are able to deal with high dimensional data where more than three measurements are used to describe each sample. Finally, techniques are available for selecting important features from a large set of measurements. Thus, studies can be performed on data sets where the exact relationships are not fully understood.

For pattern recognition analysis, each IR spectrum is represented by a data vector $x = (x_1, x_2, x_3, ….. x_j, …….x_p)$ where component x_j is the value of the j^{th} descriptor. Such a vector can be considered as a point in a high dimensional measurement space. The Euclidean distance between a pair of points in the measurement space is inversely related to the degree of similarity between samples (as represented by their spectra). Points representing samples from one class (e.g., a specific variety of an edible oil) will cluster in a limited region of the measurement space. Pattern recognition is a set of numerical methods for assessing the structure of the data, which is defined as the overall relation of each sample to every other sample in the data set.

What are the operations that must be performed in order to apply pattern recognition techniques to the authentication of edible oils? A summary of the techniques used in the studies presented in this dissertation will be discussed in the following sections. Emphasis is placed on how the data is configured for pattern recognition.

2.3.1. Data Representation

The first step in a pattern recognition study is to convert the original spectral data into computer compatible form, which is a string of scalar measurements comprising an n-tuple called a pattern vector. Each component of the pattern vector is the absorbance value at a specific wavelength (cm^{-1}). The pattern vectors, in turn, are arranged in the form of a data matrix (see Equation 2.2); the rows of the matrix represent the spectra and the columns of the matrix represent the absorbance value at a specific wavelength for the spectra. It is essential that each column encodes the same information for all samples in the data set. If

the second column is the absorbance value at 3950 cm^{-1} for sample one, it must also be the absorbance value at 3950 cm^{-1} for sample two, three ….. N.

$$\begin{bmatrix} X_{11} & X_{12} & X_{13} & \dots & X_{1P} \\ X_{21} & X_{22} & X_{23} & \dots & X_{2P} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{N1} & X_{N2} & X_{N3} & \dots & X_{NP} \end{bmatrix} \qquad (2.2)$$

2.3.2. Data Preprocessing

The next step involves scaling of the data. The scaling procedures used for a given data set depend upon the attributes of the data, the pattern recognition method used and the nature of the problem investigated. This aspect of pattern recognition has not been thoroughly investigated for IR spectral data. In the studies discussed herein, two techniques were used: normalization and autoscaling.

Normalization involves setting the sum of the squares of the components of each pattern vector equal to unity. Normalization is performed on each row of the data matrix to compensate for variations in the optical path length of the spectra. Autoscaling involves adjusting the measurements such that each has a mean of zero and a standard deviation of one (see Equation 2.3) where $\bar{x}_{i,orig}$ is the mean and $s_{i,orig}$ is the standard deviation of the original measurement variable. Autoscaling is performed on each column in the data matrix and removes any inadvertent weighing of the variables that otherwise would occur due to differences in the magnitude among the measurements. After autoscaling, all measurements have equal weight and therefore an equal effect on the analysis.

$$x_{i,new} = \frac{(x_{i,orig} - \bar{x}_{i,orig})}{s_{i,orig}} \tag{2.3}$$

2.3.3. Looking at Multidimensional Measurement Space

Graphical methods are often used by physical scientists to study data. If there are only two or three measurements per sample, the data can be displayed as points in a two- or three-dimensional measurement space. The coordinate axes of this space are defined by the measurement variables. By examining the plot, a scientist can search for similarities and dissimilarities among samples, find natural clusters and even gain information about the overall structure of the data. If there are p measurements per sample ($p > 3$), a two or three-dimensional representation of the measurement space is needed that faithfully reflects the relative position of the points in the high dimensional measurement space. The approach that is taken to solve this problem in the studies discussed herein involves using a technique called principal component analysis[2-6-2-8](PCA). This technique can be summarized as a method for transforming the original measurement variables into new uncorrelated variables called principal components. Each principal component is a linear combination of the original measurement variables. Using this procedure is analogous to finding a set of orthogonal axes that represent the directions of greatest variance in the data.

A measure of the amount of information conveyed by each principal component is its variance. (The variance is defined as the degree to which the data points are spread apart in the p-dimensional measurement space.) For this reason, the principal components are usually arranged in order of decreasing variance. Thus, the most informative principal component is the first and the least informative is the last. Often, only the first two or three

principal components are used to generate a plot representing the p-dimensional pattern space. For data sets with a large number of interrelated measurement variables, PCA is a powerful method for analyzing the structure of the data and reducing the dimensionality of the data.

PCA is performed via a decomposition of the data matrix **X** (n x p) into a score matrix **T** (n x F), a loading matrix **P** (F x p), and a residual matrix **E** (n x p), where n is the number of spectra (i.e., samples) in the data set, p is the number of measurement variables, and F is the number of principal components necessary to represent a specified fraction of the total cumulative variance in the data. Usually, F is smaller than p due to correlations among the measurement variables. The matrix equation to decompose X is

$$\mathbf{X} = (1 \times m) + \mathbf{TP} + \mathbf{E} \qquad (2.4)$$

where 1 is a column vector (n x 1) of ones and m is a (1 x p) row vector representing the mean of the samples. The coordinates of the samples (spectra) in the principal component space are provided by the score matrix, whereas the loading matrix contains the necessary information for transforming the original measurement variables into principal components. By plotting the columns of T against each other, a plot representing the distribution of the data points in the p dimensional space is obtained. The number of principal components needed to describe the signal in the data is equal to F or the number of columns in T, which in many studies is only two or three. The score and loading matrices describe the signal in the data, whereas the residual matrix describes the noise. Hence,

dimensionality reduction and separation of signal from noise in the data matrix is possible using PCA.

2.3.4. Cluster Analysis

Exploratory data analysis techniques are often quite helpful in understanding the complex nature of multivariate relationships. The importance of mapping and display techniques such as PCA to understand the structure of a complex multivariate data set was discussed in the preceding section. Cluster analysis, which attempts to determine the structural characteristics of a data set by organizing the data into subgroups or clusters, is discussed in this section. These methods are based on the principle that distances between pairs of points in the measurement space are inversely related to their degree of similarity.

Although several different types of clustering algorithms exist, hierarchical clustering[2-9 -2-11] is by far the most popular. The starting point for hierarchical clustering is the similarity matrix, which is formed from the data matrix by computing the distances between all pairs of points in the data set. Each distance is then converted into a similarity value (see Equation 2.5) where s_{ik} is the measure of similarity between samples i and k, d_{ik} is the Euclidean distance between samples i and k, and d_{max} is the distance between the two most dissimilar samples which is also the largest distance between points in the data set. The similarity values, which vary between 0 and 1, are organized in the form of a square matrix called a similarity matrix.

$$s_{A,B} = 1 - \frac{d_{A,B}}{d_{max}} \qquad (2.5)$$

The similarity matrix is scanned for the largest value, which corresponds to the two samples that are most similar, and the two samples comprising these points are combined to form a new point located midway between the two original points. After the rows and columns corresponding to the original two data points are removed, the similarity matrix is updated to include information about the similarity between the new point and the other remaining points in the data set. The similarity matrix is again scanned, the new nearest point is again identified and combined to form a single point, the rows and columns of the two data points that were combined are removed, the matrix is again recomputed to include information about the similarity between the new point and every other data point remaining. The reduction of the similarity matrix by the aggregation of the samples is repeated until all points have been linked. The results of this procedure are summarized in a diagram called a dendrogram, which is a visual representation of the relationships between samples in the data set. Interpretation of the results is intuitive (see Chapter 3), which is the major reason for the popularity of these methods.

All clustering procedures yield the same results with well separated clusters. However, the results differ when the clusters overlap because of space distorting effects. Single linkage hierarchical clustering favors the formation of large linear clusters instead of the usual elliptical or spherical clusters. As a result, poorly separated clusters are often chained together. Complete linkage hierarchical clustering or Wards method favors the formation of small spherical clusters. For this reason, it is a good idea to use at least two different clustering methods when studying a data set. If the results agree, a strong case can be made for partitioning the data into distinct sample groups. If the clustering memberships differ, the data should be investigated further using PCA. As a general rule,

it is recommended that hierarchical methods be used in tandem with PCA to detect clusters in multivariate data. Hierarchical methods of clustering are exploratory tools. The absolute validity of a dendrogram is less important than the insights and suggestions gained by the use about the structure of the data.

2.3.5. Classification

The overall goal of a pattern recognition study is to solve the class membership problem. In most pattern recognition studies, samples are classified according to a specific property using measurements that are indirectly related to that property. An empirical relationship or classification rule is developed from a set of samples for which the property of interest and the measurements are known. The classification rule is then used to predict this property in samples that are not part of the original training set. The property in question (for example) is the specific variety of an edible oil and the measurements are the mid-infrared absorbance values of the edible oil sample at specific wavelengths.

Although PCA and hierarchical clustering are powerful methods for analyzing the structure of a data set, they are not sufficient for developing a classification rule. In the final section of this chapter, a genetic algorithm (GA) for pattern recognition[2-12-2-15] which combines variable selection and classification in a single step is discussed. The pattern recognition GA identifies features that optimize the separation of the classes in a plot of the two largest principal components of the data. Because the largest principal components capture the bulk of the variance in the data, the wavelengths selected by the GA convey information primarily about differences between the classes in the data set. Hence, the principal component analysis routine embedded in the fitness function of the pattern

recognition GA acts as an information filter, significantly reducing the size of the search space as it restricts the search to wavelengths whose principal component score plots show clustering on the basis of class. In addition, the algorithm is able to focus on those classes and or samples that are difficult to classify as it trains by boosting the class and sample weights. Samples that consistently classify correctly are not as heavily weighted as samples that are difficult to classify. Over time, the algorithm learns its optimal parameters in a manner similar to a neural network. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a "smart" one -pass procedure for variable selection and classification.

A block diagram of the pattern recognition GA is shown in Figure 2.2. The GA builds a population of chromosomes, each of which represents a potential solution, i.e., a set of wavelengths. During each generation, the chromosomes are assigned a value by the fitness function, which is a measure of the quality of the proposed solution for the classification problem. Solutions with a high fitness value have a higher probability of being selected for cross over than chromosomes with a low fitness value. The power of the GA arises from crossover, which causes a structured yet randomized exchange of information between potential solutions (chromosomes) with the prospect that good solutions can generate even better ones. The new population of chromosomes, which often yields better solutions to the problem, is again evaluated using the fitness function. This entire process (evaluation, selection, crossover, mutation, and adjustment of internal parameters) is repeated until convergence is achieved or a specified number of generations has been found.

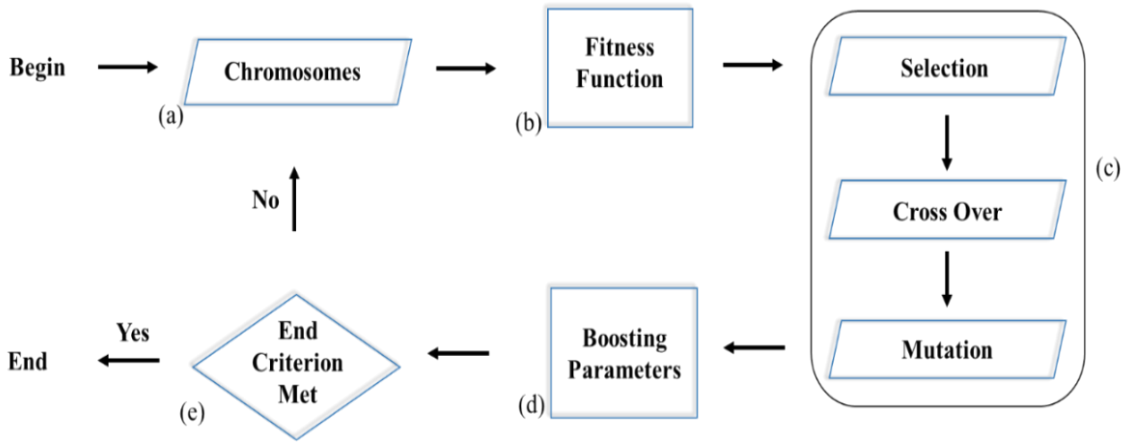

Figure 2.2. Block diagram of the pattern recognition GA.

The fitness function of the pattern recognition GA scores the principal component plots associated with each wavelength subset (chromosome) to identify a set of wavelengths that optimize the separation of the sample classes in a plot of the two largest principal components of the data. To facilitate the tracking and scoring of the principal component plots, class and sample weights, which are part of the fitness function, are computed, see Equations 2.6 and 2.7 where CW(c) is the weight of class c and $SW_c(s)$ is the weight of sample s in class c. The class weights sum to 100, and the sample weights in a class sum to a value equal to the corresponding class weight.

$$CW(c) = 100 \frac{CW(c)}{\sum_{c} CW(c)} \tag{2.6}$$

$$SW_c(s) = CW(c) \frac{SW_c(s)}{\sum_{s \in c} SW_c(s)} \tag{2.7}$$

Each principal component plot generated for each feature subset (chromosome) is scored using the K-nearest neighbor (K-NN) classification algorithm.[2-16] For a given data point, the Euclidean distance is computed between it and every other point in the principal component plot. These distances are arranged from the smallest to the largest. A poll is then taken of the point's k-nearest neighbors. For the most rigorous classification of the data, k equals the number of samples in the class to which the point belongs. (k, which is assigned by the user, varies with the class.) The number of k-nearest neighbors with the same class label as the sample point in question, the so-called sample hit count (SHC) is computed ($0 < SHC(s) < K_c$) where K_c is the number of nearest neighbors calculated for each sample in class c.) It is a simple matter to score a principal component plot, see Equation 2.8 where F(d) is the fitness function of the feature set scored, SHC(s) is the number of nearest points (samples) with the same class label as sample s and SW (s) is the weight of the sample s.

$$\sum_{c}\sum_{s\in c}\frac{1}{K_c}\times SHC(s)\times SW(s) \qquad (2.8)$$

To understand the scoring of each principal component plot by the fitness function, consider a hypothetical data set consisting of thirty samples distributed between two classes that have been initially assigned equal weights. Since class 1 has 10 samples and class 2 has 20 samples, K_1 is 10 and K_2 is 20. At generation 0 (reproduction has not yet occurred), the samples in a given class will have the same weights. Therefore, each sample in class 1 has a sample weight of 5, whereas each sample in class 2 has a weight of 2.5. If sample 3, which is in class 1, has as its nearest neighbors 7 class one samples, then $SHC/K_1 = 0.7$ and

$(SHC/K_1)*SW(3) = 0.7*5$ or 3.5. By summing up $(SHC/K_c)*SW(s)$ for each sample point in the plot, the principal component plot of the feature subset (wavelengths) is scored.

The fitness function of the pattern recognition GA is able to focus on those classes and samples that are difficult to classify by boosting the values of their weights over successive generations. In order to boost, it is necessary to first compute the sample hit rate (SHR), which is the mean value of SHC/K_c over all wavelength subsets (ϕ) produced in a generation (see Equation 2.9). SHR is calculated over the entire population of solutions in a particular generation and provides information about the difficulty in classifying a particular sample.

$$\mathrm{SHR(s)} = \frac{1}{\phi}\sum_{i=1}^{\phi}\frac{\mathrm{SHC_i}(s)}{\mathrm{K}} \tag{2.9}$$

Boosting is a two-step process. First, the class hit, which is the average sample hit rate for all samples in a class, is computed, see Equation 2.10 where CHR_g (c) is the class hit rate for class c during generation g, AVG is the average and SHR_g (s) is the sample hit rate for sample s in class c during generation g.

$$\mathrm{CHR_g(c)} = AVG(\mathrm{SHR_g(s)} : \forall_{s\in c}) \tag{2.10}$$

Second, the class and sample weights are then adjusted during each generation using a perceptron, see Equations 2.11 and 2.12 where $CW_{g+1}(c)$ is the class weight for class c during the current generation g+1, $CW_g(c)$ is the class weight for class c during the previous generation g, P is the momentum, $CHR_g(c)$ is the class hit rate for class c during generation g, $SW(s)_{g+1}$ is the sample weight for sample s during generation g+1, $SW(s)_g$ is

the sample weight for sample s during the previous generation g, and $SHR_g(s)$ is the sample hit rate for sample s during the previous generation.

$$CW_{g+1}(c) = CW_g(c) + P(1 - CHR_g(c)) \tag{2.11}$$

$$SW_{g+1}(s) = SW_g(s) + P(1 - SHR_g(s)) \tag{2.12}$$

During each generation, the class and sample weights are updated using the class and sample hit-rates from the previous generation. After a certain number of generations, the class weights become fixed. Equation 2.11 is disabled, P is halved and the sample weights are renormalized using Equation 2.7. The pattern recognition GA then focuses on the troublesome samples (see Equation 2.12).

Boosting is crucial to the successful operation of the pattern recognition GA as it allows the values of both the class and sample weights to change as the GA is training, thereby modifying the criteria used by the fitness function for a good score. This helps to minimize the problem of convergence to a local optimum. Hence, the fitness function of the pattern recognition GA changes as the population evolves towards a solution.

During each generation, the selection, crossover, and mutation operators are applied to the chromosomes to develop new and potentially better solutions (more informative wavelengths) to the class membership problem. The selection operator used by the pattern recognition GA is implemented by ordering the chromosomes, i.e., wavelength subsets, from best to worse fitness while simultaneously generating a copy of the same population and randomizing the order with respect to fitness. A fraction of the population is then selected as per the selection pressure, which is usually set at 0.5. The

top half of the ordered population is mated with the top half of the random population, guaranteeing that the best 50% are selected for reproduction, while ensuring that every chromosome in the randomized copy has an equal chance of being selected due to the randomization criteria imposed on the chromosomes in this population.

The resulting population of chromosomes, both parents and children, are sorted by their fitness (see Equation 2.8) and the top ϕ chromosomes are retained for the next generation. The new population can be expected to perform on average better than its predecessor because of the selection criteria used for the higher-ranking strings. However, the reproduction operator also assures a significant degree of diversity in the population ensuring that convergence to a local optimum does not occur.

References

2-1. Larkin, P., Instrumentation and sampling methods in infrared and Raman spectroscopy: Principles and spectral interpretation. P. Larkin (Ed), Elsevier: 2011.

2-2. Griffiths, P. R. and de Haseth, J. D. Fourier transform infrared spectrometry, John Wiley & Sons. Inc., New York (*529 pp*) **2007**.

2-3. McLachlan, G., Discriminant Analysis and Statistical Pattern Recognition, John Wiley & Sons. *Inc., New York (526 pp)* **1992**.

2-4. Brereton, R. G., *Multivariate pattern recognition in chemometrics: illustrated by case studies*. Elsevier: 1992.

2-5. Varmuza, K. *Pattern recognition in chemistry*, Springer-Verlag: 1980.

2-6. Brown, S. D., Chemical systems under indirect observation: Latent properties and chemometrics. *Applied Spectroscopy* **1995,** *49* (12), 14A-31A.

2-7. Jolliffe, I. T., Principal component analysis. *Technometrics* **2003,** *45* (3), 276.

2-8. Haswell, S., *Practical guide to chemometrics*. CRC Press: 1992.

2-9. Massart, D.; Kaufman, L., The interpretation of analytical chemical data by the use of cluster analysis. John Wiley & Sons: 1983.

2-10. Murtagh, F., A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal* **1983,** *26* (4), 354-359.

2-11. Govender, P.; Sivakumar, V., Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research* **2020,** *11* (1), 40-56.

2-12. Lavine, B. K.; Davidson, C. E.; Breneman, C.; Katt, W., Electronic van der Waals Surface Property Descriptors and Genetic Algorithms for Developing Structure− Activity Correlations in Olfactory Databases. *Journal of Chemical Information and Computer Sciences* **2003,** *43* (6), 1890-1905.

2-13. Lavine, B. K.; Davidson, C.; Moores, A. J., Innovative genetic algorithms for chemoinformatics. *Chemometrics and Intelligent Laboratory Systems* **2002,** *60* (1-2), 161-171.

2-14. Lavine, B. K.; Davidson, C.; Moores, A. J.; Griffiths, P., Raman spectroscopy and genetic algorithms for the classification of wood types. *Applied Spectroscopy* **2001,** *55* (8), 960-966.

2-15. Lavine, B. K.; Davidson, C.; Moores, A. J., Genetic algorithms for spectral pattern recognition. *Vibrational Spectroscopy* **2002,** *28* (1), 83-95.

2-16. James, M. *Classification algorithms*. Wiley-Interscience: 1985.

CHAPTER III

Classification and Adulteration of Edible Oils

3.1. Introduction

Edible oils are food substances other than dairy products that are manufactured from fats and oils.[3-1] Edible oils are composed of triglycerides, which are esters formed between glycerol and saturated, mono-unsaturated or poly-unsaturated fatty acids. Edible oils play an important role in the human diet as they add flavor and color to foods and provide health benefits due to the presence of antioxidants such as tocopherol, phenolic compounds and phytosterol that assists in the metabolism of fat-soluble vitamins.[3-2] In addition, edible oils also supply omega-3 and omega- 6 fatty acids, which are not synthesized by the human body, but are essential to brain function.[3-3]

Adulteration, the addition of inferior or foreign substances into food, is of great concern to the food industry as these substances may pose a serious health risk to consumers and adversely impact product quality. Extra virgin olive oil (EVOO), in particular, has been subject to adulteration using less expensive edible oils such as rapeseed oil, soy oil, and canola oil.[3-4 – 3-5] In the southern region of the European Union (EU), the European commission has reported that the economic cost of EVOO adulteration by hazelnut oil is estimated to be 4 million euros per year.[3-6] Authentication of expensive

edible oils, such as EVOO, is crucial for maintaining both the quality and safety of food, and protecting consumers from the sale of fraudulent products.

Separation techniques such as capillary column gas chromatography/mass spectrometry and liquid chromatography/mass spectrometry have been used successfully to authenticate edible oils.[3-7 – 3-8] However, these methods are time consuming, labor intensive, expensive and are restricted to large laboratories. In contrast, Fourier transform infrared (FTIR) spectroscopy is fast, does not require sample preparation, and can be applied (on-line) to monitor the quality of edible oils such as EVOO during its production. Although infrared (IR) spectra of edible oils contain hundreds of compounds with overlapping bands that contribute to the complexity of the spectra, an IR spectrum can serve as a chemical fingerprint of an edible oil.[3-9] Analysis of these fingerprints by pattern recognition methods is crucial to ensure the effective extraction of qualitative and quantitative information necessary to verify the authenticity of an edible oil and to detect adulteration. The combination of FTIR spectroscopy and multivariate analysis can discriminate chemically similar edible oils and to model mixtures of edible oils.[3-10 – 3-13] In these studies, principal component analysis[3-14] has been demonstrated to be an effective tool to analyze FTIR spectra of edible oils for the presence of adulterants. Specifically, principal component analysis (PCA) and related methods have been used to cluster known edible oil samples and to classify unknown oil samples with similar properties.[3-15 – 3-17]

The application of FTIR spectroscopy and pattern recognition methods to the problem of discriminating edible oils by type (e.g., canola oil versus corn oil) is the focus of this study. Previously published studies on discrimination of edible oils by FTIR spectroscopy have been limited to approximately thirty samples spanning five or six

distinct varieties of edible oils obtained from a single manufacturer over a production year range of less than one year.[3-18 – 3-26] In this study, FTIR spectra obtained from ninety-seven samples spanning twenty distinct varieties (types) of edible oils (collected over a three-year period encompassing multiple brands representing supplier to supplier variation as well as seasonal and batch variation within a supplier) were analyzed using the four main types of pattern recognition methodology: mapping and display, cluster analysis, variable selection, and classification. A genetic algorithm (GA) for variable selection[3-27 – 3-30] was applied to the FTIR spectra of the edible oils to discriminate these oils by type. Using a hierarchical classification scheme, the twenty commercial varieties of edible oils can be divided into four distinct groups. The nature of the clustering and the edible oils comprising each group are similar to the results obtained in two previous studies on the characterization of edible oils using Raman spectroscopy.[3-31, 3-32] Edible oils from different groups can be reliably discriminated from each other, whereas discrimination of the edible oils within the same group can be problematic. Adulteration of plant based edible oils by other edible oils from the same group (e.g., EVOO by almond oil) cannot be reliably detected by FTIR spectroscopy, whereas adulteration of the edible oils by other oils from different groups (e.g., EVOO adulterated by corn or canola oil) can be detected at concentration levels as low as 10% (v/v) which is consistent with the results obtained in previously published studies using PLS regression.[3-33, 3-34]

A unique aspect of this work is the incorporation of edible oils collected systematically over several years into the data cohort, which introduces a heretofore unseen variability in the chemical compositions of these oils. This work also demonstrates that previously published studies (which rely on a single source to represent each type of edible

oil – a single brand, often a single bottle from that brand, to represent each variety of the edible oils in the study) provide an overly optimistic estimate of the capability of FTIR spectroscopy to classify plant based edible oils by variety and to detect the presence of adulterants in these edible oils.

3.2. Materials and Methods

Ninety-seven edible oil samples (see Table 3.1) from multiple brands spanning twenty distinct varieties of plant based edible oils were purchased over three years from supermarkets in Newark, DE to account for seasonal and batch variations within each supplier as well as variations between suppliers. Binary mixtures of edible oils to simulate adulteration were prepared by mixing a more expensive edible oil (e.g., extra virgin olive oil) with a less expensive one (e.g., corn oil) in varying amounts using a digital pipette (Eppendorf). FTIR absorbance spectra (4000 cm^{-1} to 400 cm^{-1}) of the commercial edible oils (see Tables 3.1 and 3.2) and their binary mixtures (see Table 3.3) were measured in triplicate or quadruplicate (64 scans each) using an iS50 FTIR spectrometer (Thermo-Nicolet, Madison, WI) equipped with a deuterated triglyceride sulfate (DTGS) detector. The FTIR spectrometer was operated in ATR mode using a diamond crystal.

The ATR spectrum of a corn oil sample is shown in Figure 3.1. This sample has intense absorption bands at 2922 cm^{-1} which is attributed to asymmetric $-C-H$ stretching of $-CH_2-$. Other intense absorption bands for this sample include symmetric $-C-H$ stretching of $-CH_2-$ at 2853 cm^{-1}, $-C=O$ stretching of ester at 1743 cm^{-1} and $-C-O$ stretching and $-CH_2-$ bending at 1160 cm^{-1}. The spectral region between 2200 and 2000 cm^{-1} which corresponds to the absorbance by the diamond crystal was excluded from the

analysis as no meaningful information about edible oils is contained in this region. Other bands observed in the spectrum include a shoulder at 2956 cm^{-1} (-CH_2 stretching of alkane), a peak at 3009 cm^{-1} (-CH stretching of cis alkene) and another at 1657 cm^{-1} (alkene double bond of cis olefins). The peaks for -CH_2 bending appear at 1378 cm^{-1}, 1241 cm^{-1} and 1160 cm^{-1}. The peaks at 1122 cm^{-1} and 1102 cm^{-1} are attributed to C-O stretching of an ester.

Table 3.1. Composition of the training set of the pure edible oils

Edible Oil Type	Number of Samples	Number of Spectra
EVOO	26	68
ELOO	7	19
Olive	8	21
Avocado	2	5
Peanut	4	10
Sweet Almond	2	5
Almond	4	11
Safflower	2	6
Hazelnut	2	6
Avocado-Olive-Flaxseed	1	3
Sunflower	1	3
Canola	9	27
Canola-Vegetable	1	3
Extra Virgin Sesame	3	8
Toasted Sesame	1	3
Canola-Sunflower-Soybean	1	8
Corn	9	29
Grapeseed	8	22
Vegetable	4	10
Walnut	2	6
Total	97	273

Table 3.2. Composition of the prediction set of the pure edible oils

Edible Oil Type	Number of Samples	Number of Spectra
EVOO	7	11
ELOO	1	1
Olive	3	5
Avocado	2	4
Peanut	3	5
Sweet Almond	1	1
Almond	2	4
Safflower	1	3
Hazelnut	1	3
Avocado-Olive-Flaxseed	0	0
Sunflower	1	0
Canola	3	9
Canola-Vegetable	0	0
Extra Virgin Sesame	2	7
Toasted Sesame	0	0
Canola-Sunflower-Soybean	1	1
Corn	3	10
Grapeseed	5	11
Vegetable	1	4
Walnut	2	4
Total	39	83

Table 3.3. Binary mixtures of edible oils

Oil Type	Adulterant	Number of Spectra
EVOO	Corn	36
EVOO	Canola	51
EVOO	Almond	30

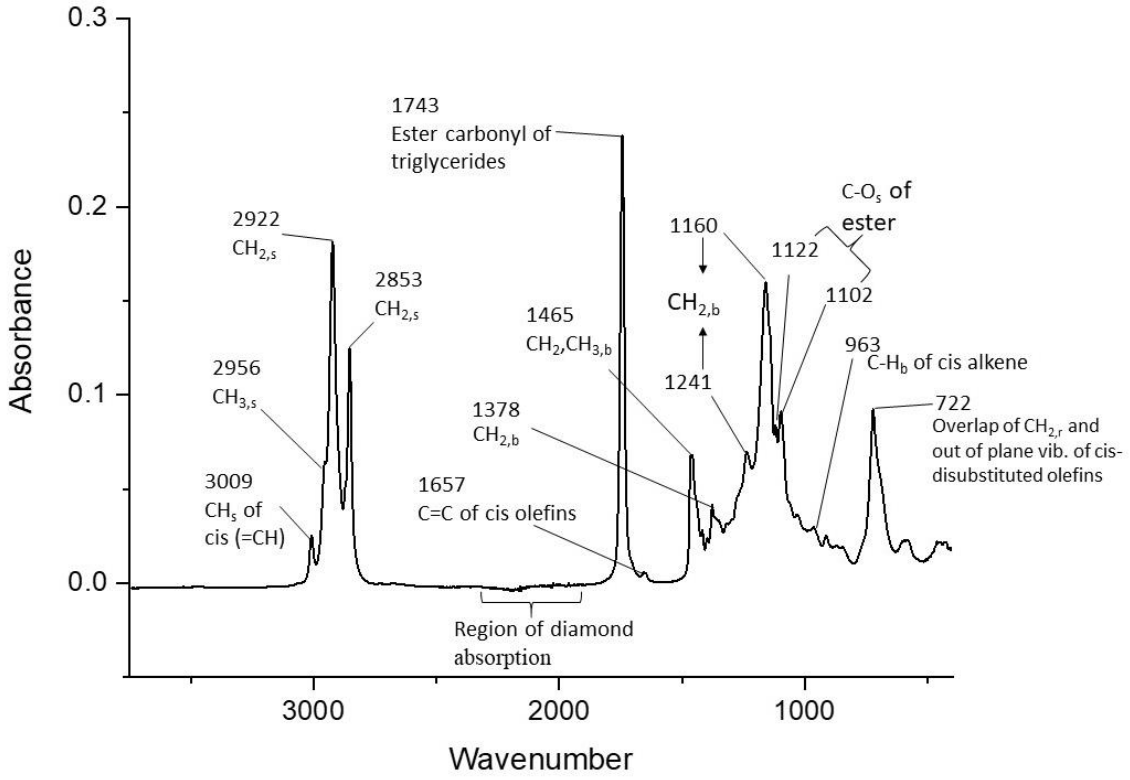

Figure 3.1. A representative FTIR spectrum of an edible oil sample (corn oil). Fundamental vibration frequencies are indicated.

3.2.1. Pattern Recognition Analysis

Each FTIR spectrum was baseline corrected using OMNIC (Thermo Nicolet, Madison, WI) and normalized to unit length with MATLAB (Math Works, Natick, MA). For pattern recognition analysis, each baseline corrected and normalized FTIR spectrum, which consisted of 6921 points, was represented as a data vector, $x = (x_1, x_2, x_3 \ldots x_{6921})$ where x_1 is the absorbance at 4000 cm^{-1} and x_{6921} is the absorbance at 400 cm^{-1}. Therefore, each IR spectrum can be considered as a point in a high dimensional measurement space. A basic assumption inherent in this study is that the distance between pairs of points in the measurement (pattern) space is inversely related to their degree of similarity. The points

representing FTIR spectra from one class (e.g., corn oil) will cluster in a limited region of this space distant from the points corresponding to the other class (e.g., extra virgin olive oil). Pattern recognition is a set of methods for investigating data represented in this manner to assess its general structure, which is the overall relation of each sample to every other sample in the data.

The IR spectra of the pure edible oils were divided into a training set of 273 spectra (see Table 3.1) and a prediction set of 83 spectra (see Table 3.2). Spectra comprising the prediction set were chosen by random lot. In this study, principal component analysis (PCA) was used to analyze the IR spectra comprising the training set. PCA is a method for transforming the original measurement variables into new, uncorrelated variables called principal components. Each principal component is a linear combination of the original measurement variables. Using this procedure is analogous to finding a set of orthogonal axes that represent the directions of greatest variance in the data. (The variance is defined as the degree to which the data points are scattered in the high dimensional measurement space.) Often, only two or three principal components are necessary to explain the information content of a data set when there are a large number of correlated measurement variables. By analyzing the spectra in the training set using principal component (PC) score plots, it is possible to identify classes (i.e., distinct sample groups) in the data and to detect the presence of outliers (i.e., discordant observations).

As outliers have the potential to adversely affect the performance of pattern recognition methods, outlier analysis was performed on each class (i.e., variety of edible oil) in the training set prior to the application of pattern recognition methods. Using PC score plots to identify discordant observations in the data, two samples (one corn oil and

the other grapeseed oil) were identified as outliers and excluded from the analysis. These two samples were found to have peroxide values (as determined using a Milwaukee Lab Mi490 Photometer) that far exceeded those of the other edible oil samples in the training set.

To identify wavelengths characteristic of each class (i.e., variety of edible oil), a genetic algorithm (GA) for pattern recognition analysis.[3-35, 3-36] was applied to the training set data. The pattern recognition GA identifies the smallest set of variables (i.e., absorbances at specific wavelengths) that optimize the separation of the classes in a plot of the two or three largest principal components of the data. Because principal components maximize variance, the bulk of the information encoded by these variables is about differences between the classes in the training set. With this approach to variable selection, an eigenvector projection of the data is formulated that discriminates between the classes in the data set by maximizing the ratio of between to within group variance through selection of the appropriate variables. Although a principal component score plot is not a sharp knife for discrimination, if a score plot shows clustering on the basis of the class membership of the samples, then our experience is that we will be able to predict robustly using this set of wavelengths.

For many types of chemical measurements, noise reduction and better class separation can be achieved when principal component analysis is used to characterize the information content of each variable subset selected by the pattern recognition GA. Furthermore, the problem of chance or spurious classification, which is always of concern when using a variable selection technique, is mitigated by the pattern recognition because of the more stringent criterion used for variable selection. Variables that contain

discriminatory information about a specific classification problem are often correlated which is why variable selection methods that utilize principal component analysis often work best.

To evaluate and compare each chromosome (i.e., variable subset), an object function must be applied to the data that quantifies the fitness of each chromosome comprising the population of potential solutions generated by the pattern recognition GA in each generation. The fitness (i.e., object) function used in this study is called PCKaNN [3-37] which incorporates PCA and the K-nearest neighbor (K-NN) classification algorithm [3-38] to score each variable (i.e., wavelength) subset in the population. A unique attribute of the PCKaNN fitness function is the incorporation of boosting using class and sample weights to modify the fitness function as the population is evolving towards a solution. Thus, the problem of convergence to a local optimum will be minimized. Evaluation and boosting as well as reproduction are repeated by the genetic algorithm for the population of chromosomes until a specified number of generations have been executed or a feasible solution has been found (i.e., a specific variable subset in the population that achieves a score of 100%). Further details about the pattern recognition GA and the PCKaNN fitness function can be found in Chapter 2.

3.3. Results and Discussion.

A hierarchical classification scheme was developed to discriminate the twenty varieties of edible oils investigated in this study. The twenty varieties were first divided into distinct groups, with the edible oils in each group then differentiated by type (i.e., variety). To implement this scheme, the average IR spectrum of each variety of edible oil was computed. The twenty average IR spectra were then analyzed using PCA and cluster

analysis. Both hierarchical clustering [3-39] (see Figure 3.2a) and PCA (see Figure 3.2b) yielded the same results - four distinct groups of edible oils depicted as Groups A, B, C and D in Figures 3.1a and 3.1b.

A visual comparison of the average IR spectra revealed that spectra from the same edible oil group were more similar to each other than spectra from different edible oil groups. Therefore, the 97 edible oil samples were divided into four distinct groups (see Table 3.1). Group A contains eleven varieties of edible oils: extra virgin olive oil (EVOO), extra light olive oil (ELOO), pure olive oil, avocado, peanut, safflower, hazelnut, sunflower, sweet almond, almond and avocado-olive-flaxseed. Group B contains five varieties of edible oils: canola, extra virgin sesame, toasted sesame, canola-sunflower-soybean and canola-vegetable. Group C contains three varieties: corn, grapeseed and vegetable, whereas Group D is represented by only walnut oil.

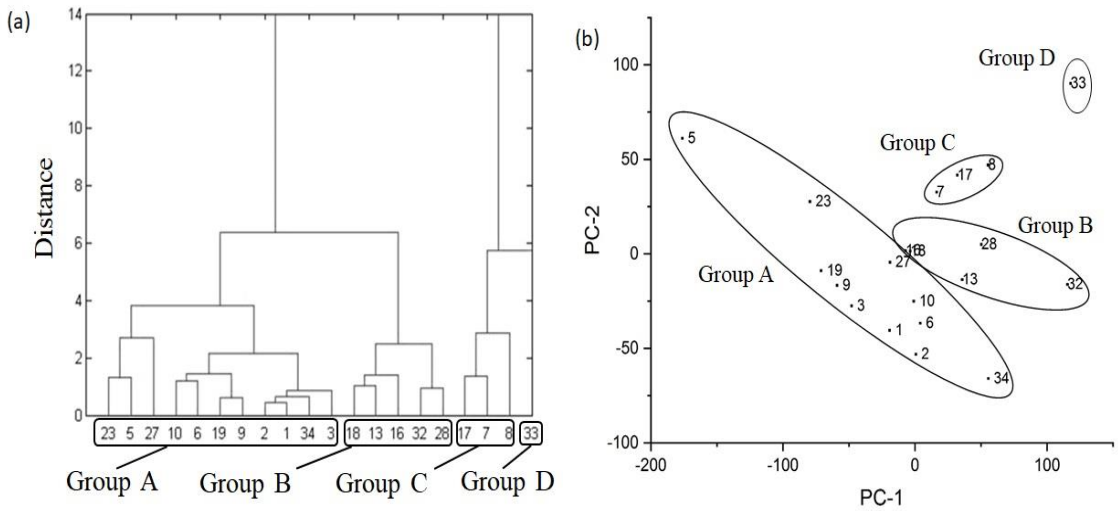

Figure 3.2. (a) Hierarchical clustering (farthest linkage) of the 20 averaged FTIR spectra; (b) Principal component analysis of spectra obtained from the 20 averaged IR spectra. Group A = 1 (EVOO), 2 (ELOO), 3 (olive oil), 5 (avocado oil), 6 (peanut oil), 9 (safflower oil), 10 (hazelnut oil), 19 (sunflower oil), 23 (sweet almond oil), 27 (almond oil) and 34 (avocado-olive-flaxseed oil). Group B = 13 (canola oil), 16 (canola vegetable oil), 18 (canola-sun-soybean oil), 28 (extra virgin sesame oil) and 32 (toasted sesame oil). Group C = 7 (corn oil), 8 (grapeseed oil) and 17 (vegetable oil). Group D = 33 (walnut oil).

Figure 3.3 displays a PC plot of the 273 FTIR spectra and the 6921 features comprising the training set. Each IR spectrum is represented as a point in the plot. The four oil groups are not well separated in the PC plot. Therefore, variable selection was the next step. Deletion of uninformative spectral features can ensure that discriminatory information about the class (edible oil group) is the major source of variation in the data. For this reason, the pattern recognition GA was applied to the 273 IR spectra of the training set to uncover wavelengths characteristic of the spectral profile of each edible oil group. K_c in the fitness function of the pattern recognition GA (see Chapter 2, Equation 2.8) was assigned a value equal to the number of IR spectra in each class. The number of chromosomes comprising the population was 10,000, and the mutation operator was set at 0.4.

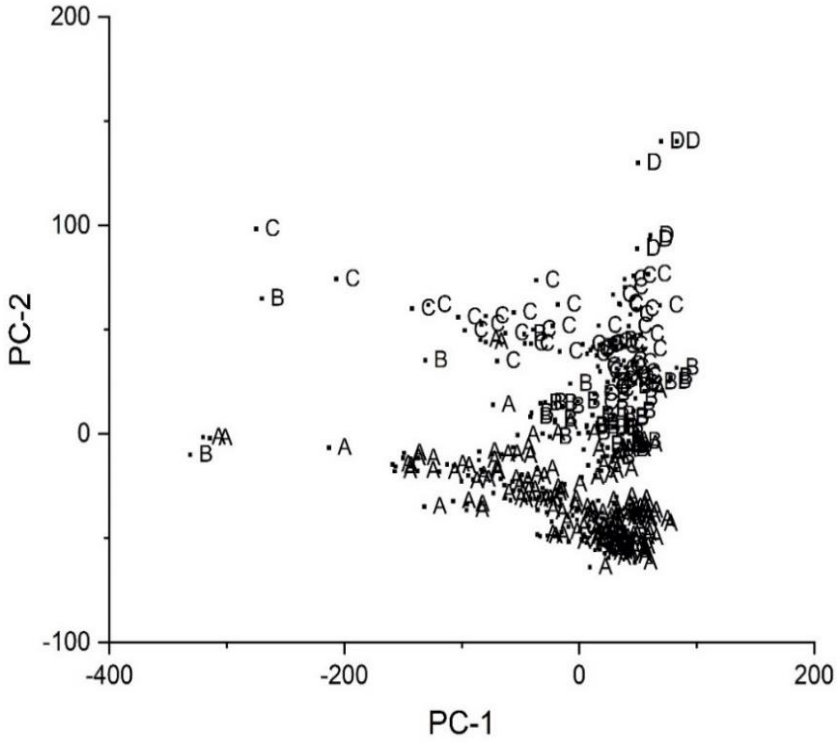

Figure 3.3. PC plot of the 273 IR spectra and 6921 features comprising the training set. Each IR spectrum is represented as a point in the plot. A = Group A, B = Group B, C = Group C, and D = Group D. The total cumulative variance explained by the two largest principal components for this training set data is 86.17%.

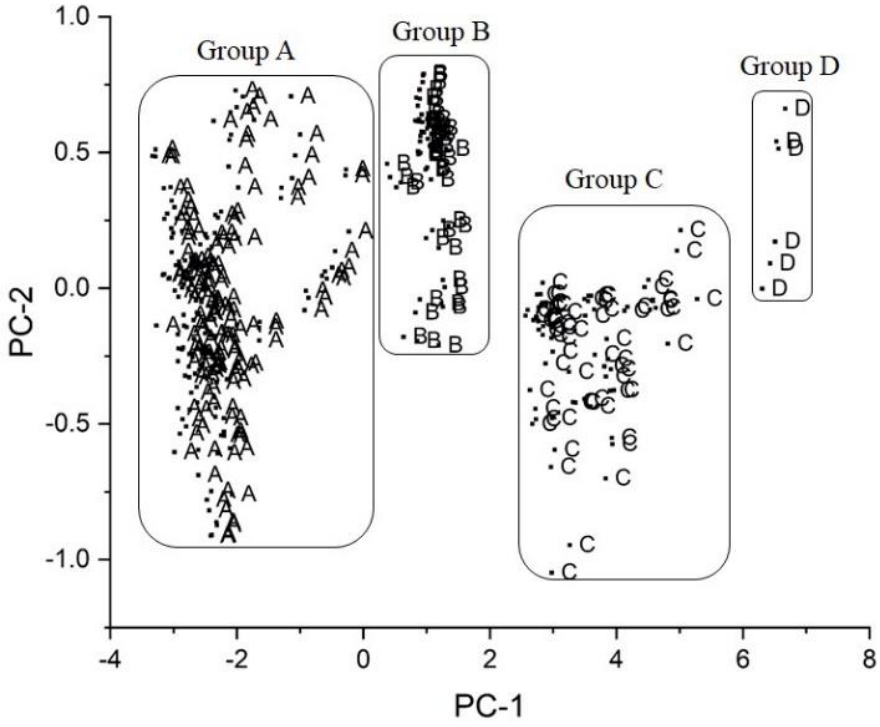

Figure 3.4. PC plot of the 273 IR spectra comprising the training set and the three spectral features identified by the pattern recognition GA. A = Group A. B = Group B. C = Group C. D = Group D.

The pattern recognition GA identified specific wavelengths correlated to oil group by sampling key feature subsets, scoring the corresponding PC plots, and tracking those samples and/or classes that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 200 generations, the pattern recognition GA identified 3 wavelengths whose PC plot showed clustering of the IR spectra by edible oil group (see Figure 3.4). Each oil group is well separated from the others in the PC plot. The three spectral features identified by the pattern recognition GA correspond to fundamental vibrational modes (see Table 3.4). The first principal component appears to be correlated to the amount of monounsaturated fatty acid content in each edible oil sample as the amount of monounsaturated fatty acids in these edible oils decreases from Group A to Group D [3-40 – 3-42]. Linolenic acid, which is low for Group A

edible oils, increases across the oil groups.[3-43, 3-44] As for Group D (walnut oil), it differs from the other nineteen edible oils due to its high omega 3 content.[3-45] In addition, walnut oil has the lowest monounsaturated fatty acid content and the highest linoleic acid content.

Table 3.4. Spectral features identified by the genetic algorithm.

Feature	Wavenumber (cm^{-1})	Assignment
609	695.20	Overlap of CH_2 rocking and out-of-plane vibrations of cis-disubstituted olefins
1457	1104.00	C-O stretching vibration
5415	3012.30	C-H stretching vibration of cis-double bond (=CH)

A prediction set of 83 IR spectra (see Table 3.2) was used to assess the predictive ability of the three spectral features identified by the pattern recognition GA. The 83 IR spectra were directly mapped onto the PC plot defined by the 273 FTIR spectra and the three spectral features identified by the pattern recognition GA. Figure 3.5 shows the IR spectra from the prediction set projected onto the PC plot developed using the training set data. All IR spectra from the prediction set were correctly classified as each spectrum is located in a region of the PC plot containing spectra that are tagged with the same class label.

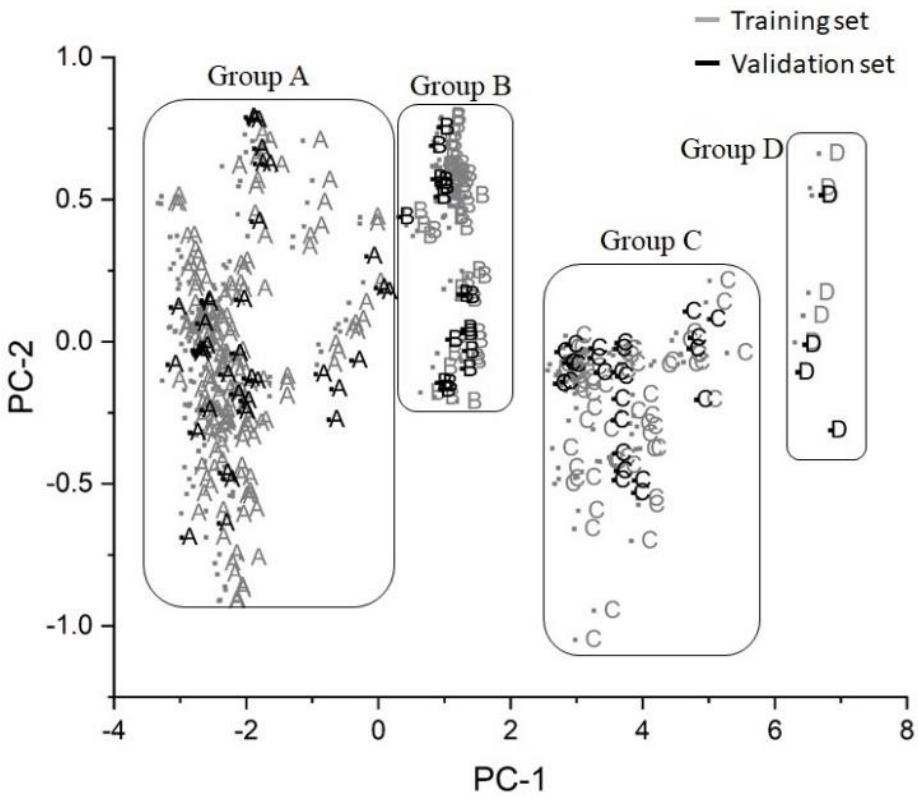

Figure 3.5. Projection of the 83 prediction set spectra onto the PC plot developed from the 273 spectra and the 3 features identified by the pattern recognition GA.

The next step was differentiating the edible oils in each group by type. For this reason, the pattern recognition GA was applied directly to the spectra in each edible oil group to identify wavelengths that discriminate the edible oils by variety within each group. The edible oils comprising Group A overlap in the PC plot of the data (see Figure 3.6 and Table 3.5), whereas the edible oils comprising Group B (see Figure 3.7 and Table 3.6) or Group C (see Figure 3.8 and Table 3.7) appear separated in the PC plot. From the results of the pattern recognition analysis and the differences in the chemical composition of each edible oil (e.g., monounsaturated fatty acids and linolenic acid content), it is evident that IR spectra of edible oils from different oil groups can be reliably differentiated (e.g., EVOO from Group 1 versus corn oil from Group 3), whereas IR spectra of the edible oils from the

same group (e.g., canola oil versus canola vegetable oil) would be more difficult to discriminate by type due to the similarity in their chemical composition.

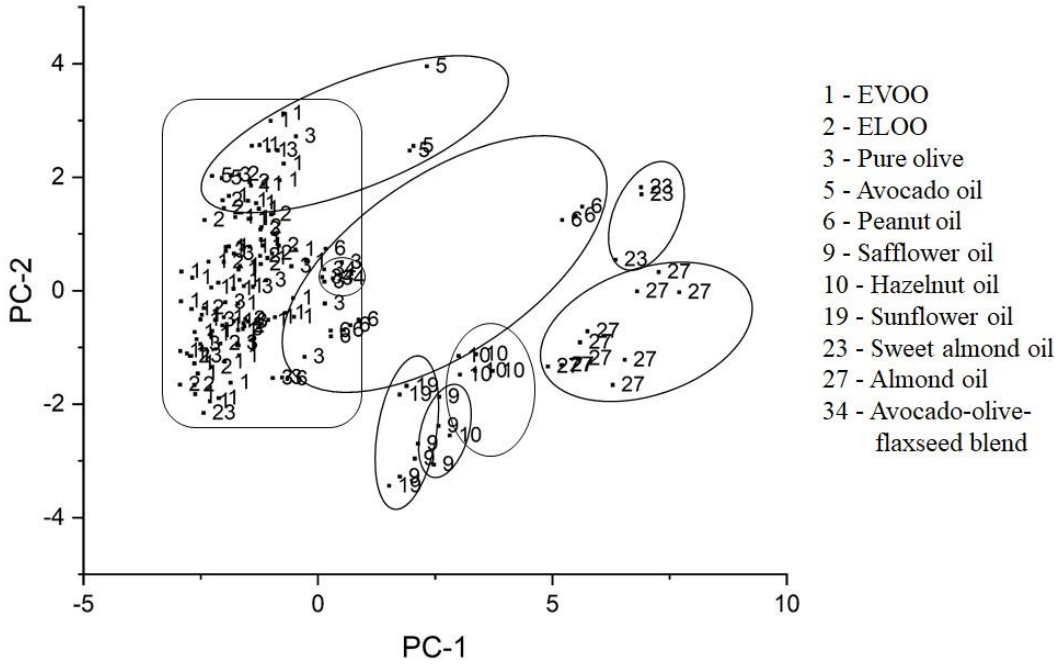

Figure 3.6. PC plot of the 157 IR spectra and the nine spectral features identified by the pattern recognition GA for Group A. The total cumulative variance explained by the two largest principal components for this training set data is 71.53 %.

Table 3.5. Spectral features identified for Group A oils by the genetic algorithm

Feature	Wavenumber (cm^{-1})	Assignment
596	688.95	C-C bending out of plane
1117	940.13	-HC=CH- (trans) bending out of plane
1333	1044.27	-C-O stretching
1402	1077.53	-C-O stretching
1467	1108.87	-C-O stretching
1800	1269.41	-C-O stretching, $-CH_2-$ bending
2194	1459.37	-C-H bending (asymmetric) of CH_3
2804	1753.46	-C=O stretching of ester group
5408	3008.89	=C-H (cis-) stretching

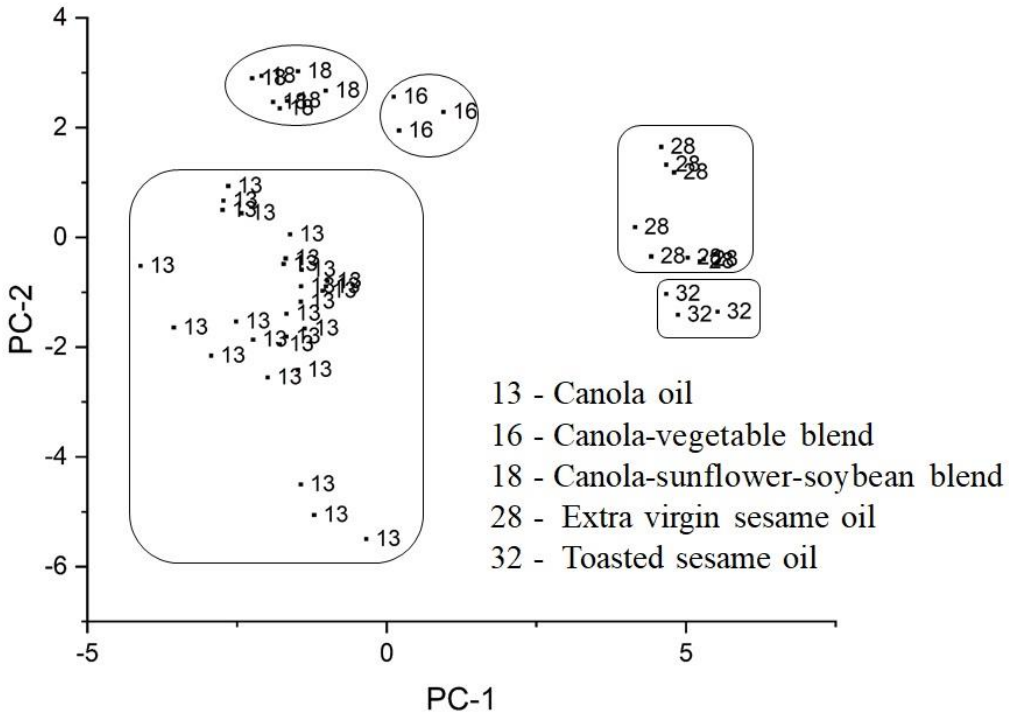

Figure 3.7. PC plot of the 49 IR spectra and the nine spectral features identified by the pattern recognition GA for Group B. The total cumulative variance explained by the two largest principal components for this training set data is 88.22%.

Table 3.6. Spectral features identified for Group B oils by the genetic algorithm

Feature	Wavenumber (cm^{-1})	Assignment
687	732.82	Overlap of CH_2 rocking and out of plane vibration of cis-disubstituted olefins
1052	908.79	-HC=CH- (cis) bending out of plane
1165	963.27	-HC=CH- (trans) bending out of plane
1167	964.23	-HC=CH- (cis) bending out of plane
1404	1078.50	-C-O stretching
1452	1101.64	-C-O stretching
1454	1102.60	-C-O stretching
2066	1397.66	=C-H bending
2190	1457.44	-C-H (CH_2) bending (scissoring)

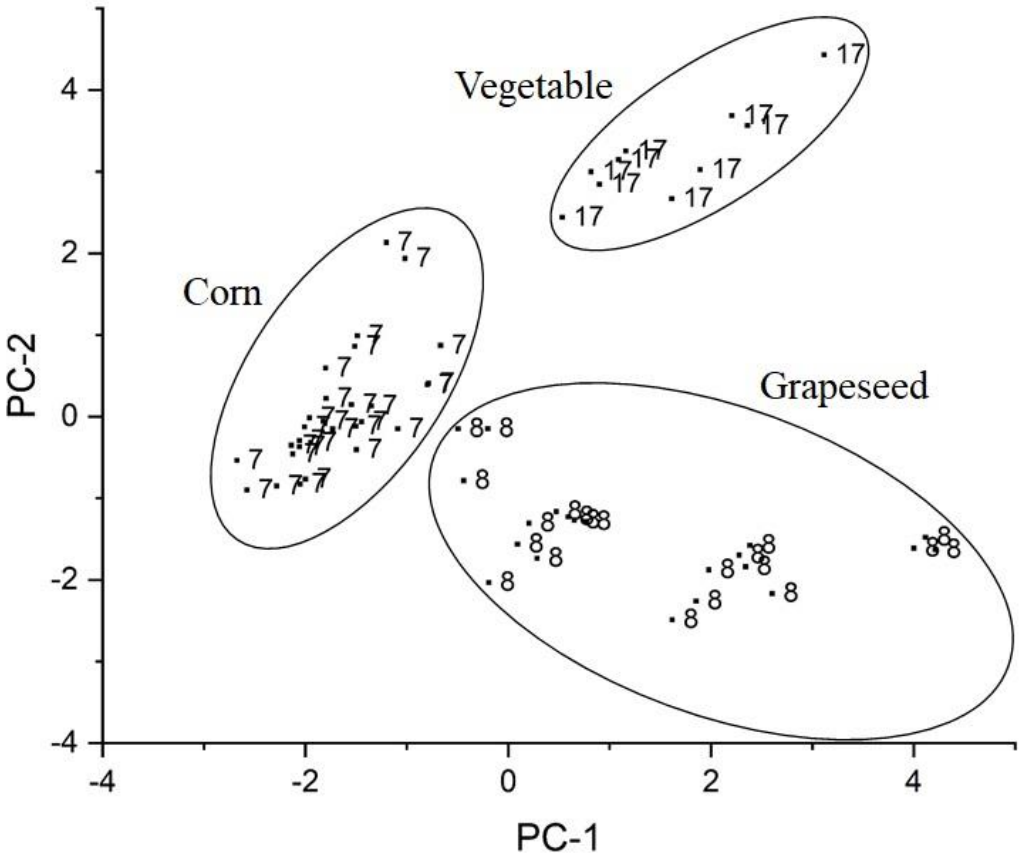

Figure 3.8. PC plot of the 61 IR spectra and the ten spectral features identified by the pattern recognition GA for Group C. The total cumulative variance explained by the two largest principal components for this training set data is 62.54%.

Table 3.7. Spectral features identified by the genetic algorithm for Group C

Feature	Wavenumber (cm^{-1})	Assignment
673	726.07	$-(CH_2)_n$ - rocking and -HC=CH- (cis-) bending out of plane
811	792.60	-C-H bending out of plane
1175	968.09	-HC=CH- (trans-) bending out of plane
1380	1066.92	-C-O stretching
1472	1111.28	-C-O stretching
1501	1125.26	-C-O stretching
2170	1447.80	-C-H bending (asymmetric) of CH_3
2207	1465.64	-C-H bending (scissoring) of CH_2
2843	1772.26	-C=O stretching of ester
5434	3021.43	=C-H (trans-) stretching

To determine whether two edible oils can be differentiated by FTIR spectroscopy, we compared the IR spectra of the pure edible oils to IR spectra of their mixtures (which simulate adulterated edible oils) using pattern recognition techniques. The edible oils compared are from the same oil group or from different oil groups. Specifically, IR spectra of EVOO were compared to IR spectra of EVOO (Group A) mixed with either corn oil (Group C), canola oil (Group B) or almond oil (Group A) in known amounts. Corn oil and canola oil were selected because these two edible oils have been previously reported as adulterants in EVOO [3-46], whereas almond oil, which is a member of the same edible oil group as EVOO, was selected because it appears to be the only edible oil separated from EVOO in a PC plot of the Group A edible oils (see Figure 3.6). Each comparison was formulated as a three-way classification problem: EVOO, adulterated EVOO, and adulterant. The goal was to obtain an overview of the dominant patterns present in the data. The EVOO-adulterant mixtures used for each training set and prediction set span a large concentration range.

For these studies, it is assumed that IR spectra of the mixtures of EVOO and corn oil, canola oil, or almond oil adhere to a linear mixture model. Consider, for example, the IR spectrum of an EVOO-corn oil mixture which is expected to be a combination of the IR spectrum of EVOO and the IR spectrum of corn oil with the weights of the constituents defining the mixing proportion of each pure edible oil that comprises the mixture. If the IR spectrum of EVOO can be differentiated from the IR spectrum of the EVOO-adulterant mixture, then differences between the IR spectra of the two pure edible oils (e.g., EVOO versus corn oil) are of sufficient magnitude to ensure that discrimination of these two edible oils by FTIR is viable.

To identify the wavelengths (in each three-way classification problem) that convey information about the degree of adulteration, K_c in the fitness function of the pattern recognition GA (see Chapter 2, Equation 2.8) is assigned a value equal to five for the adulterated EVOO mixtures, whereas K_c for EVOO or for corn oil, canola oil, or almond oil is assigned a value equal to the number of IR spectra comprising the class. K_c is an important parameter influencing the performance of the pattern recognition GA. By way of default, K_c is usually set equal to the number of samples in each class. This represents the most stringent criterion for developing a variable subset to classify the data as a variable subset can only receive a score of 100% from the fitness function when all the samples from the same class are closer to each other than to the samples from other classes. However, a value of K_c that is too large can introduce an "activation energy" thereby making the goal of identifying the optimal variable subset problematic. Decreasing the value of K_c can improve the classification of the samples comprising the training set. Smaller values of K_c are usually assigned to classes that are neither compact nor well separated in the pattern space to ensure the identification of the most efficacious variable subsets.

Each data set (EVOO/Corn, EVOO/Canola, and EVOO/Almond) was analyzed by the pattern recognition GA using the same set of parameters (number of chromosomes, selection pressure, configuration of initial population and K_c). The pattern recognition GA was developed to solve classification problems that are linearly separable. Since the concentration of the constituents comprising the adulterated edible oils is related to infrared absorbance through Beer's law, the application of the pattern recognition GA for variable selection in this phase of the classification study is both appropriate and logical. As in all

pattern recognition studies, each class must be well represented in the training set. The study design used ensured that there were enough samples to obtain meaningful results.

Figure 3.9 shows a plot of the two largest principal components of the 120 IR spectra (see Table 3.8) and the nine spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, corn oil and binary mixtures of EVOO and corn oil. EVOO, corn oil and the EVOO-corn oil mixtures cluster in separate regions of the PC plot. Furthermore, the first principal component appears to be correlated to the amount of corn oil in each edible oil sample.

The predictive ability of the nine spectral features identified by the pattern recognition GA was assessed using an external prediction set of eighteen IR spectra of EVOO and corn oil mixtures whose composition varied from 0% to 60% corn oil. The plant based edible oil samples used to prepare the EVOO-corn oil mixtures comprising the prediction set had been excluded from the training set for this classification problem. Figure 3.9 also shows the 24 prediction set spectra projected onto the PC plot of the 120 IR spectra comprising the training set and the nine spectral features identified by the pattern recognition GA. All FTIR spectra in the prediction set are correctly classified. Clearly, EVOO can be differentiated from corn oil using FTIR spectroscopy. The detection limit for corn oil in EVOO from the PC plot of the training set data is approximately 10% and is in agreement with the detection limit previously reported for corn oil in EVOO using PLS.[3-24]

Figure 3.10 shows a PC plot of the 119 IR spectra comprising the training set (see Table 3.9) and the thirteen spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, EVOO-canola oil, and canola oil. The IR

spectra of EVOO, canola oil, and mixtures of EVOO and canola oil can be differentiated using the discriminating relationship in the PC plot developed from the thirteen spectral features identified by the pattern recognition GA. Again, the first principal component appears to be correlated to the amount of adulterant (i.e., canola oil) in each sample. The discriminating relationship developed from these thirteen spectral features was successfully validated using the 22 IR spectra that comprised the prediction set (see Figure 3.10 and Table 3.9). The edible oil samples used to prepare the EVOO-canola oil mixtures that constitute the predictions set are (again) not the same samples used to prepare the edible oil mixtures used for the training set. This suggests that the IR spectral profiles of these edible oil mixtures possess features that are common to each constituent (EVOO and canola oil). The detection limit for canola oil in EVOO using FTIR spectroscopy is again 10%[3-47] (see Figure 3.10).

Figure 3.11 shows a PC plot of the 106 IR spectra comprising the training set (see Table 3.10) and the eleven spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, almond oil, and EVOO-almond oil mixtures. Although the samples comprising EVOO, almond oil, and the EVOO-almond oil mixtures cluster in different regions of the PC plot, the first principal component cannot be correlated to the amount of almond oil in the mixtures. Furthermore, only six of fifteen prediction set samples are correctly classified. Although the PC plot of the training set data for Group A (see Figure 3.6) suggests that EVOO and almond oil can be discriminated, the absence of spectral features in the data that can differentiate EVOO from EVOO adulterated with almond oil would indicate that almond oil and EVOO have similar IR spectra. This would

also be consistent with the first principal component not being correlated to the amount of almond oil in the samples.

Table 3.8. Composition of training and prediction sets for detection of corn oil in EVOO

	Number of spectra for training	Number of spectra for prediction
EVOO	67	5
Corn	35	1
EVOO-corn	18	18
Total	120	24

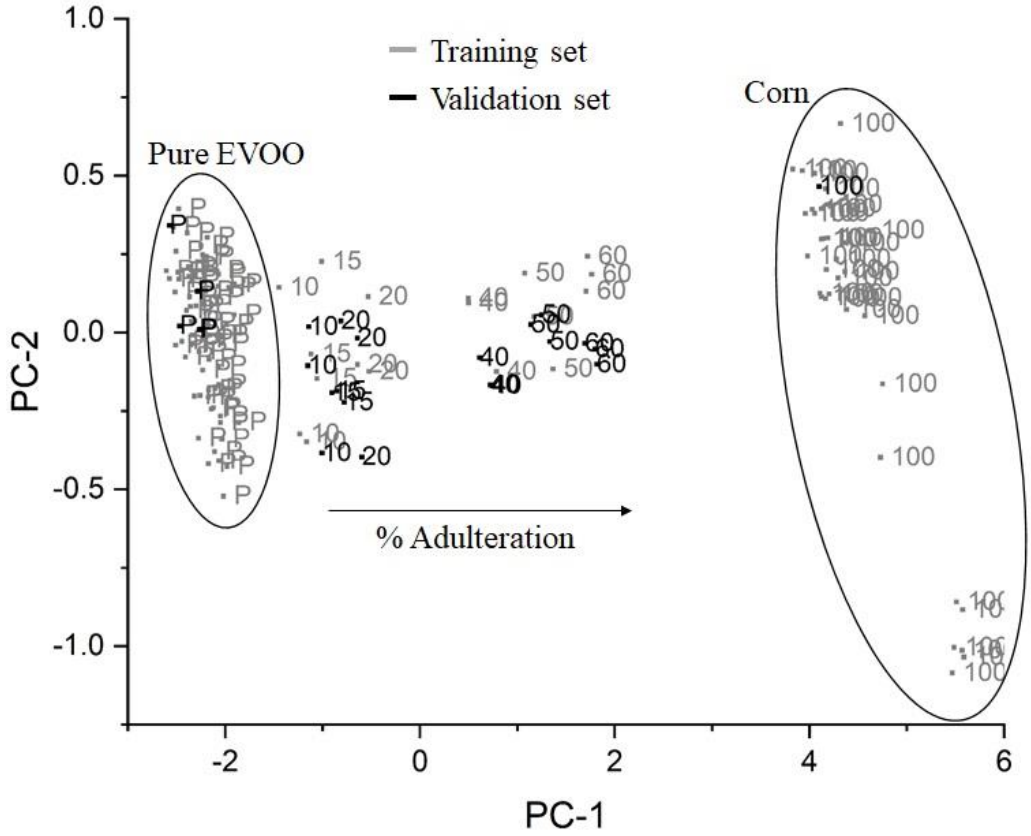

Figure 3.9. Plot of the two largest principal components of the 120 IR training set spectra (grey) and the nine spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, corn oil and EVOO-corn oil mixtures (10% corn oil to 60% corn oil). The prediction set spectra are represented in black color. The total cumulative variance explained by the two largest principal components for this training set data is 98.57%.

Table 3.9. Composition of training and prediction sets for detection of canola oil in EVOO

	Number of spectra for training	Number of spectra for prediction
EVOO	50	7
Canola	33	0
EVOO-canola	36	15
	119	22

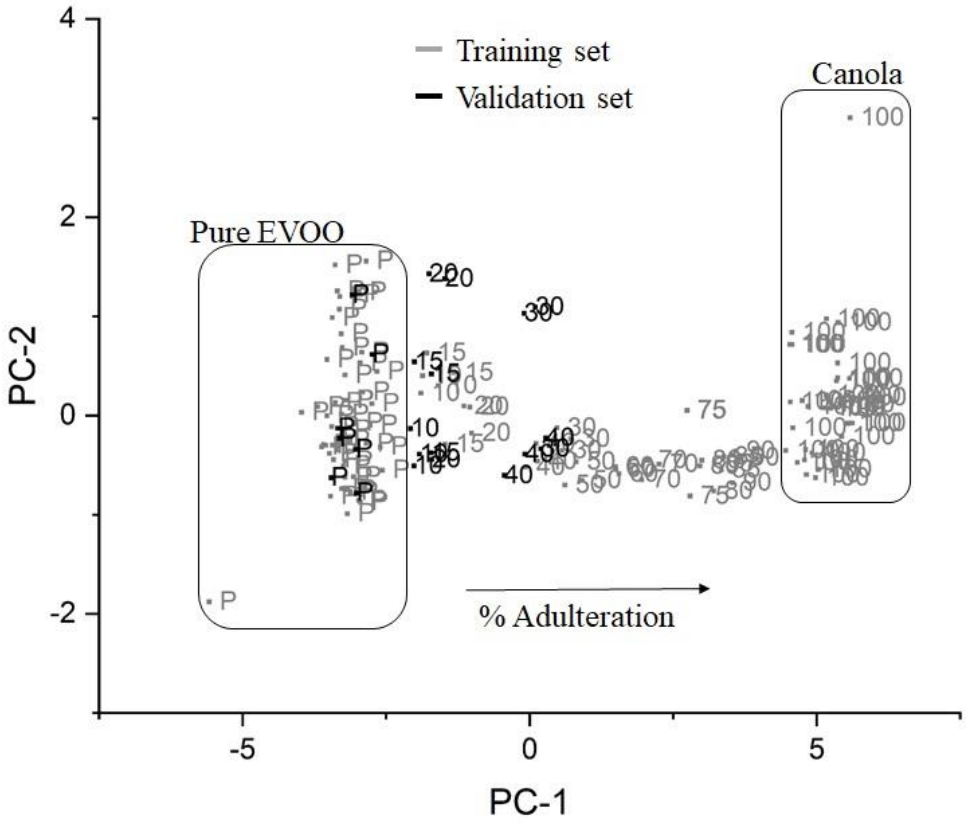

Figure 3.10. Plot of the two largest principal components of the 119 IR training set spectra (grey) and the thirteen spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, canola oil and EVOO-canola oil mixtures (10% canola oil to 90% canola oil). The prediction set spectra are represented in black color. The total cumulative variance explained by the two largest principal components for this training set data is 94.06%.

Table 3.10. Composition of training and prediction sets for detection of almond oil in EVOO

	Number of spectra for training	Number of spectra for prediction
EVOO	76	0
Almond	15	0
EVOO-almond	15	15
	106	15

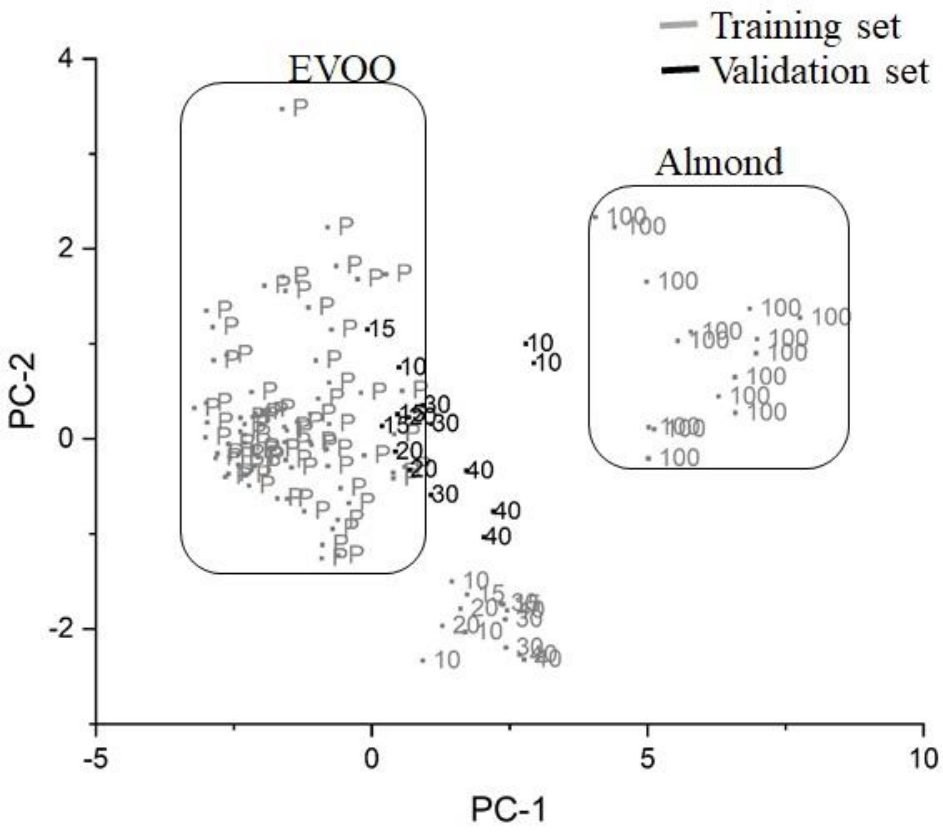

Figure 3.11. Plot of the two largest principal components of the 106 IR training set spectra (grey) and the eleven spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, almond oil and EVOO-almond oil mixtures (10% almond oil to 40% almond oil). The prediction set spectra are represented in black color. The total cumulative variance explained by the two largest principal components for this training set data is 85.29%.

3.4. Conclusions

This study was not designed to be a quantitative determination of the sources of variability spanning edible oils. Rather, it highlighted the challenges encountered when taking into account the different sources of variability that can impact the IR spectra of each of the twenty varieties of the edible oils surveyed in this study.

To differentiate the twenty-plant based edible oils investigated as part of this study, it was necessary to develop a hierarchical classification scheme. Each edible oil was first assigned to an oil group with the oils in each group then differentiated by type. From the adulteration studies that were undertaken, the composition of the edible oils that are in the same group as reflected by their IR spectra are similar to each other and therefore would be difficult to discriminate by type as opposed to differentiating edible oils that are from different groups. For example, an edible oil sample assigned to Group A can be readily differentiated from an edible oil sample from Groups B, C and D, whereas the IR spectra of edible oils within the same group more closely resemble each other when sample variability has been incorporated into the experimental design. Furthermore, detecting the adulteration of an edible oil by a less expensive edible oil from the same group would appear to be problematic. Thus, FTIR spectroscopy can differentiate edible oils provided that the oils are from different groups.

References

3-1. Cooperative Patent Classification. Edible oils or fats. In *CPC Definition- Subclass A23D, version 2018.05*, United States Patent and Trademark Office (USPTO): **2018**.

3-2. Gouilleux, B.; Marchand, J.; Charrier, B.; Remaud, G. S.; Giraudeau, P. High-throughput authentication of edible oils with benchtop Ultrafast 2D NMR. *Food Chem.* **2018**, *244*, 153-158.

3-3. Ng, T. T.; Li, S.; Ng, C. C. A.; So, P. K.; Wong, T. F.; Li, Z. Y.; Chan, S. T.; Yao, Z. P. Establishment of a spectral database for classification of edible oils using matrix-assisted laser desorption/ionization mass spectrometry. *Food Chem.* **2018,** *252*, 335-342.

3-4. Simopoulos, A.P. Evolutionary Aspects of Diet: The Omega-6/Omega-3 Ratio and the Brain. *Mol Neurobiol*. **2011**, 44**,** 203–215.

3-5. NPR, Olive Oil Fraud Rampant as Demand Skyrockets. In *NPR Daily Newsletter*, August 7, **2007**.

3-6. CORDIS, Development and assessment of methods for the detection of adulteration of olive oil with hazelnut oil. European Commision: Brussels, Belgium, 2001.

3-7. Andrikopoulos, N. K.; Giannakis, I. G.; Tzamtzis, V. Analysis of Olive Oil and Seed Oil Triglycerides by Capillary Gas Chromatography as a Tool for the Detection of the Adulteration of Olive Oil. *Journal of Chromatographic Science* **2001**, *39* (4), 137-145.

3-8. Aparicio, R.; Aparicio-Ruíz, R. Authentication of vegetable oils by chromatographic techniques. *Journal of Chromatography A* **2000**, *881* (1), 93-104.

3-9. López-Díez, E. C.; Bianchi, G.; Goodacre, R. Rapid Quantitative Assessment of the Adulteration of Virgin Olive Oils with Hazelnut Oils Using Raman Spectroscopy and Chemometrics. *Journal of Agricultural and Food Chem.* **2003**, *51* (21), 6145-6150

3-10. Baeten, V.; Meurens, M.; Morales, M. T.; Aparicio, R. Detection of Virgin Olive Oil Adulteration by Fourier Transform Raman Spectroscopy. *J. Agric. Food Chem.* **1996**, *44* (8), 2225-2230.

3-11. Downey, G.; McIntyre, P.; Davies, A. N. Detecting and quantifying sunflower oil adulteration in extra virgin olive oils from the Eastern Mediterranean by visible and near-infrared spectroscopy. *J. Agric. Food Chem.* **2002**, *50* (20), 5520-5525.

3-12. Karoui, R.; Downey, G.; Blecker, C. Mid-infrared spectroscopy coupled with chemometrics: A tool for the analysis of intact food systems and the exploration of their molecular structure-quality relationships-A review. *Chem. Rev.* **2010**, *110* (10), 6144-6168.

3-13. Vanstone, N.; Moore, A.; Martos, P.; Neethirajan, S. Detection of the adulteration of extra virgin olive oil by near-infrared spectroscopy and chemometric techniques. *FSQR* **2018**, *2* (4), 189-198.

3-14. Wold, S.; Esbensen, K.; Geladi, P., Principal component analysis. *Chemometr Intell Lab Syst.* **1987**, *2* (1), 37-52.

3-15. Obeidat, S. M.; Khanfar, M. S.; Obeidat, W. M. Classification of edible oils and uncovering adulteration of virgin olive oil using FTIR with the aid of chemometrics. *AJBAS* **2009**, *3* (3), 2048-2053.

3-16. Moore, J. C.; Lipp, M.; Griffiths, J. C., Preventing the adulteration of food protein with better analytical methods: Avoiding the "next melamine". *INFORM - International News on Fats, Oils and Related Materials* **2011**, *22* (6), 373-375.

3-17. de la Mata, P.; Dominguez-Vidal, A.; Bosque-Sendra, J. M.; Ruiz-Medina, A.; Cuadros-Rodríguez, L.; Ayora-Cañada, M. J. Olive oil assessment in edible oil blends by means of ATR-FTIR and chemometrics. *Food Control* **2012**, *23* (2), 449-455.

3-18. Rohman, A.; Che Man, Y. B., The use of Fourier transform mid infrared (FT-MIR) spectroscopy for detection and quantification of adulteration in virgin coconut oil. *Food Chem.* **2011**, *129* (2), 583-588.

3-19. Rohman, A.; Man, Y. B. C., Fourier transform infrared (FTIR) spectroscopy for analysis of extra virgin olive oil adulterated with palm oil. *Int. Food Res. J.* **2010**, *43* (3), 886-892.

3-20. Wójcicki, K.; Khmelinskii, I.; Sikorski, M.; Sikorska, E. Near and mid infrared spectroscopy and multivariate data analysis in studies of oxidation of edible oils. *Food Chem.* **2015**, *187*, 416-423.

3-21. Luna, A. S.; da Silva, A. P.; Ferré, J.; Boqué, R. Classification of edible oils and modeling of their physico-chemical properties by chemometric methods using mid-IR spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2013**, *100*, 109-114.

3-22. Lerma-García, M. J.; Ramis-Ramos, G.; Herrero-Martínez, J. M.; Simó-Alfonso, E. F. Authentication of extra virgin olive oils by Fourier-transform infrared spectroscopy. *Food Chem.* **2010**, *118* (1), 78-83.

3-23. Jiménez-Carvelo, A. M.; Osorio, M. T.; Koidis, A.; González-Casado, A.; Cuadros-Rodríguez, L. Chemometric classification and quantification of olive oil in blends with any edible vegetable oils using FTIR-ATR and Raman spectroscopy. *LWT - J. Food Sci. Technol.* **2017**, *86*, 174-184.

3-24. Vlachos, N.; Skopelitis, Y.; Psaroudaki, M.; Konstantinidou, V.; Chatzilazarou, A.; Tegou, E. Applications of Fourier transform-infrared spectroscopy to edible oils. *Analytica Chimica Acta* **2006**, *573-574*, 459-465.

3-25. Tay, A.; Singh, R. K.; Krishnan, S. S.; Gore, J. P. Authentication of Olive Oil Adulterated with Vegetable Oils Using Fourier Transform Infrared Spectroscopy. *LWT - J. Food Sci. Technol.* **2002**, *35* (1), 99-103.

3-26. Zhang, Q.; Liu, C.; Sun, Z.; Hu, X.; Shen, Q.; Wu, J. Authentication of edible vegetable oils adulterated with used frying oil by Fourier Transform Infrared Spectroscopy. *Food Chem.* **2012**, *132* (3), 1607-1613.

3-27. Lavine, B. K.; White, C. G.; Ding, T.; Gaye, M. M.; Clemmer, D. E. Wavelet based classification of MALDI-IMS-MS spectra of serum N-Linked glycans from normal controls and patients diagnosed with Barrett's esophagus, high grade dysplasia, and esophageal adenocarcinoma. *Chemometr Intell Lab Syst.* **2018**, 176,74 - 81**.**

3-28. Perera, U. D. N; Nishikida K.; Lavine, B. K. Development of Infrared Library Search Prefilters for Automotive Clear Coats from Simulated ATR Spectra. *J. Appl. Spectrosc.* **2018**, 186, 662-669.

3.29. Lavine, B. K.; White, C. G.; DeNoyer, L.;Mechref, Y. Multivariate classification of disease phenotypes of esophageal adenocarcinoma by pattern recognition analysis of MALDI-TOF mass spectra of serum N-linked glycans. *Microchemical Journal* **2017**, 132, 83-88.

3-30. Lavine, B. K.; Mirjankar, N; Delwiche, S. Classification of the waxy condition of durum wheat by near infrared reflectance spectroscopy using wavelets and a genetic algorithm. *Microchemical Journal* **2014**, 117, 178-182.

3-31. Kwofie, F.; Lavine, B. K.; Ottaway, J.; Booksh, K. Incorporating brand variability into classification of edible oils by Raman spectroscopy. *J. Chemom.* **2019**, 34 (7), 1-14

3-32. Kwofie, F.;, Lavine, B. K.; Ottaway, J.; Booksh, K. Differentiation of edible oils by type using Raman spectroscopy and pattern recognition methods. *Appl. Spec.* **2020**, 74(6), 645-654.

3-33. Dong, W.; Zhang, Y.; Zhang, B.; Wang, X. Quantitative analysis of adulteration of extra virgin olive oil using Raman spectroscopy improved by Bayesian framework least squares support vector machines. *Anal. Methods* **2012**, 4 (9), 2772-2777.

3-34. Zhang, X. F.; Zou, M. Q.; Qi, X. H.; Liu, F.; Zhang, C.; Yin, F. Quantitative detection of adulterated olive oil by Raman spectroscopy and chemometrics. *Journal of Raman Spectroscopy* **2011**, 42 (9), 1784-1788.

3-35. Lavine, B. K.; Davidson, C. E.; Moores, A. J.; and Griffiths, P. R. Raman spectroscopy and genetic algorithms for the classification of wood types. *Applied Spectroscopy* **2001**, 55 (8), 960 - 966.

3-36. Lavine, B. K.; Davidson, C. E.; Moores, A. J. Genetic algorithms for spectral pattern 3-recognition. Vibrational Spectroscopy , **2002**, 28 (1), 83-95.

3-37. Lavine, B. K.; Moores, A. J. Genetic algorithms for pattern recognition analysis and fusion of sensor data. In Pattern recognition, chemometrics and imaging of optical environmental monitoring. Siddiqui, A.; Eastwood, D., Eds.; Proceedings of SPIES, **1999**, 103-112.

3-38. Tou, J. T.; Gonzalez, R. C. Pattern recognition principles, Addison Wesley Publishing 3-Company, Reading, MA **1974**.

3-39. Lavine, B. K.; Mirjankar, N. Clustering and Classification of Analytical Data, in Encyclopedia of Analytical Chemistry, John Wiley & Sons, Ltd. 2012.

3-40. Savage, G. P.; Dutta, P. C.; McNeil, D. L. Fatty acid and tocopherol contents and oxidative stability of walnut oils. *J. Am. Oil Chem.' Soc.* **1999,** *76* (9), 1059-1063.

3-41. Zwarts, G. P.; Savage, G. P.; McNeil, D. L. Fatty acid content of New Zealand-grown walnuts (Juglans regia L.). *Int. J. Food Sci. Nutr.* **1999,** *50* (3), 189-194.

3-42. Fasina, O. O.; Hallman, H.; Craig-Schmidt, M.; Clements, C. Predicting temperature-dependence viscosity of vegetable oils from fatty acid composition. *J. Am. Oil Chem.' Soc.* **2006,** *83* (10), 899.

3-43. Azizian, H.; Mossoba, M. M.; Fardin-Kia, A. R.; Delmonte, P.; Karunathilaka, S. R.; Kramer, J. K. G. Novel, Rapid Identification, and Quantification of Adulterants in Extra Virgin Olive Oil Using Near-Infrared Spectroscopy and Chemometrics. *Lipids* **2015,** *50* (7), 705-718.

3-44. Maguire, L. S.; O'Sullivan, S. M.; Galvin, K.; O'Connor, T. P.; O'Brien, N. M. Fatty acid profile, tocopherol, squalene and phytosterol content of walnuts, almonds, peanuts, hazelnuts and the macadamia nut. *Int. J. Food Sci. Nutr.* **2004,** *55* (3), 171-178.

3-45. United States Department of Agriculture. Provisional table on the content of omega-3 fatty acids and other fat components in selected foods. In: Simopoulos A. P., Kifer R. R., Martin RE, eds. Health effects of polyunsaturated fatty acids in seafoods. Orlando, FL: Academic Press, 1986; 453–458.

3-46. Maninder , M.; Qianxi, C.; Baojun, X. A critical review on analytical techniques to detect adulteration of extra virgin olive oil. *Trends Food Sci. Technol.* **2019**, 91, 391 - 408.

3-47. Gurdeniz, G; Ozen, B. Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data. *Food Chem.* **2009**, 116, 519 - 525.

CHAPTER IV

Authentication of Edible Oils Using an Infrared Spectral Library and Digital Sample Sets

4.1. Introduction

Edible oils are an important component of the human diet due to their high nutritional value serving as a major source of fatty acids and fat-soluble vitamins in many diets.[4-1, 4-2] These oils are primarily composed of triglycerides which contain saturated, monounsaturated and polyunsaturated fatty acids.[4-3] The relative quantity of each fatty acid is related to the specific variety of the edible oil.[4-4] For example, safflower oil has more polyunsaturated fatty acids than extra virgin olive oil. Edible oils are typically used in cooking and are also ingredients in many preprocessed foods because of their sensory characteristics.

Adulteration of edible oils is an important chemical analysis problem as the most frequently adulterated food is extra virgin olive oil (EVOO).[4-5] Adulteration of a more expensive edible oil by either substitution or blending with less expensive cooking oils is of concern to government and regulatory officials. Adulterated EVOO cannot meet the International Olive Council's standards for the composition of monounsaturated fatty acids, free fatty acids, trans fatty acids, peroxides, and esterified fatty acids[4-6] and cannot be detected by either the consumer or retailer as the adulterated cooking oil is often comparable in appearance and flavor to EVOO. In addition, adulteration of EVOO by

less expensive edible oils such as peanut oil (which contains allergens) poses a serious health risk. The successful classification of edible oils by variety (e.g., discrimination of EVOO from peanut oil) is a crucial first step in solving this problem. In the preceding chapter, a pattern recognition study was reported describing the FTIR analysis of ninety-seven edible oil samples from twenty plant-based varieties collected over a three-year period. The ninety-seven edible oil samples that were selected encompassed multiple brands and manufacturers representing supplier to supplier variation as well as seasonal and batch variation within a supplier. Using a hierarchical classification scheme, the twenty plant-based varieties of edible oils could be divided into four distinct groups. Edible oils from different oil groups were reliably discriminated, whereas the discrimination of edible oils within the same group was problematic. Adulteration of the plant-based edible oils by other oils in the same group (e.g., EVOO by almond oil) could not be reliably detected using FTIR spectroscopy, whereas adulteration of edible oils by other edible oils that were not part of the same oil group (e.g., EVOO adulterated by corn or canola oil) could be detected at concentration levels as low as 10% (v/v) which was consistent with the results reported in previously published studies using partial least squares regression. A unique aspect of this study was the incorporation of edible oils collected systematically over three years, which introduced a heretofore unseen variability in the chemical composition of the edible oils. This work also demonstrated that previously published studies (which relied on a single sample or brand to represent each variety of edible oil) provide an overly optimistic estimate of the capability of FTIR spectroscopy to discriminate plant based edible oils by variety as well as detect the presence of adulterants in edible oils.

In this chapter, a potential method to determine whether two varieties of edible oils can be differentiated is proposed using digitally generated data of adulterated edible oils from an IR spectral library. The first step is the evaluation of the digitally blended data sets which is the focus of this chapter. Specifically, IR spectra of adulterated edible oils are computed from digitally blended experimental data of the IR spectra of an edible oil and the corresponding adulterant using the appropriate mixing coefficients for the spectra to achieve the desired level of adulteration. To determine whether two edible oils can be differentiated by FTIR spectroscopy, pure IR spectra of the two edible oils were compared to IR spectra of the two edible oils that were digitally mixed using pattern recognition techniques to solve a ternary classification problem. If the IR spectra of the two edible oils and their binary mixtures are differentiable, then differences between the IR spectra of these two edible oils are of sufficient magnitude to ensure that a reliable classification of these two edible oils by FTIR spectroscopy can be obtained. Using this approach, the feasibility of authenticating edible oils such as EVOO directly from library spectra has been demonstrated. For this study, both digital and experimental data were combined to generate training and validation data sets to assess detection limits for adulterants.

4.2. Edible Oil Spectral Library

An IR spectral database of 3720 IR spectra of both pure and adulterated edible oils have been collected using an iS50 Thermo-Nicolet IR spectrometer equipped with a diamond ATR accessor and a DTGS detector. The pure edible oil samples (99 in total) comprising the library spanned 20 distinct plant-based edible oil varieties (see Table 4.1). The sample cohort was obtained from supermarkets in the greater metropolitan Newark,

DE area over three years to account for brand, lot, year, storage and seasonal variability for a particular manufacturer. To further characterize these edible oils, their peroxide value was measured using a spectrophotometric method (see Chapter 3). Table 4.1 lists the twenty oil types that comprise the 99 samples collected, the number of samples collected for each edible oil and the number of spectra per sample. For each resolution, a total of 377 IR spectra were collected with 1508 IR spectra collected in total for the pure edible oils.

Each pure and adulterated edible oil sample was analyzed at 4 cm^{-1}, 6 cm^{-1}, 8 cm^{-1}, and 16 cm^{-1} resolution. All adulterated samples were prepared by mixing EVOO, extra light olive oil (ELOO) or sesame oil with less expensive edible oils (corn, canola, almond, peanut, sunflower, hazelnut, grapeseed, safflower, and vegetable) using a digital pipette to prepare adulterated mixtures by v/v in known amounts from 5% to 90%. For example, a 10% adulterated mixture of EVOO with corn oil as the adulterant was prepared by mixing 900 μL of extra virgin olive oil and 100 μL of corn oil in a 15 mL sterile falcon tube using a Thermolyne MaxiMixPlus vortex mixer. 416 IR spectra of EVOO, ELOO, or sesame oil adulterated by corn oil, canola oil, almond, peanut, sunflower, hazelnut, grapeseed, safflower, and vegetable oils were also collected at 4 cm^{1}, 6 cm^{-1}, 8 cm^{-1}, and 16 cm^{-1} resolution for a total of 1664 IR spectra (see Table 4.2). Ternary mixtures (which consist of two adulterants added to EVOO, ELOO, or sesame oil) were also prepared. The adulterants used in the ternary mixtures included corn, canola, almond, hazelnut, vegetable, grapeseed and safflower oils. 128 FTIR spectra of the ternary mixtures were also collected at 4 cm^{-1}, 6 cm^{-1}, 8 cm^{-1}, and 16 cm^{-1} resolution for a total of 512 spectra (see Table 4.3). The FTIR IR spectra (4000 cm^{-1} to 400 cm^{-1}) of each pure and adulterated edible oil sample were collected in triplicate, quadruplicate, or quintuplicate, each at 64 scans. All FTIR

spectra in the database were baseline corrected using OMNIC and normalized to unit length with MATLAB. Apodization of the spectra was performed using OMNIC and the Happ-Genzel function.

The FTIR spectral database is a flexible platform as it allows analytical chemists to test new data analysis methodologies. Experimental designs can be constructed with very similar edible oils (e.g., EVOO and sunflower oil) or oils with relatively distinct spectra (e.g., EVOO and corn oil). One can progress from simple classifications of mixtures (e.g., extra virgin olive oil that contains peanut oil), quantitative mixture analysis (relative concentrations of adulterants in edible oils) to quantitative determinations of intrinsic properties of edible oils (e.g., peroxide number to assess rancidity). The effect of spectral resolution on the outcome of the classification or calibration for constructing models to determine rancidity can also be assessed using this database.

Table 4.1. Composition of the pure edible oils in the IR library

Pure edible oil	Oil Type ID	Number of samples	Number of spectra
Extra virgin olive oil	1	26	83
Extra light olive oil	2	8	27
Olive oil	3	8	26
Avocado oil	5	2	9
Peanut oil	6	4	19
Corn oil	7	9	42
Grapeseed oil	8	9	36
Safflower oil	9	2	9
Hazelnut oil	10	2	9
Canola oil	13	9	36
Canola-vegetable blend	16	1	3
Vegetable oil	17	4	14
Canola-sunflower-soybean blend	18	1	9
Sunflower oil	19	1	3
Sweet almond oil	23	2	6
Almond oil	27	4	15
Extra virgin sesame oil	28	3	15
Toasted sesame oil	32	1	3
Walnut oil	33	2	10
Avocado-olive-flaxseed blend	34	1	3
		99	377

Table 4.2. Composition of the binary edible oil mixtures in the IR library

Binary mixtures	Oil Type ID	Number of spectra
ELOO-corn mixture	40	45
EVOO-corn mixture	44	60
EVOO-peanut mixture	45	24
Sesame-sunflower mixture	47	24
Sesame-canola mixture	48	21
Sesame-corn mixture	49	26
EVOO-almond mixture	51	36
Sesame-grapeseed mixture	55	18
ELOO-hazelnut mixture	54	18
Sesame-vegetable mixture	58	18
EVOO-canola mixture	60	78
ELOO-canola mixture	61	24
ELOO-safflower mixture	62	24
		416

Table 4.3. Composition of the ternary mixtures in the IR library

Ternary mixtures	Oil Type ID	Number of spectra
ELOO-corn-canola mixture	43	32
EVOO-corn-canola mixture	46	24
Sesame-corn-canola mixture	50	12
EVOO-almond-hazelnut mixture	52	12
Sesame-grapeseed-corn mixture	56	12
ELOO-almond-hazelnut mixture	53	12
Sesame-corn-safflower mixture	57	12
Sesame-vegetable-safflower	59	12
		128

4.3. Preparation of digitally blended data from IR spectra of edible oils

Digital blending refers to the mixing proportion of each edible oil that comprises the oil mixture. Digital blending was performed on the unprocessed IR spectra. To obtain a digital blend representing an 80% EVOO and 20% corn oil mixture, the IR spectrum of an EVOO sample is multiplied by 0.8 and added to an IR spectrum of a corn oil sample that is multiplied by 0.2. Gaussian distributed noise was added to the IR spectrum of each digital blend to homogenize the spectral data. For each spectrum, noise was added to the regions which contained IR bands (402 cm^{-1} to 1525 cm^{-1}, 1600 cm^{-1} to 1850 cm^{-1} and 2750 cm^{-1} to 3150 cm^{-1}). For a training set of digitally blended IR spectra, the largest absorbance value at each wavelength was identified, and one thousandth of this value was multiplied by Gaussian distributed random noise which had a mean of zero and standard deviation of one. If the largest absorbance value was less than or equal to zero, noise was not added to the blended spectrum at that particular wavelength. For the pattern recognition studies that were undertaken to demonstrate equivalency between real data and digitally blended data, the full spectral range (4000 cm^{-1} to 400 cm^{-1}) was used.

4.4. Validation of Digitally Blended Data

To determine whether two edible oils can be differentiated by FTIR spectroscopy, the IR spectra of the pure edible oils was compared to the IR spectra of their mixtures (which simulate an adulterated edible oil). The focus of these studies was EVOO (which is frequently a target of adulteration), and each comparison was formulated as a three-way classification problem: EVOO, adulterated EVOO and adulterant (corn oil, canola oil or almond oil). The EVOO-adulterant mixtures used for the training and validation set for

both the experimental and digitally blended data span a large concentration range. For each comparison, it is assumed that the IR spectra of the adulterated mixture can be represented by the IR spectra of EVOO and the other edible oil, with the weights of the constituents defining the mixing proportion of each edible oil that comprises the mixture. If the IR spectra of EVOO and the adulterant adhere to a linear mixture model then the results of the three-way classification study for the experimental and blended data would be similar.

To identify the wavelengths in each three-way classification problem that convey information about the degree of adulteration for both the experimental and digitally blended data, the pattern recognition GA was applied to each of these data sets. Modifications to the PCKaNN fitness function was undertaken to allow incorporation of model inference into the variable selection process. The goal is to identify variables that minimize the error across the entire model. This was accomplished by assessing the uncertainty of the sample scores in the principal component plot using the jackknife[4-7] to generate estimates of dispersion. During each generation, the fitness function of the pattern recognition GA evaluates thousands of principal component plots, one for each feature subset (i.e., chromosome) in the population of solutions. For each principal component score plot, the corresponding training set samples are removed one at a time, and the score matrix and loading matrix for the resampled (i.e., jackknifed) training set is recomputed. (Due to the rotational ambiguities of PCA, the loading matrix for each resampled training set must be rotated using a Procrustean rotation[4-8] to match the loading matrix associated with the score plot containing all the samples.) For each training set sample, scores across all leave-one-out score plots will be projected onto the original principal component plot

of the feature subset which will then be scored using PCKaNN. Thus, information about the level of confidence in the classification of each training set sample is directly incorporated into the variable selection process with the jackknifed scores for each sample effectively comprising an error cloud to depict the uncertainty associated with each training set sample.

Each data set (EVOO/Corn, EVOO/Canola, and EVOO/Almond) was analyzed by the pattern recognition GA using the same set of parameters (number of chromosomes, selection pressure, configuration of initial population, and K_c). Figures 4.1 and 4.2 show the plots of the two largest principal components of the 118 FTIR spectra (see Table 4.4) and the 8 and 17 spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, EVOO-corn oil mixtures, and corn oil. EVOO, corn oil, and the binary mixtures of EVOO-corn oil cluster in separate regions of the PC plot for both the experimental and digitally blended data. The first principal component appears to be correlated to the amount of adulterant (i.e., corn oil) in each sample.

The predictive ability of the 8 and 17 spectral features identified by the pattern recognition GA was assessed using an external prediction set of 12 spectra of the EVOO-corn oil mixtures whose composition varied from 0% to 40% corn oil. The plant-based edible oil samples used to prepare the EVOO-corn oil mixtures comprising the prediction set were excluded from the training set. Figures 4.3 and 4.4 show the plots of the 12 prediction set spectra projected onto the principal component score plot of the 118 IR spectra comprising the training set and the 8 and 17 spectral features identified by the pattern recognition GA. All 12 FTIR spectra in the prediction set for both the experimental and digitally blended data were correctly classified as each spectrum is located in a region

of the principal component score plot that contain samples tagged with the same class label. Clearly, EVOO can be differentiated from corn oil. The detection limit for corn oil in EVOO from the principal component score plot is approximately 10% for both the experimental and digitally blended data which is in agreement with the detection limits previously reported for corn oil using PLS.[4-9] Furthermore, the agreement between the results obtained for the experimental and digitally blended data suggests that digitally blended data can be used to assess whether two different varieties of edible oils (e.g., EVOO versus corn oil) can be differentiated by FTIR spectroscopy using a ternary classification study.

Table 4.4. Training and prediction set for experimental and blended data

	Number of spectra in training set/prediction set
EVOO	73/0
Corn	33/0
EVOO-corn	12/12
Total	118/12

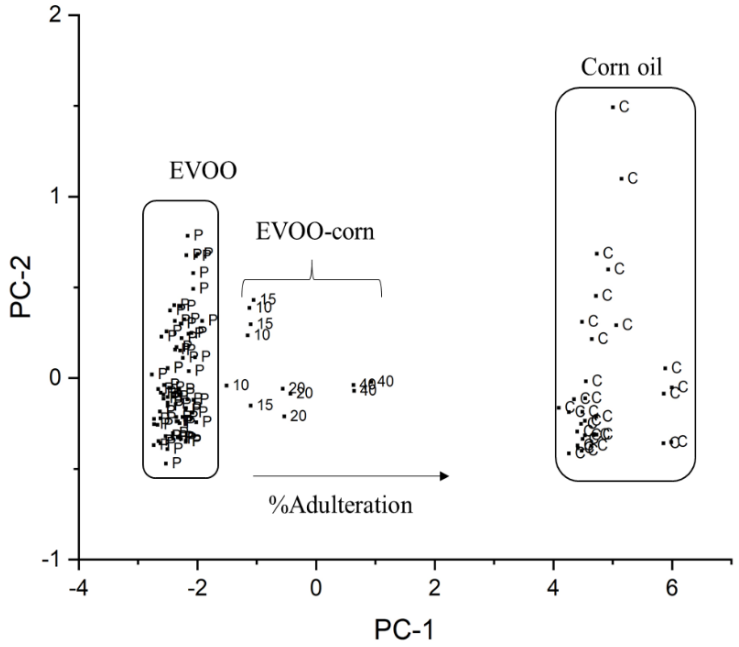

Figure 4.1. Plot of the two largest principal components of the 118 IR training set spectra (black) and the 8 spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, corn oil and EVOO-corn oil mixtures (10% corn oil to 40% oil). The total cumulative variance explained by the two largest principal components for the experimental data is 98.57%. P = EVOO, C = corn oil, 10 = 10% corn oil, 15 = 15% corn oil, 20 = 20% corn oil, and 40 = 40% corn oil.

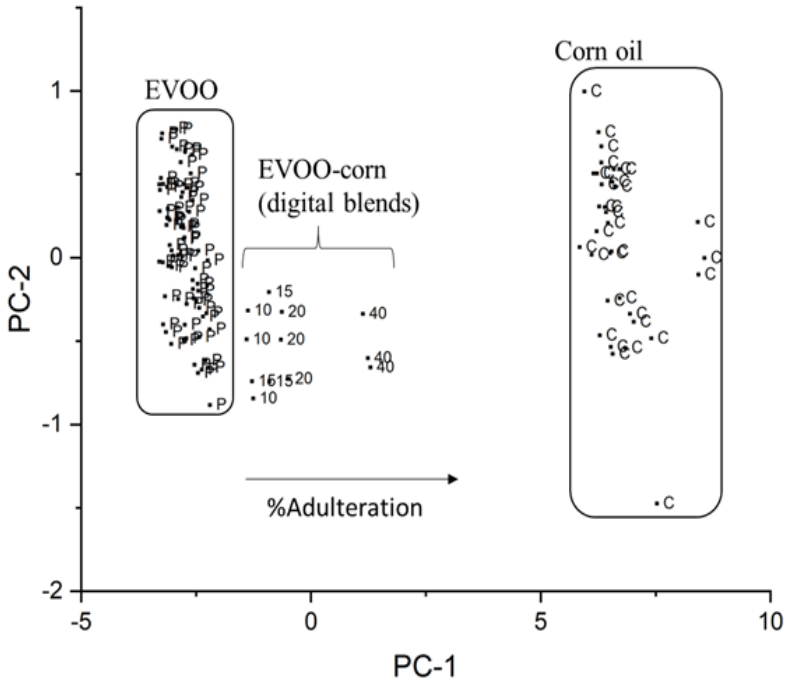

Figure 4.2. Plot of the two largest principal components of the 118 IR training set spectra (black) and the 17 spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, corn oil and EVOO-corn oil mixtures. The total cumulative variance explained by the two largest principal components for the digitally blended data is 97.18%. P = EVOO, C = corn oil, 10 = 10% corn oil, 15 = 15% corn oil, 20 = 20% corn oil, and 40 = 40% corn oil.

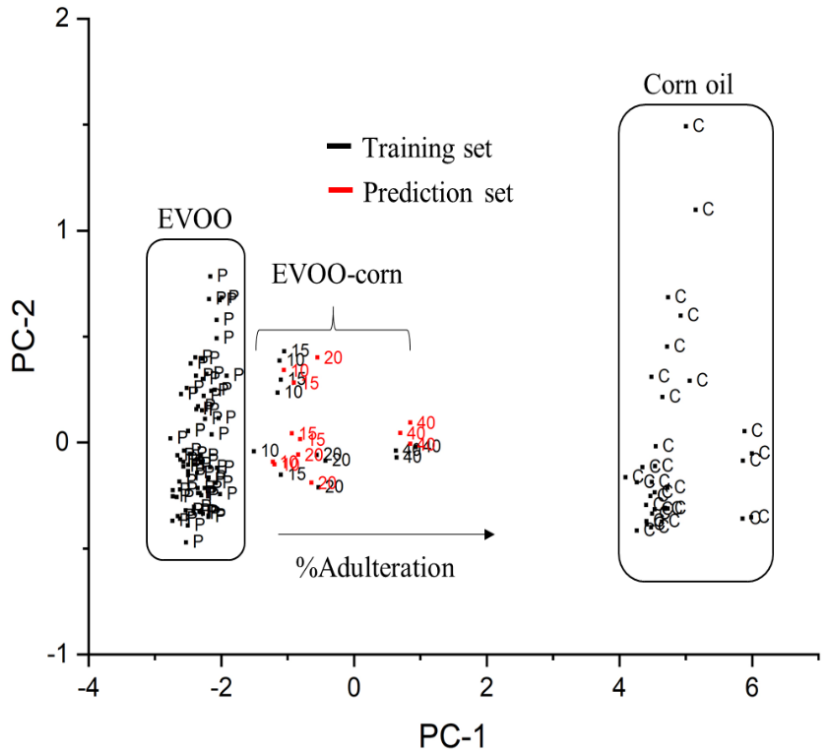

Figure 4.3. Projection of the 12 prediction set spectra (red) onto the PC-plot developed from the 118 training set spectra and 8 features identified by the pattern recognition GA for the experimental data. P = EVOO, C = corn oil, 10 = 10% corn oil, 15 = 15% corn oil, 20 = 20% corn oil, and 40 = 40% corn oil.

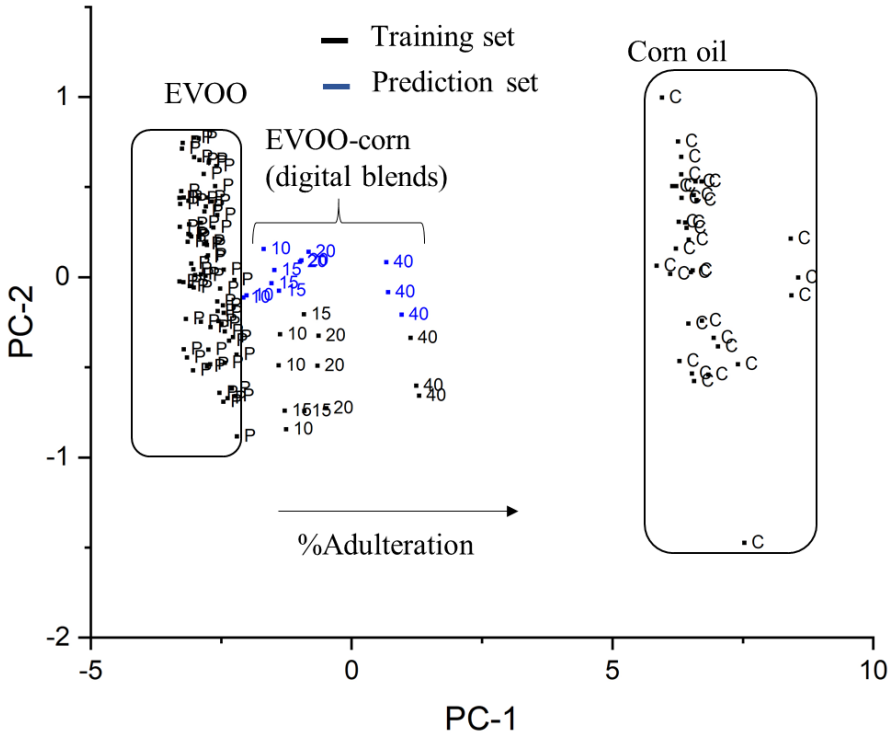

Figure 4.4. Projection of the 12 prediction set spectra (blue) onto the PC-plot developed from the 118 training set spectra and 17 features identified by the pattern recognition GA for the digitally blended data. P = EVOO, C = corn oil, 10 = 10% corn oil, 15 = 15% corn oil, 20 = 20% corn oil, and 40 = 40% corn oil.

Figures 4.5 and 4.6 show the plots of the two largest principal components of the 115 IR spectra comprising the training set (see Table 4.5) and the 9 and 11 spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, EVOO-canola oil mixtures, and canola oil. EVOO, canola oil, and the binary mixtures of EVOO-canola oil cluster in separate regions of the PC plot for both the experimental and digitally blended data. Again, the first principal component appears to be correlated to the amount of adulterant (i.e., canola oil) in each sample. The discriminating relationship developed from the 9 and 11 spectral features was successfully validated using the 15 FTIR spectra comprising the prediction set for both the experimental and digitally blended data (see Figures 4.5 and 4.6). The plant-based edible oil samples used to prepare the adulterated EVOO mixtures comprising the prediction set were again excluded from the training set. The detection limit for canola oil in EVOO from the principal component score plot of the FTIR spectra is again 10%[4-10].

Table 4.5. Training and prediction set for experimental and blended data

	Number of spectra in training set/prediction set
EVOO	73/0
Canola	27/0
EVOO-canola	15/15
Total	115/15

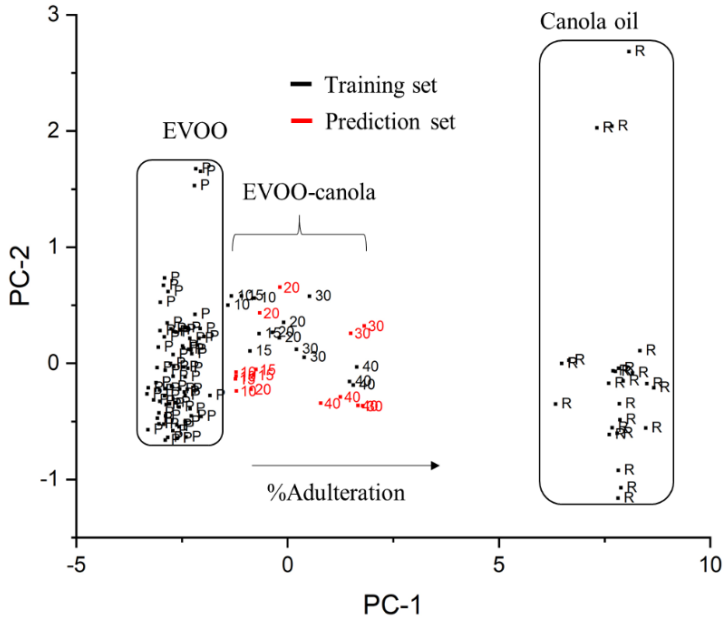

Figure 4.5. Plot of the two largest principal components of the 115 IR training set spectra (black) and the 9 spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, canola oil and EVOO-canola oil mixtures (10% canola oil to 40% oil). The total cumulative variance explained by the two largest principal components for the experimental data is 96.82%. The prediction set spectra are represented in red. P = EVOO, R = canola oil, 10 = 10% canola oil, 15 = 15% canola oil, 20 = 20% canola oil, 30 = 30% canola oil and 40 = 40% canola.

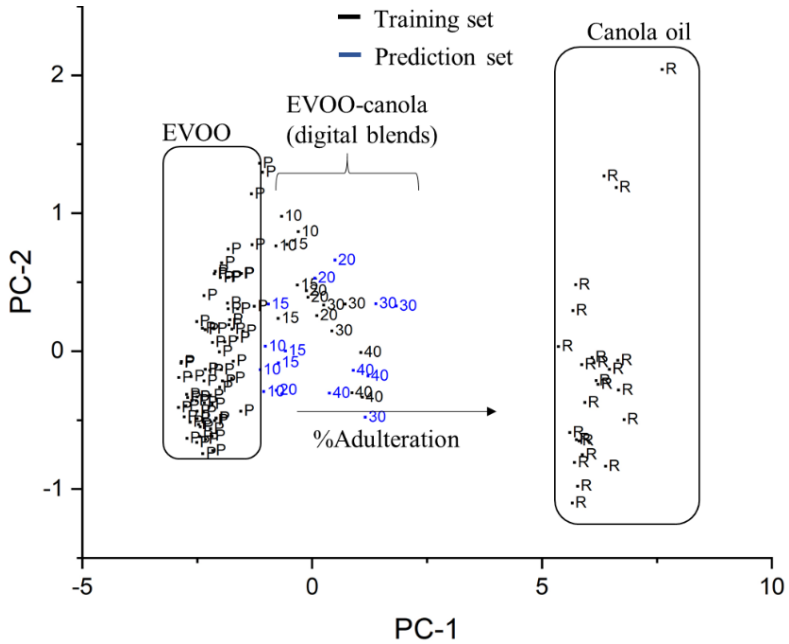

Figure 4.6. Plot of the two largest principal components of the 115 IR training set spectra (black) and the 11 spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, canola oil and EVOO-canola oil mixtures (10% canola oil to 40% oil). The total cumulative variance explained by the two largest principal components for the blended data is 95.62%. The prediction set spectra are represented in red. P = EVOO, R = canola oil, 10 = 10% canola oil, 15 = 15% canola oil, 20 = 20% canola oil, 30 = 30% canola oil and 40 = 40% canola.

Figures 4.7 and 4.8 show the principal component score plots of the 100 IR spectra comprising the training set (see Table 4.6) and the 5 and 25 spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, EVOO-almond oil, and almond oil. The first principal component does not appear to be well correlated to the amount of almond oil in the mixtures. Furthermore, several EVOO-almond oil samples in the training and prediction set for both the experimental and digitally blended data are not correctly classified. The absence of spectral features in the experimental and digitally blended data that can differentiate EVOO from EVOO adulterated with almond oil and the first principal component being weakly correlated to the amount of almond oil in the samples would indicate that EVOO and almond oil have similar IR spectra and would be difficult to discriminate by FTIR. Detecting adulteration of EVOO by almond oil would also be problematic. The concordance of the results for the experimental and digitally blended data for EVOO-corn, EVOO-canola, and EVOO-almond oil would suggest that we can discriminate two edible oils by variety using FTIR spectroscopy if the pure IR spectra of the two edible oils and their digitally mixed spectra can be discriminated in a ternary classification problem. For this comparison, the samples representing each edible oil must account for seasonal and batch variations within each supplier as well as variations between suppliers.

Table 4.6. Training and prediction set for experimental and blended data

	Number of spectra in training set/prediction set
EVOO	73/0
Almond	12/0
EVOO-almond	15/15
Total	100/15

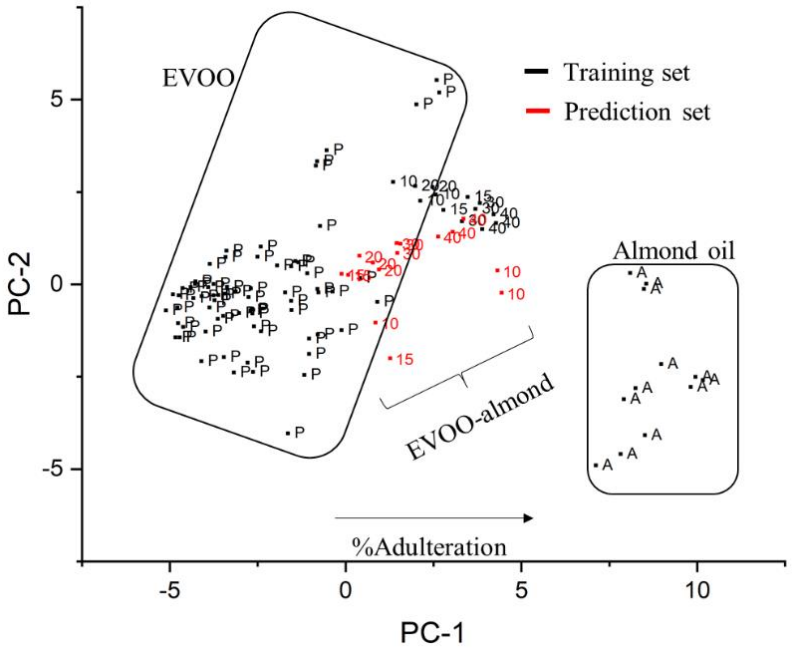

Figure 4.7. Plot of the two largest principal components of the 100 IR training set spectra (black) and the 5 spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, almond oil and EVOO-almond oil mixtures (10% almond oil to 40% oil). The total cumulative variance explained by the two largest principal components for the experimental data is 85.29%. The prediction set spectra are represented in red. P = EVOO, A = almond oil, 10 = 10% almond oil, 15 = 15% almond oil, 20 = 20% almond oil, 30 = 30% almond oil and 40 = 40% almond.

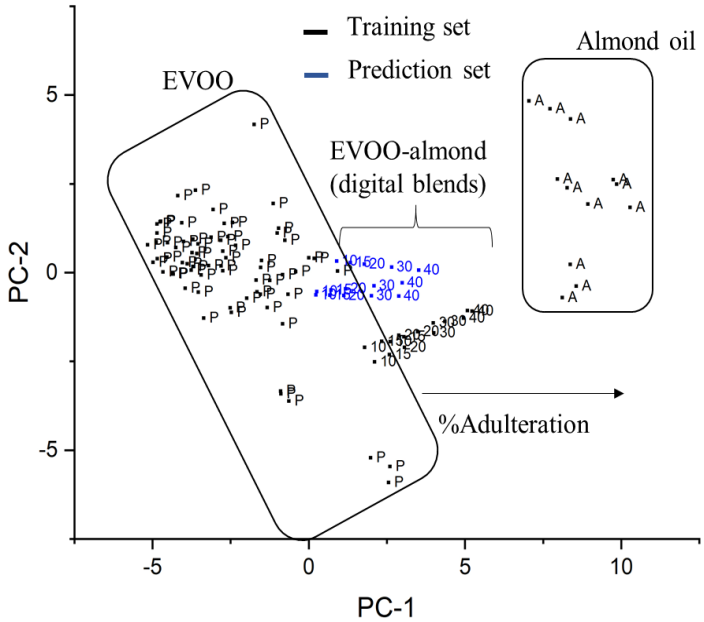

Figure 4.8. Plot of the two largest principal components of the 100 IR training set spectra (black) and the 25 spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, almond oil and EVOO-almond oil mixtures (10% almond oil to 40% oil). The total cumulative variance explained by the two largest principal components for the experimental data is 78.7%. The prediction set spectra are represented in blue. P = EVOO, A = almond oil, 10 = 10% almond oil, 15 = 15% almond oil, 20 = 20% almond oil, 30 = 30% almond oil and 40 = 40% almond.

4.5. Uncalibrated Adulterants

Variable selection is both critical to predictive modeling and general for edible oil identification. Selection of informative variables can impact most of the modeling done for the classification of edible oils by FTIR spectroscopy. However, a challenging aspect to this problem which has been largely ignored is encountering new contributions to the signal that we seek to model. The focus in the present study has been the selection of variables from a set of multivariate responses with the goal of making rapid estimates of class membership for a specific type of edible oil (e.g., EVOO). In the final phase of this study, our goal is to reduce the effects of uncalibrated contributors that will then help us to identify wavelengths that are useful for recognizing adulterants in edible oils. The pattern recognition GA using the PCKaNN fitness function generalized to allow for incorporation of model inference into the variable selection process has been evaluated in this context. To investigate the problem of uncalibrated interferents, two different classification models were developed and are described in detail below. The methodology used to detect uncalibrated interferents is similar to the methodology previously developed to assess the suitability of discriminating two edible oils using FTIR spectroscopy.

In the first problem to be discussed (see Table 4.7), the training set consisted of 144 IR spectra of EVOO, canola oil and digitally blended mixtures of EVOO and canola oil. The prediction set was comprised of digitally blended mixtures of EVOO and corn oil. Figure 4.9 shows a plot of the two largest principal components of the 144 training set spectra and the 14 spectral features identified by the pattern recognition GA. The discriminating relationship developed from the 14 spectral features was successfully validated using the 24 digitally blended FTIR spectra comprising the prediction set. Thus,

EVOO adulterated by corn oil could be recognized using a classifier developed for canola oil. Furthermore, the digitally blended data yielded similar results to experimentally generated data for the same training and prediction set (see Figure 4.10).

Table 4.7. EVOO-Canola Data Set

	Number of spectra in training set/prediction set
EVOO	73/0
Canola	27/0
EVOO-canola	44/0
EVOO-corn	0/24
Total	144/24

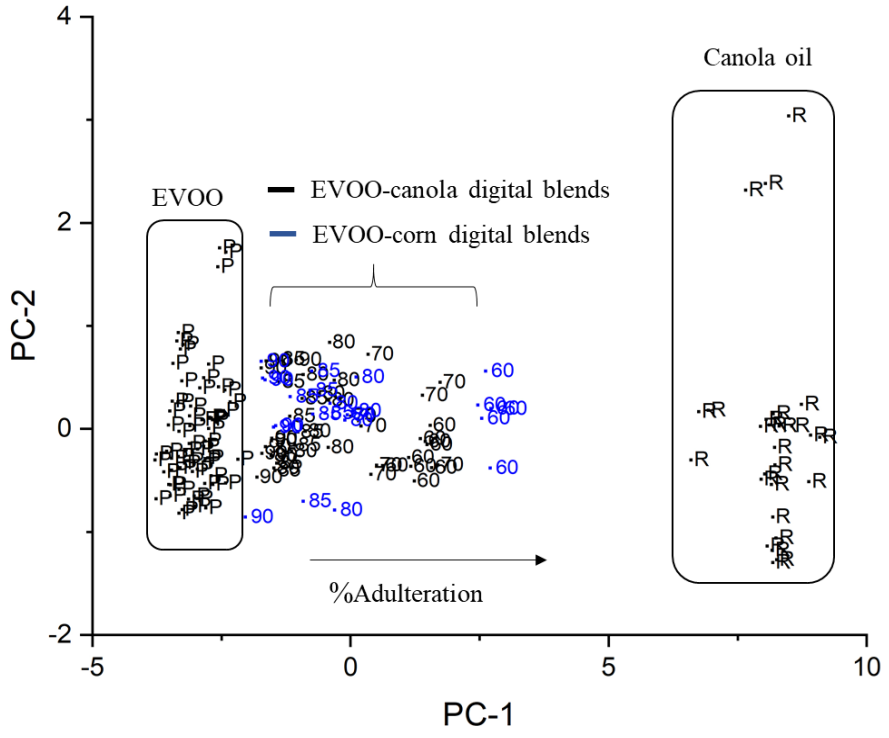

Figure 4.9. Plot of the two largest principal components of the 144 IR training set spectra (black) and the 14 spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, canola oil and the digital blends of EVOO and canola. The prediction set samples (blue) are digital blends of EVOO and corn oil spectra. The total cumulative variance explained by the two largest principal components is 98.2%.

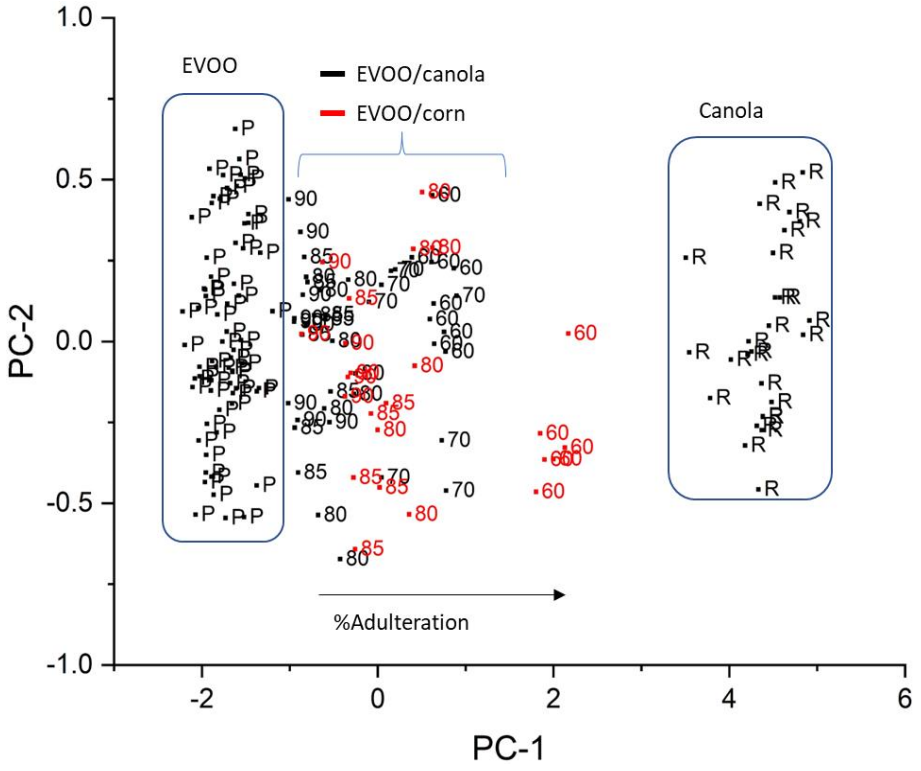

Figure 4.10. Plot of the two largest principal components of the 144 IR training set spectra (black) and the five spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, canola oil and EVOO-canola oil. The prediction set samples (red) are EVOO-corn mixtures prepared using a digital pipette. The total cumulative variance explained by the two largest principal components is 96.9%.

In the second problem (see Table 4.8), the training set consisted of 124 IR spectra of EVOO, corn oil and digitally blended mixtures of EVOO and corn oil. The prediction set was comprised of digitally blended mixtures of EVOO and canola oil. Figure 4.11 shows a plot of the two largest principal components of the 112 training set spectra and the 18 spectral features identified by the pattern recognition GA. The discriminating relationship developed from the 18 spectral features was successfully validated using the 15 digitally blended FTIR spectra of the EVOO-canola oil mixtures comprising the prediction set. EVOO adulterated by canola oil could be recognized using a classifier developed for corn oil. Furthermore, the digitally blended data yielded similar results to experimentally generated data for the same training and prediction set (see Figure 4.12).

Table 4.8. EVOO-corn data set

	Number of spectra in training set/prediction set
EVOO	73/0
Corn	27/0
EVOO-corn	12/0
EVOO-canola	0/15
Total	112/15

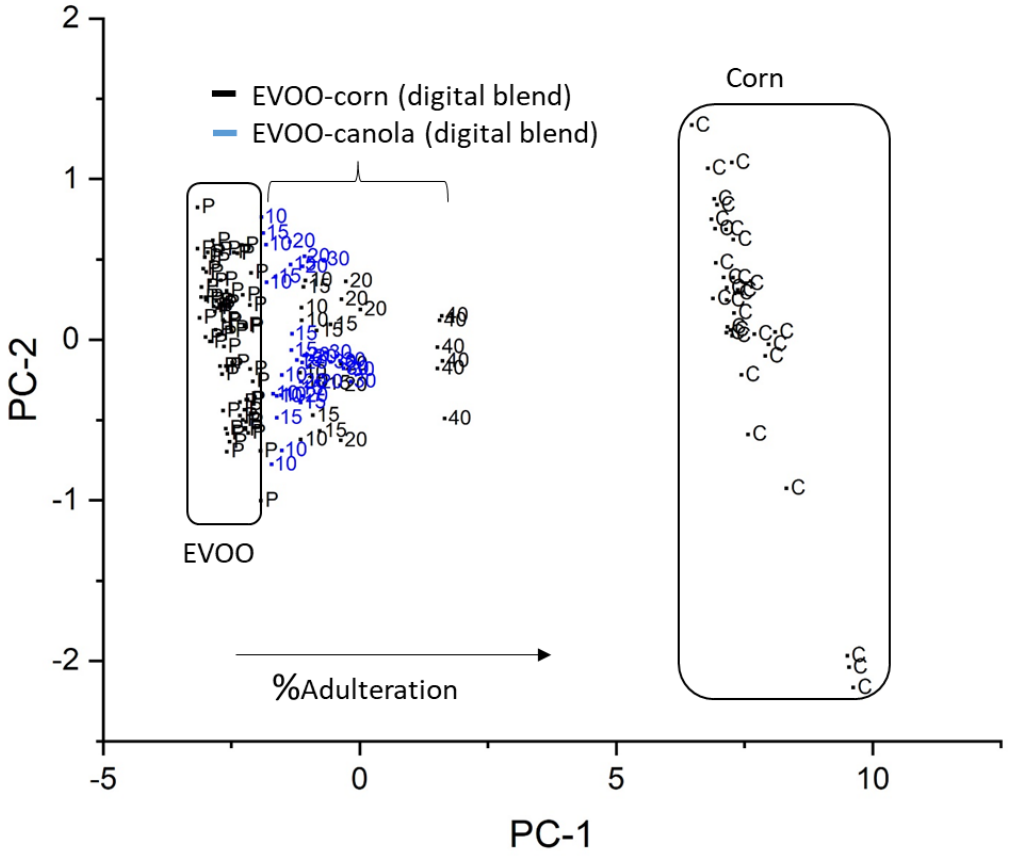

Figure 4.11. Plot of the two largest principal components of the 112 IR training set spectra (black) and the 18 spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, corn oil and digital mixtures of EVOO-corn oil spectra. The prediction set (blue) are digital blends of EVOO and corn oil spectra. The total cumulative variance explained by the two largest principal components is 96.44%.

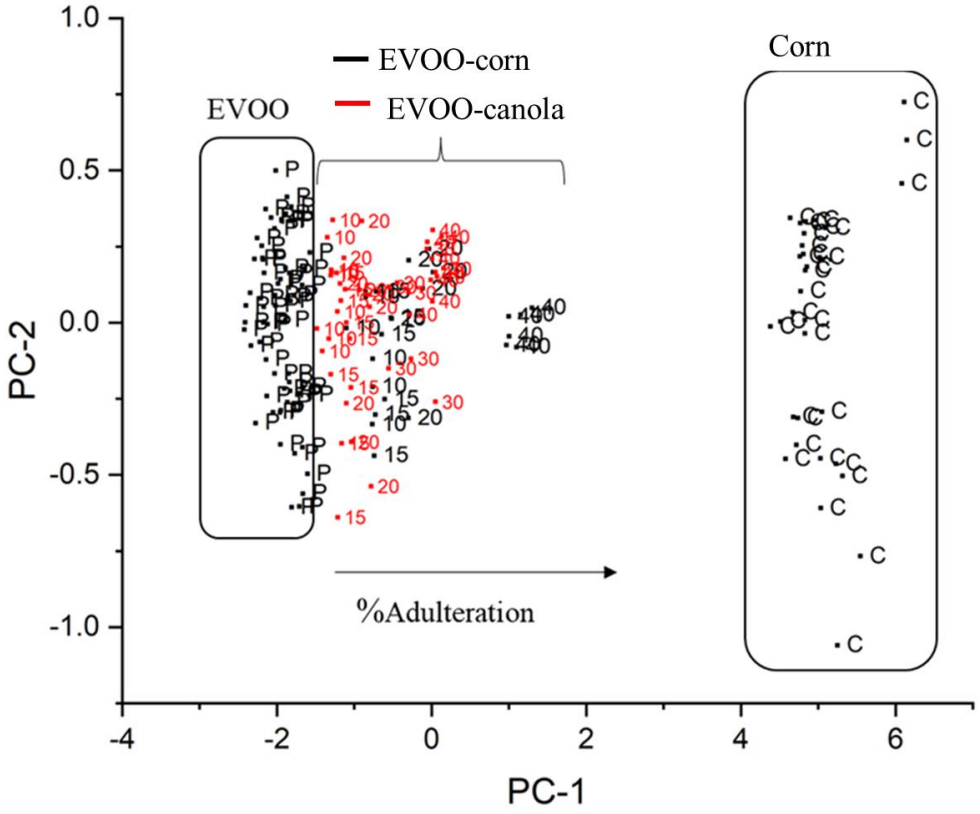

Figure 4.12. Plot of the two largest principal components of the 118 IR training set spectra (black) and the 11 spectral features identified by the pattern recognition GA for the three-way classification problem: EVOO, corn oil and EVOO-corn oil mixtures. The prediction set samples (red) are EVOO-canola oil mixtures prepared using a digital pipette. The total cumulative variance explained by the two largest principal components is 97.43%.

4.6. Conclusions

In this chapter, a basic methodology for assessing the suitability of discriminating two edible oils from their FTIR spectra was described. The FTIR spectra of the pure edible oils were compared to the FTIR spectra of their mixtures using library spectra and digitally blended data. For this comparison, the samples representing each edible oil must account for seasonal and batch variations within each supplier as well as variations between suppliers. Each comparison was formulated as a three-way classification problem. The edible oil mixtures used for the training and validation sets for both the experimental and digitally blended data should span a large concentration range. For each comparison, it is

assumed that the IR spectra of the mixture can be represented by the IR spectra of the two oils, with the weights of the constituents defining the mixing proportion of each edible oil that comprises the mixture. If the IR spectra of the two edible oils and their digitally blended mixtures are differentiable, then differences between the IR spectra of these two edible oils will be of sufficient magnitude to ensure that a reliable classification of these two edible oils by FTIR spectroscopy can be obtained. Using this approach, the feasibility of authenticating edible oils such as EVOO directly from library spectra has been demonstrated. Furthermore, the suitability of authenticating edible oils containing uncalibrated interferents (i.e., adulterants) was also demonstrated using the edible oil spectral library.

References

4-1. A. Patel, U. Rova, P. Christakopoulos, L. Matsakas, Introduction to essential fatty acids, John Wiley & Sons, Hoboken, NJ, **2022**, pp. 1-22.

4-2. J. Orsavova, L. Misurcova, J. V. Ambrozova, R. Vicha, J. Mlcek. Fatty acids composition of vegetable oils and its contribution to dietary energy intake and dependence of cardiovascular mortality on dietary intake of fatty acids. *Int. J. Mol. Sci*. **2015**; 16 (6):12871–12890.

4-3. A. H. Lichtenstein, Fats and Oils in Encyclopedia of Human Nutrition, 3rd edition, B. Caballero (Ed.) Academic Press, Waltham, MA **2013**: pp. 201-208.

4-4. K. Chowdhury, L.A. Banu, S. A. Khan, and A. Latif. Studies on the fatty acid composition of edible oil, *Bangladesh J. Sci. Ind. Res.* **2007**; 42: 311-316.

4-5. J. Spink, and D. C. Moyer, Defining the public health threat of food fraud. *J. Food Sci.*, **2011**; 76 (9): R157-R163.

4-6. J. C. Moore, J. Spink, and M. Lipp, Development and application of a database of food ingredient fraud and economically motivated adulteration from 1980 to 2010. *J. Food Sci.*, **2012**; 77(4): R118-R126.

4-7. M. Meloun, J. Militky, and M. Forina, Chemometrics for analytical chemistry. volume 1: PC aided statistical data analysis, Ellis Horwood, NY, **1992**, pp. 138-139.

4-8. B.G.M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. DeJong, P. J. Lewi, and J. Smeyers-Verbeke, Handbook of chemometrics and qualimetrics, Elsevier, Amsterdam, **1998**, pp. 310.

4-9. Vlachos, N.; Skopelitis, Y.; Psaroudaki, M.; Konstantinidou, V.; Chatzilazarou, A.; Tegou, E. Applications of Fourier transform-infrared spectroscopy to edible oils. *Analytica Chimica Acta* **2006,** *573-574*, 459-465.

4-10. Gurdeniz, G; Ozen, B. Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data. *Food Chem.* **2009**, 116, 519 - 525.

CHAPTER V

Summary

In the preceding chapters, the use of IR spectroscopic data to classify a diverse set of edible oils via principal component analysis and the use of genetic algorithms to perform wavelength selection was described. The study was undertaken to show that for practical applications, when one is attempting to differentiate edible oils by type (e.g., extra virgin olive oil versus canola oil), it is necessary to account for all variations within a source including seasonal variations by purchasing edible oils under different brand names and over several years. This was recognized at the start of the investigation as this was the working hypothesis. The goal was to capture as much variation as possible to simulate real world conditions. The novelty of the study arises from the incorporation of supplier-to-supplier variation as well as seasonal variation within a supplier in the research design. Supplier to supplier variation and possibly seasonal variation within a supplier can be greater than the within supplier variation. varieties of the edible oils surveyed. This is the first time that so many different edible oils and commercially available brands have been classified simultaneously. By comparison, previous studies, which were restricted to five or six varieties of edible oils and relied on only one brand and often a single bottle from

that brand. for each type of edible oil, provided an overly optimistic estimate of the ability to classify edible oils as to variety or to detect low levels of adulterants in edible oils using FTIR spectroscopy.

As it is not possible to simultaneously classify 20 varieties of edible oil by a single classifier, a hierarchical classification of the FTIR spectra of the edible oils was undertaken. The twenty edible oil varieties were divided into four groups as determined by hierarchical clustering and principal component analysis of the average IR spectrum of each of the twenty edible oil varieties. The successful classification of the twenty oil varieties into four groups was demonstrated using a pattern recognition GA, thereby supporting the conclusions about the data from the cluster analysis. Differentiation of edible oils within each group based on the classification results obtained by the pattern recognition GA as well as the differences in the chemical composition of each edible oil within a group (e.g., monounsaturated fatty acids and linolenic acid content) suggest that edible oils in the same group are more difficult to discriminate than edible oils in different groups. The three adulteration studies described in Chapter 3 are consistent with this premise. More importantly, the findings that some large subsets of edible oils can be parsed using the variable selection methodology and that validation sample subsets can be correctly classified support model validity, and that it is not possible to develop a single classifier that can separate all possible edible oil types from each other are also supported by the evidence presented. Finally, the hierarchical classification scheme proposed in this dissertation enables an analyst to determine whether a specific problem in the detection of adulterants (e.g., adulteration of a more expensive edible oil by blending with a less expensive oil) can be solved using FTIR spectroscopy.

The IR spectra of the different varieties of edible oils surveyed in this study are very similar. For example, the most dissimilar IR spectral pair in this data set has a hit quality index (HQI) value of 92%. For this reason, variable selection is important. To identify wavelengths characteristic of each edible oil variety, a pattern recognition GA was applied to the IR spectral data. The approach underlying variable selection by the pattern recognition GA is based on a simple idea - identify the smallest set of variables (i.e., absorbances at specific wavelengths) that optimize the separation of the edible oils in a plot of the two or three largest principal components of the data. Because principal components maximize variance, the bulk of the information encoded by these variables is about differences between the edible oils in the training set. Using this approach for variable selection, an eigenvector projection of the data is formulated that discriminates between the edible oils in the data set by maximizing the ratio of between to within group variance through selection of the appropriate variables. Although a principal component score plot is not a sharp knife for discrimination, if a score plot shows clustering on the basis of class membership of the samples, then our experience (as well as the experience of other workers) is that one will be able to predict robustly using this set of wavelengths.

The authenticity of edible oils has a very direct impact on human health. In this dissertation, FTIR spectroscopy has been combined with pattern recognition to establish a rapid method for the discrimination of edible oils. From the experimental results presented in the studies described in Chapters 3 and 4, the proposed methodology has been proven to be successful.

Isio Sota-Uba

Candidate for the Degree of

Doctor of Philosophy

Dissertation: AUTHENTICATION OF EDIBLE OILS USING FOURIER TRANSFORM INFRARED SPECTROSCOPY AND PATTERN RECOGNITION METHODS

Major Field: Chemistry

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Chemistry at Oklahoma State University, Stillwater, Oklahoma in July 29, 2022.

Completed the requirements for the Master of Science in Chemistry at University of Ibadan, Ibadan, Oyo/Nigeria in 2012.

Completed the requirements for the Bachelor of Science in Chemistry at University of Ibadan, Ibadan, Oyo/Nigeria in 2007.

Previous Work Experience:

Chemistry teacher at Al-Azim College in Lagos, Nigeria from 2008 to 2011.

Technologist at Multi-disciplinary Central Research Laboratory, Ibadan, Nigeria from 2012 to 2013.

Chemistry department teaching and research assistant at Oklahoma State University from 2017 to 2022.

Current Position:

Chemometrician, PPG Industries, Springdale, PA from August 23, 2022.