OIL SPECTRA ANALYSIS AND ADULTERATION DETECTION IN MIR SPECTROSCOPY DATA USING MACHINE LEARNING

V.T.BALA MURUGAN[1], S.M.SANGEETHA[2], R.RESHMA[3], R.SIVA RANI[4]

[1] Assistant Professor, EIE Department, Bannari Amman Institute of Technology, Tamilnadu, India
[2] BE Student, EIE Department, Bannari Amman Institute of Technology, Tamilnadu, India
[3] BE Student, EIE Department, Bannari Amman Institute of Technology, Tamilnadu, India
[4] BE Student, EIE Department, Bannari Amman Institute of Technology, Tamilnadu, India

---

Abstract -In daily eating regimenvegetable oils have a major commitment as cooking oil, frying oil, salad oil or in nourishment items plans. Oils are more essential for nutritious perspectives. Some edible oils, for example, castor oil, coconut oil, sesame oil, and groundnut oil are so costly which makes enticing to adulterate them with other lower value vegetable oils and fats to accomplish more benefits. Adulteration of oil with other vegetable oils and fats can lead to many health issues like cardio-vascular disease, epidemic dropsy, capillaries dilation and causing millions of death annually. Thus to distinguish the contamination it is more important to utilize the progressed and reasonable strategies. There is a need for sustenance related association to create and use dependable techniques to recognize such corruptions, which can make purchasers and markets progressively sure on authenticity and virtue of edible oils. It is proposed to provide low cost oil adulteration detection method for the benefit of consumers. The techniques involve including MIR spectroscopy and Machine Learning algorithms like Principle Component Analysis and K-Means Clustering.

***Key Words*:**Oil Adulteration, MIR Spectroscopy, Machine Learning Algorithms,Principle Component Analysis, K-Means Clustering.

1.INTRODUCTION

Adulteration in food products means the addition of prohibited substance either partially or entirely due to the state of financial gain or lack of hygienic processing and storage conditions that lead to the consumer being cheated. In our eating standard, vegetable oils and fats plays an important role. Oils are more essential for nutritious perspectives. Edible oils are food substance which produced entirely or in portion for human consumption. Because of decreased net regional supply, India is an importer of edible oil. The total edible oil production in India in 2015-16 was 25.3 million tons, and total edible oil area was 26.13 million hectare. Total production of edible oils from 28.05 million hectares was recorded in 2013-14 which was 32.75 million tons. India imported 148.2 tons of edible oils in 2015-16, and 86.37 tons of net domestic supply. Due to their increased demand in the national and international markets, high price oil is adulterated with the low price oil. Oil adulteration causes many chronic diseases. In the human diet, the intake of adulterated oils and transfats has adverse health effects including cardiovascular disease, causing millions of death each year. For the sake of consumers there is an immediate need for authentication and prevention of adulteration. Thus, it is more important to utilize progressed and reasonable strategies to distinguish the adulterated oil. Adulteration can cause few issues in eatable oils application. There is a need for sustenance related association to create and use dependable techniques to recognize such corruptions, which can make purchasers and markets progressively sure on authenticity and virtue of edible oils.

Adulteration is more advanced today. Therefore, efficient and suitable methods are needed to detect adulteration. Various methods are available for detecting adulteration in oil. Making these approaches fast, efficient and cost effective is a real challenge. In general, it is no longer practical to detect adulteration using the physical properties such as the calorimetric reaction based method as well as viscosity density refractive index measurement computer vision method, saponification value, iodine value, acid value and specific gravity. There are also different methods for detecting adulteration such as liquid chromatography and gas chromatography. Those properties are well designed to mask the adulteration in oil. Edible oils and fats are composed of major and minor components. Triacylglycerol is the major component in edible oils and fats. Minor components are sterols, carotenoids, tocopherol, chlorophyll and other minor compounds.

To detect adulteration of edible oils and fats, both the major and minor components can be used as a detection tool. Because each oil may have specific components at a known level, its presence and quantities should be considered as a tool for detection. Therefore, a cost-effective and accurate method is required. The development of non-destructive technologies that are capable of monitoring in real-time assessment is the significant interest. In this, the data is obtained by mid infrared spectroscopy and the adulteration is detected by using machine learning techniques based on k-means clustering and principal component analysis algorithm.

2.METHODOLOGY

2.1Mir Spectroscopy

Mid-infrared spectroscopy is a very significant and commonly used sample characterization and analytical

method. The sample is a gas, liquid, paste-like, or solid. This approach is commonly used in applications involving qualitative analysis, including either functional group or structural information of a sample and its composition or identification of a material. MIR spectroscopy depends on the absorption of light. Incident infrared light onto a sample induces vibration of biochemical bonds. MIR spectroscopy is a vibrational spectroscopy technique that identifies chemicals based on the mid-infrared region (400 to 4000 cm^{-1}) interaction of molecules with electromagnetic radiation. Infrared spectroscopy identifies chemicals that are based on molecule absorbing specific wavelength of mid-infrared light.

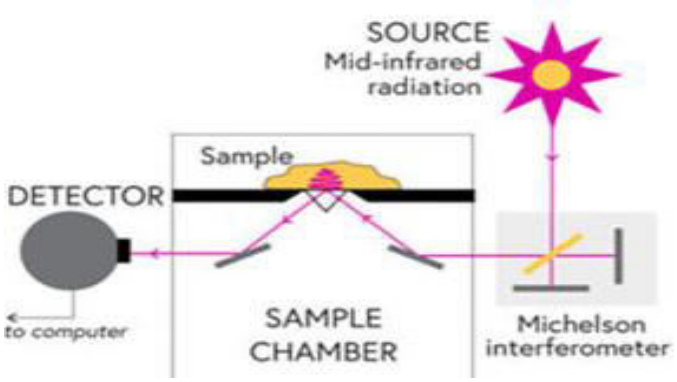

Fig-1: Block Diagram Of Mir Spectroscopy

In mid-infrared light, the molecular absorbance induce molecular rotations and vibrations and are categorized by functional chemical groups. This causes the absorption of specific amounts of energy from the incident IR beam by vibrational molecular modes, reducing the intensity of the initially detected infrared beam. The energy difference between incident and detected IR radiation results in complex interferogram which is deconvoluted. It distinguishes the individual wavelength of the measured IR range into wavelength of the materials, creating a spectrum of wave numbers. The resulting MIR spectrum produced is characteristic for a given molecule.

2.1.1 Beer Lambert's Law

The Beer-Lambert law gives the relationship between the attenuation of light and the properties of the same substance.

The relationship can be expressed as

$$A = \varepsilon cl$$

Where A is absorbance (no units)

ε is the molar absorptivity (L mol^{-1} cm^{-1})

l is the sample's path length (cm)

c is the compound concentration (mol L^{-1})

2.1.2 Attenuated Total Reflection

Attenuated Total Reflection (ATR) is shown in the figure 2. is based on the principle of total internal reflection. It produces an evanescent wave penetrating electromagnetic field whose intensity quickly decays as it moves away from its source. The depth of the penetration into sample is in the range of 0.5 and 2 micrometers with accurate value calculated by wavelength, the incident angle and the refractive indices for the crystal and their medium is being tested. It is the distance of the point at the amplitude of the evanescent wave has been reduced to thirty seven percent of its highest value. The reflection that altered by changing the angle of incidence. The beam interacts and absorbs energy from the sample. So the intensity of the reflected wave reaching the detector is reduced.This evanescent effect works best when the crystal is made of an optical material with a higher refractive index than the sample being studied. Otherwise light is lost to the pattern. In the case of a liquid pattern, pouring a little quantity over the surface of the crystal is enough. In the case of a stable pattern, samples are firmly clamped to make contact and to eliminate with trapped air that would reduce signal depth. The signal to noise ratio acquired depends on the wide variety of reflections however also on the total length of the optical path which reduce the intensity.

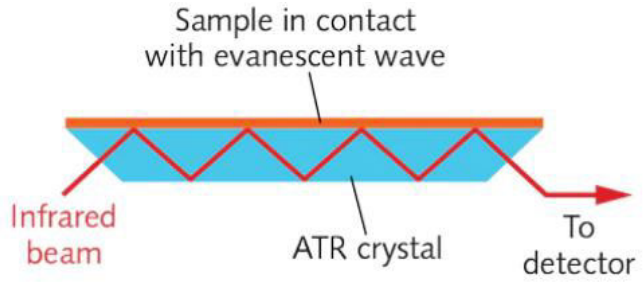

Fig-2: Attenuated Total Reflection

An Attenuated Total Reflection (ATR-MIR) is the most popular sampling technique used in most laboratories, allowing for quick analysis of solid and liquid materials. MIR spectra can be searched against large reference spectral databases, making this technique a powerful instrument to identify unknown chemicals. ATR generally allows little or no preparation of samples which greatly accelerates sample analysis. It allows the IR beam penetrate into the sample in very thin path length and depth. It is useful for samples which are too thick to be examined during transmission and those which absorb radiation strongly.

2.1.3 Absorbance And Transmittance

Absorbance is a measure of the amount of light with a defined wavelength which prevents a given material from going through it. Transmittance is defined as the ratio of the transmitted intensity over the light intensity.

$$T = \frac{I}{I_0}$$

Where T is the transmittance (no unit)

I is the incident light intensity (candela)

I_0 is the reference light intensity (candela)

2.1.4 Mirror Software Specifications

The mirror software is used for storing the data in the excel sheet of wavelength and intensity of the oil. It has various specifications and their values are fixed. Some of the specifications are as follows.

Sensor and measurement - 200

Pulse rate (Hz) - 8

Light intensity - 100

Warm up scans - 40

Delay should be set as zero seconds.

2.2 Principal Component Analysis

The principal component analysis is unsupervised method in machine learning. It is used to make analyze the data easily and visualize the difference between the data values. It is a quantitative method that uses an orthogonal transformation to translate a set of data of potentially correlated variables into a set of values of sequentially uncorrelated variables. It is a sensitive method. In this, the first principal component and the succeeding component have the greatest possible variance below the constraint which is orthogonal to the preceding components. The following vectors are an uncorrelated orthogonal basis set. It is the simplest of true multivariate analyzes dependent on the eigenvector. This procedure can often be considered to show the data's internal structure that describes the data variance. In a high-dimensional data space, a set of coordinates of multivariate dataset is visualized and it provides the lower dimensional image to the user. The first few principal components are achieved so that the transformed data dimension is reduced. PCA is closely associated with the factor analysis. It is a dimensionality reduction technique that allows identifying the correlations and patterning in a data set so that it can be transformed into a significantly lower dimensional data set without loss of any important information. PCA is intended to determine the patterns and correlations between different features in the data set. A final decision is made to reduce the dimensions of the data in such a way that the significant data is still retained in order to find a string correlation between different variables. Such a process is very necessary in solving complex data-driven issues involving the use of high-dimensional data sets. PCA is achieved through a series of steps. The following figure 3.3 represents the flow chart of principal component analysis

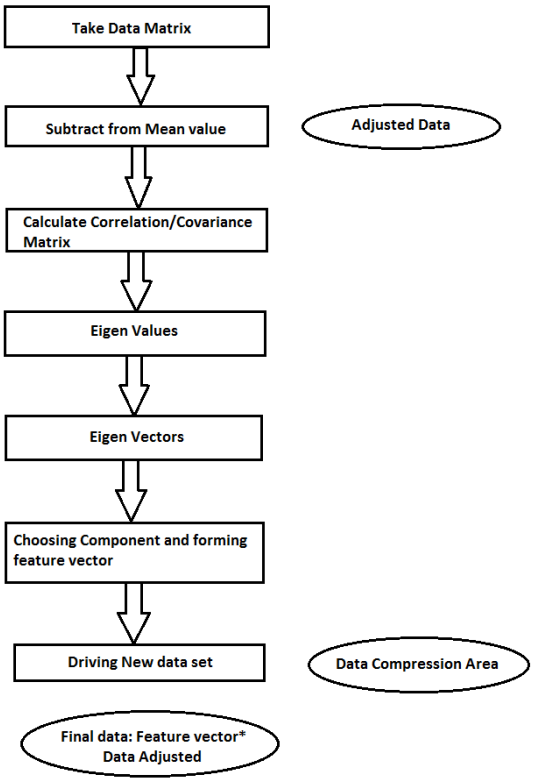

Fig-3: Flow Chart of Principal Component Analysis

2.2.1 Standardization of The Data

Standardization is about subtract all the values in the column from the mean. Then variance should be maximized because PCA will load on large variance. So, to maximize the variance standardization should be done. It produces the dataset whose mean is zero. It can be calculated by

$$z = \frac{variable\ value - mean}{standard\ deviation}$$

2.2.2 Computing The Covariance Matrix

The correlation between the different variables in the data set is expressed by a covariance matrix. Identifying heavily dependent variables is essential since they contain biased and obsolete information which reduces the overall performance. Each entry in the matrix represents covariance of the corresponding variables. It can be calculated by,

$$\begin{bmatrix} cov(a,a) & cov(a,b) \\ cov(b,a) & cov(b,b) \end{bmatrix}$$

2.2.3 Calculating The Eigenvectors And Eigenvalues

Eigenvectors and eigenvalues are the mathematical construction to be calculated. To determine the principal components of the data set they are calculated from the covariance matrix.

To calculate Eigen values,

$$\det(\lambda I - A) = 0$$

Where, I - Identity matrix
det - Determinant of the matrix
λ – Eigen value of the matrix
To calculate Eigen vectors,

$$(\lambda I - A)\, v = 0$$

Where, v is Eigen vector.

2.2.4 Computing the Principal Components

This is a dimension reduction part. If there are N variables, it should have N Eigen values and N Eigen vectors. Eigen vector corresponding to highest Eigen value is the principle component. The principal components are the set of new variables obtained from the initial set of variables. The principal components are calculated in such a way that the variables which are obtained newly are highly significant and independent of each other. They compress and possess most of the valuable data that was distributed between the initial variables. After the eigenvector and eigenvalues are calculated, they are arranged in a descending order where the highest eigenvalue of the eigenvector is most significant and thus it constitutes the first principal component. To reduce the dimension of the data the principal component of lower significances can be removed. Then the principal components are computed to form the matrix known as the feature matrix which contains all the important data variables that have the maximum data information.

$$\text{Feature matrix} = (eig_1, eig_2)$$

2.2.5 Reducing The Dimensions of The Data Set

The original data is re-arranged with the calculated principal components which constitutes the data set's most significant and maximum information. The newly formed principal components are replaced with the original data axis by multiply the transpose of the original data set by the transpose of the feature vector that has been obtained.

$$\text{New Data} = \text{Feature matrix}^T \times \text{Scaled data}^T$$

Where, New data - Matrix consisting of the principal components
Feature matrix - Matrix formed using Eigen vectors
Scales data - Scaled version of original dataset

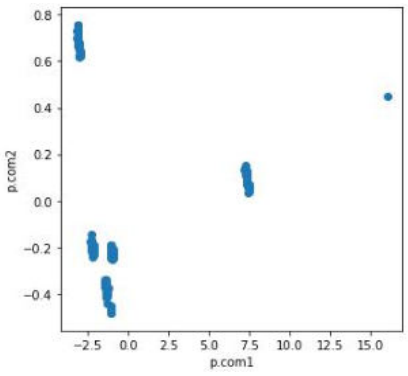

Fig-4:PCA Response for Reduced Data

2.3 K-Means Clustering

K-Means is an unsupervised Machine Learning algorithm. It is a vector Quantization. Clustering is one of the popular methods of exploratory data analysis used to get fixed pattern about the data's structure. This appears to be defined as the assignment of identifying subgroups in data set information with the end goal that the data set information with the end goal that the data focus in a specific subgroup is essentially the same as when data focus in different groups is completely different. When a set of object is given, they are placed into the similarity based group. Based on some similarity, clustering algorithm used to classify natural data groups. It locates the data points on the centroid. The algorithm assesses the distance between each point from the cluster's centroid to perform successful clustering. Clustering is used to determine the underlying grouping of unlabeled data in a set of data. Because of its ubiquitousness, they are called as k-means algorithm and also referred as Lloyd's algorithm in the computer network. It also called as naive k-means due to faster alternatives are exist there. K-means algorithm is widely used in many areas ranging from unsupervised neural network learning, pattern recognition, classification analysis, artificial intelligence, machine vision and many others. The aim of K-means clustering is to analysis the high quality clusters. To find similar type of data's are formed in one cluster or otherwise formed as inter clusters.

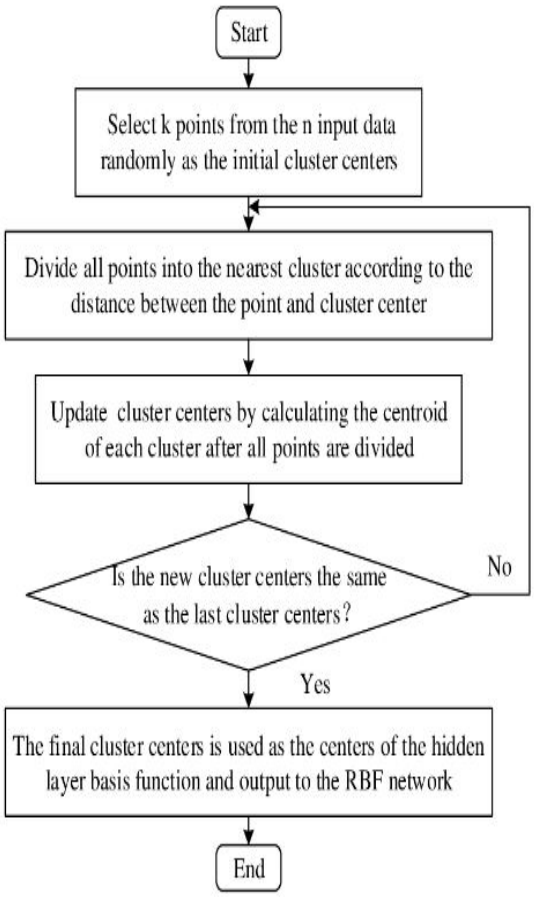

Figure-5: Flow Chart of K-Means Clustering

The algorithm is iterative used to group the data to the predefined, independent, non-overlapping subgroups in which the data point belongs to a single group only. It makes the data points between clusters. This assigns data to a group in such a way that the sum of the squared distance among the datasets and the centroid of the cluster is at the minimum. The variation is less in clusters, the more similar the data points are in the same cluster. It is useful for the identification of undirected information and relatively simple.

The operation of the k means algorithm is as follows:

1. Select the cluster centers as 'c' randomly.
2. Calculate the distance between center of the cluster and the data point.
3. Keep iterating until centroids don't change i.e., assigning the cluster data points is not changing.
4. Measure the sum of the square distance between all centroids and data points.
5. Assign each data set to the nearest cluster.
6. Calculate the centroids for the clusters by taking all points that belong to each cluster on average.

3.RESULTS

The given data is reduced using Principle Component Analysis algorithm. The data is then given to K-Means Clustering Algorithm and it is clustered as adulterated and unadulterated oil. The following figure 4.1 represents the clustering of five pure oil.

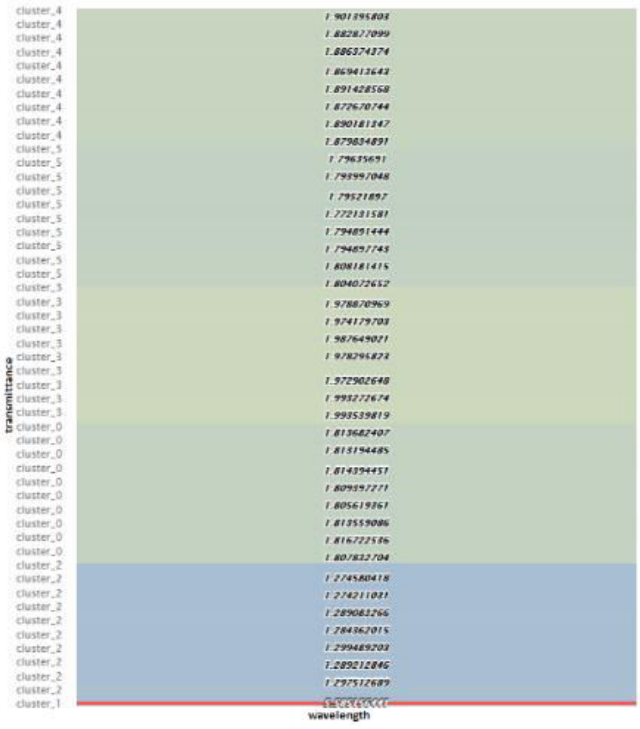

Fig-5: Clustering of Five Pure Oil

4.CONCLUSION

In this project the five different oil data is taken and the data is processed .The processed data is analysed by using Principal component analysis in which the dimension of the dataset is reduced and the processed data is clustered by using Kmeans clustering algorithm.Here the pure oil is clustered into five different clusters.

5.FUTURE SCOPE

The adulteration in the oil is detected by using Principal component analysis algorithm and K-means clustering algorithm. When the PCA is applied, the large set of data is reduced. Then the reduced data is clustered by means of K-means clustering algorithm to determine whether the oil is adulterated or not. This technique also implemented to provide the low cost handheld oil adulteration detection device. This device will display the percentage of the adulterated and unadulterated content present in the oil.

REFERENCES

1. Downey, G., McIntyre, P., & Davies, A. N. (2002). Detecting and quantifying
sunflower oil adulteration in extra virgin olive oils from the eastern
mediterranean by visible and near-infrared spectroscopy. Journal of agri-
cultural and food chemistry, 50 , 5520–5525.

2. Christy, A. A., Kasemsumran, S., Du, Y., & Ozaki, Y. (2004). The detection
and quantification of adulteration in olive oil by near-infrared spectroscopy
andchemometrics. Analytical Sciences, 20 , 935–940.

3.Catharino RR, Haddad R, Cabrini LG, Cunha IBS, Sawaya ACHF, Eberlin MN. Characterization of vegetable oils by electrospray ionization mass spectrometry fingerprinting: Classification, quality, adulteration, and aging. American Chemical Society Analytical Chemistry, 2005; 77:7429-7433.

4. Fragaki, G., Spyros, A., Siragakis, G., Salivaras, E., & Dais, P. (2005). Detection of extra virgin olive oil adulteration with lampante olive oil and refined olive oil using nuclear magnetic resonance spectroscopy and multivariate statistical analysis. Journal of agricultural and food chemistry, 53 , 2810–2816.

5.Gurdeniz, G., &Ozen, B. (2009). Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data. Food Chemistry, 116 , 519–525.

6. Zou MQ, Zhang XF, Qi XH, Ma HL, Dong Y, Liu CW et al. Rapid authentication of olive oil adulteration by Raman spectrometry. Journal of agricultural and food chemistry. 2009; 57(14):6001-6006.

7.Gromadzka J, Wardencki W. Trends in edible vegetable oils analysis. Part B. Application of different analytical techniques. Journal of Food and Nutrition Sciences. 2011; 61(2):89-99.

8.Cataldo, A., Piuzzi, E., Cannazza, G., & Benedetto, E. D. (2012). Classificationand adulteration control of vegetable oils based on microwavereflectometry analysis. Journal of Food Engineering, 112 , 338 – 345.

9. Huo, Q., Jin, X.-B., & Zhang, H. (2012). Multi-label classification for oil authentication.In 9th International Conference on Fuzzy Systems and Knowledge Discovery (pp. 711–714).

10. Amereih S, Barghouthi Z, Marowan O. Detection and quantification of adulteration in olive oil using a uv-spectrophotometric method. Palestine Technical University Research Journal. 2014; 2:l14-19.

11. Zhang L, Li P, Sun X, Wang X, Xu B, Ma F et al. Classification and adulteration detection of vegetable oils based on fatty acid profiles. Journal of Agricultural and Food Chemistry.2014; 62:8745-8751.

12. Damirchi SA, Torbati M. Adulterations in some edible oils and fats and their detection methods. Journal of food quality and hazard control.2015; 2:38-44.

13. Tony George, Elizabeth Rufus, Zachariah C. Alex (2017). Artificial Neural Network Based Ultrasonic Sensor System For Detection Of Adulteration In Edible Oil. Journal of Engineering Science and Technology Vol. 12, No. 6 (2017) 1568-1579.

14.Neha S. Dhande, Rupesh D. Sushir (2018). Detection and Estimation of Adulteration in Oil Sample Using Digital Image Processing. IJSRST, Volume 4, Issue 2, ISSN: 2395-6011

15.S.A.Antora, M.N.Hossain, M. Rahman, M. A. Alim and M. Kamruzzaman. Detection of Adulteration in Edible Oil Using FT-IR Spectroscopy and Machine Learning. International Journal of Biochemistry Research & Review 26(1): 1-14, 2019; Article no.IJBCRR.49028 ,ISSN: 2231-086X, NLM ID: 101654445.