Application of Partial Least Squares Discriminant Analysis for Discrimination of Palm Oil

Mas Ezatul Nadia Mohd Ruah[1], Nor Fazila Rasaruddin[1], Sim Siong Fong[2] and Mohd Zuli Jaafar[1,3]

[1]Universiti Teknologi MARA, Kampus Kuala Pilah, 72000 Kuala Pilah, Negeri Sembilan, Malaysia
[2]Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[3]Faculty of Applied Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
Email: ezatul_n@yahoo.com

ABSTRACT

This paper outlines the application of chemometrics and pattern recognition tools to classify palm oil using Fourier Transform Mid Infrared spectroscopy (FT-MIR). FT-MIR spectroscopy is used as an effective analytical tool in order to categorise the oil into the category of unused palm oil and used palm oil for frying. The samples used in this study consist of 28 types of pure palm oil, and 28 types of frying palm oils. FT-MIR spectral was obtained in absorbance mode at the spectral range from 650 cm^{-1} to 4000 cm^{-1} using FT-MIR-ATR sample handling. The aim of this work is to develop fast method in discriminating the palm oil by implementing Partial Least Square Discriminant Analysis (PLS-DA), Learning Vector Quantisation (LVQ) and Support Vector Machine (SVM). Raw FT-MIR spectra were subjected to Savitzky-Golay smoothing and standardised before developing the classification models. The classification model was validated by finding the value of percentage correctly classified using test set for every model in order to show which classifier provided the best classification. In order to improve the performance of the classification model, variable selection method known as *t*-statistic method was applied. The significant variable in developing classification model was selected through this method. The result revealed that PLS-DA classifier of the standardised data with application of

t-statistic showed the best performance with highest percentage correctly classified among the classifiers.

Keywords: *palm oil, chemometrics, FT-MIR, t-statistic, Partial Least Squares Discriminant Analysis (PLS-DA)*

INTRODUCTION

Public interests in food quality and methods of food production have increased significantly in the recent decade due to changes in eating habits, consumer behaviour, and the increased industrialisation and globalisation of food supply chain [1]. There are different types of vegetable oils commonly used for cooking predominantly in deep-frying food. However, the oil experiences degradation due to its chemical properties and the exposure to heat during the frying process [2].

Deep frying is one of the popular methods for food preparation, especially in terms of preparing fast food. It is common practice to use cooking oil for food preparation at home and in commercial industries. A large portion of Malaysian use palm oil as Malaysia is one of the main producers of palm oil in the world. Cooking oil such as palm oil is mainly used as a heat exchange medium in cooking but when used for deep frying, they contribute to the organoleptic quality of the fried product [3].

The used palm oil is considered as a good renewable resource to be used as raw material for the production of biodiesel. However, it has become a threat to humans due to some unscrupulous traders who add it to new edible vegetable oil to increase the quantity which would bring in high profits. The government does not have the corresponding legal regulations to curb this matter. Adulteration of edible oil has been a chronic illness in food industry for many years. It not only causes a potential harm or threat to the health of consumers, but also undermines the integrity and stability of the economy. There are many ways of adulterating the oil, for example, highly priced oil adulterated with lower priced oil, edible oil adulterated with non-edible oil and qualified vegetable oils adulterated with waste cooking oil [4].

According to the National Poison Centre, consuming used cooking oil repeatedly can cause hypertension, affects the liver and may lead to cancer if used for long periods. Besides Jayabalan [5], said that cooking oil should not be used more than twice. When used repeatedly, the concentration of hydrocarbon in the oil increases and this can clog and stiffen arteries, causing hypertension and also affect the liver [6]. A research group highlighted that hypertension is related to the degradation of dietary frying oils. During the process of frying, the oil is heated to high temperatures and at the same time is exposed to the air, which results in a complex series of reactions which can increase the risk of hypertension [7].

Traditional analyses for food authentication, based on chemical and physical methods have several drawbacks, the most significant of which are low speed, the necessity for pre-treatments, a requirement for highly skilled people and destruction of the sample. Fourier Transform Infrared (FT-IR) spectroscopy has proven to be a successful analytical method for the analyses of a variety of food products [8].

Thus, it is important to develop a classification model based on a simple, fast and non-destructive instrumentation which is FT-MIR spectral data. It can be applied by industries in order to find out if there are any problems due to the adulteration of cooking oils. Combination of the chemometrics method and variables selection through t-statistic becomes interesting because it is an easy and friendly algorithm and it involves a minimal number of tuneable parameters and would be useful for analysis of a large and complicated FT-MIR datasets [9]. In this study, the focus will be classifying "pure" and "frying" cooking oil by using different model include PLS-DA, SVM and LVQ. Pure palm oil is referring to an unused cooking oil, while the term "frying" cooking oil is referred to as used cooking oil which is the palm oil that was used to fry the French potato at 180 degrees C until the color changed to yellow-brownish. The result showed that PLS-DA is a superior classification model compared to other models. This method yielded the highest percentage of correctly classified results when compared to other models.

Therefore, the combination of FT-MIR spectroscopy and chemometrics techniques provide a powerful tool to monitor a large variety of processes. The interest in enhancing the quality control procedures is increasing,

because it is a fast technique, there is no (or little) need of sample pre-treatment. The combination of FT-MIR spectroscopy and chemometrics techniques also offers good results to discriminate and determine physicochemical properties of cooking oils. On the other hand, it is necessary for the specialist on chemometrics in the research group or in the industry to manipulate the calculations or to interpret the results [1].

EXPERIMENTAL

Sample and Datasets

The datasets consist of 56 FT-IR spectra of cooking oil consisting of 28 types of pure palm oil and 28 different types of frying palm oil. All these cooking oil (samples) were purchased from local supermarkets around Negeri Sembilan. The samples were labelled as Class 1 in blue colour (pure) and Class 2 in red colour (frying) for further analysis. The frying oil was prepared by frying French fries at 180 degrees C until the colour changes to yellow-brownish. The oil used in the frying process was used as samples. The summary of the datasets used is shown in Table 1. The Matlab R2008a software was used to build the model.

Table 1: Summary of FT-IR datasets used for classifications pure and frying cooking palm oil

Cooking Oil	Number of Samples	Dimension
Pure Palm Oil	28 (Class 1, Blue)	28 x 3351
Frying Palm Oil	28 (Class 2, Red)	28 x 3351

FT-MIR Analysis

The FT-MIR spectra of samples was recorded in absorbance mode at the spectral range from 650cm^{-1} to 4000cm^{-1} with 4 cm^{-1} spectral resolution and 4 scan number using Perkin Elmer 100 FTIR Spectrometer. The Attenuated Total Reflectance (ATR) was used at sample station for sample handling. The background spectrum was first collected. No sample was placed on the ATR plate. The spectra were subtracted against background air spectrum. After every scan, a new reference air background spectrum was taken. The ATR plate was carefully cleaned in site by applying methanol

and dried with soft tissue before filling in the next sample, and made it possible to dry the ATR plate [10]. Figure 1 shows FT-IR spectra of pure and frying palm oil.

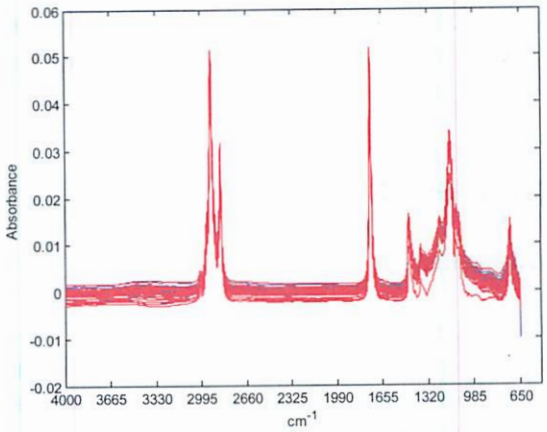

Figure 1: FT-IR spectra of pure palm oil (blue) and frying palm oil (red)

Chemometrics Methods

Previous studies indicated that several research group used chemometrics technique in predicting palm oil parameter which is the iodine value [11]. The use of chemometrics to classify pure palm oil and frying palm oil has not been widely used in other similar researches. Therefore, this study focuses more on classifying palm oil based on classes in which Class 1 (blue) is for pure palm oil and Class 2 (red) is for frying palm oil. Prior to beginning the designing of the classification model, it is necessary to prepare the data before performing the PCA [12]. In this paper, the raw data was pre-processed using standardisation method. In order to find the best model of classification, the data was split into two sets which were the training set and test set. The total number of samples were randomly divided into two sets, a training set (around two-thirds of the samples) and a test set (around one-third of the samples). Duplex splitting method was used to split the original data. Then, the training set was pre-processed and the information was incorporated to pre-process the test set. PCA scores and loadings were extracted from pre-processed training set and the loadings were used to extract the scores of the test set. The pre-processed original data, training and test set were further analysed using three different classifiers

which are LVQ, SVM and PLS-DA. Percentage correctly classified (% CC) was used as performance indicator to evaluate each classifier. The higher the percentage enabled more samples to be assigned to their correct groups. Besides that, the variable selection method in which t-statistic also was used to classify the significant variable of pure palm and frying-palm cooking oil of FT-MIR datasets. It helps to improve the performance of classification model by giving a high number of % CC.

RESULT AND DISCUSSION

Data Pre-processing

In this study some pre-treatment were used in the various steps of the analysis and is presented in a summarised figure below. The column standardisation for pre-processing method was used on the original data. Column standardisation puts all variables on an equal scale. This is useful when the variables measured are absolute quantities of various compounds. Before that, the original data was smoothed and filtered using second derivatives Savitzky-Golay by using 7 gap sizes with 3 polynomial degrees [13].

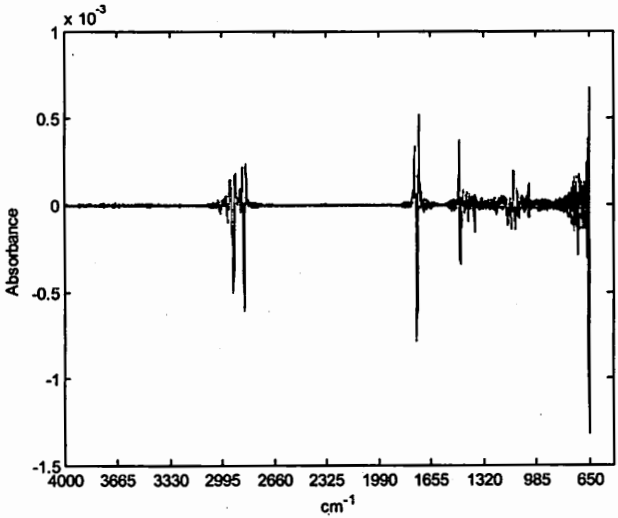

Figure 2: Second derivatives Savitzky – Golay plot of FT-MIR spectral data

Principle Components Analysis

Principle Components Analysis was plotted for the smoothed and filtered data in order to visualize the pattern of Class 1 and Class 2 cooking oils. Class 1 (blue) is for pure palm oil while Class 2 (red) is for frying palm oil. Figure 3 shows PCA plot of pure and frying palm oil. It shows that Class 1 and Class 2 are less overlapping among class.

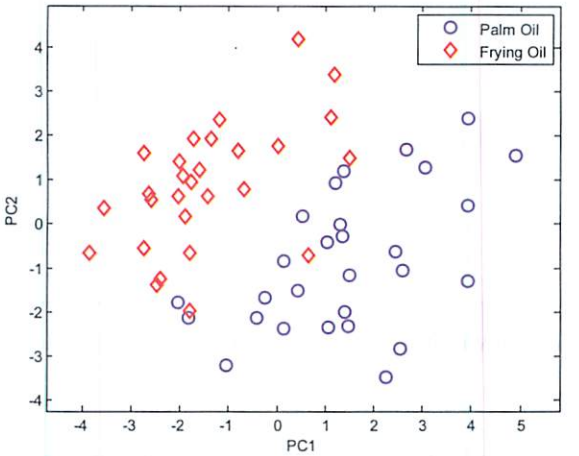

Figure 3: PCA plot of standardised data of pure (blue) and frying palm oil (red)

Variable Selection

Variable selection enhances the understanding and interpretability of multivariate classification models [14]. Before beginning to design a classification method, when many variables are involved, only those variables that are really required should be selected; that is, the first step is to eliminate the less significant variables from the analysis [15]. Variable selection method used in this work was t-statistic algorithm [9]. The t-statistic plays an important role in this study since it was used to identify and obtain the most important variable from the original variable of newly generated data as much as possible [16]. It is because before variable selection was made on raw datasets, the PCA plot was not really good and had a lot of overlapping of Class 1 and Class 2. Thus, after applying

t-statistic, PCA plot showed such a good result. After significant variable had been selected, it is necessary to plot again PCA plot for new generated data. Comparison of PCA plot before and after variable selection is showed in Figure 4. The selected variables chosen and ranked using *t*-statistic are shown in Table 1.

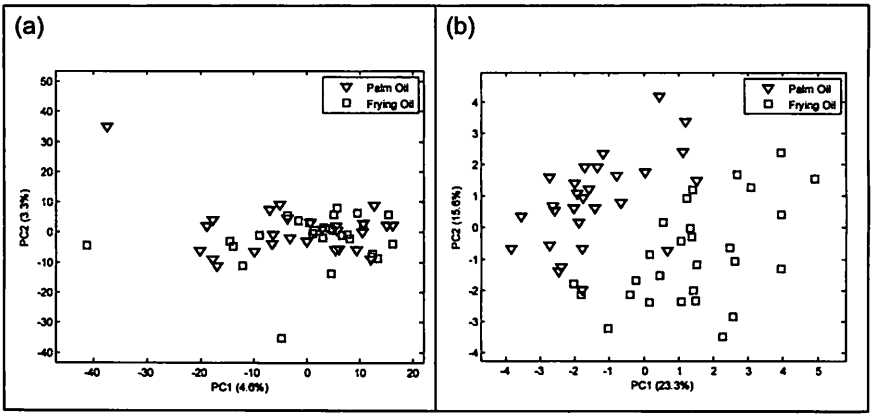

Figure 4: PCA plot without (a) and with (b) variable selection using column standardised data

Table 1: Most 10 significant variables ranked and chosen by *t*-statistic from the analysis

Rank	Wavenumber (cm^{-1})	*t*-statistic values
1	2266	4.7886
2	2265	4.7840
3	3246	-3.6169
4	1331	-3.5980
5	459	-3.4519
6	3221	-3.3309
7	2162	3.1895
8	3250	3.1825
9	3243	-3.1770
10	3324	-3.1462

Classification Model

The scores from PCA that were obtained using column standardised data were used in the classification models which are LVQ, SVM and PLS-DA. PLS-DA is an extension of the PLS method. In this case, the oil was classified into two classes which is "pure" and "fry" palm oil. The value in the y vector, for example "pure" was replaced by +1 or -1("fry").The predicted class of test set samples was identified using the values close to -1 or +1 and 0 was used as the threshold. The investigation of the PLS model's performance can also be extended if this threshold is varied. LVQ generates the codebooks vectors for each class in the training set to classify the oils. For SVM, it primarily creates a complex decision boundary between two classes with good classification ability [17]. Figure 5 shows the class boundary plot for each model. From visualisation, PLS-DA shows the best classification between Class 1 and Class 2 with less misclassified samples compared to other models [18].

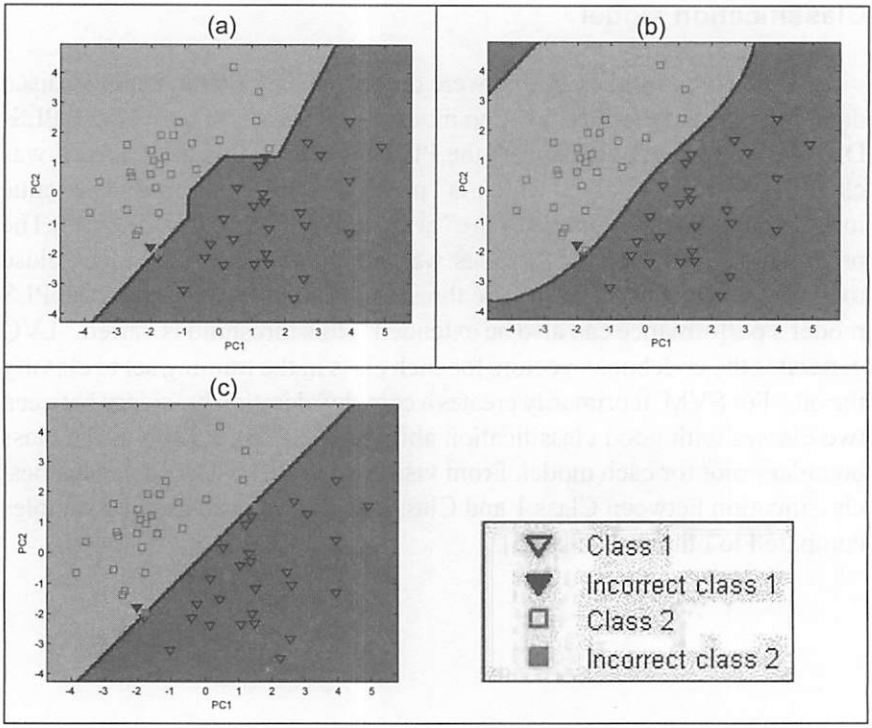

Figure 5: LVQ (a), SVM (b) and PLS-DA (c) boundary plot produced using two principle components

Model Performance Indication

Table 2 shows the overall effectiveness of the classification models which was evaluated by percentage correctly classified for the test set before variable selection and after variable selection was carried out. Figure 5 shows that all three classifiers show clear classification between Class 1 "pure palm oil" and Class 2 "frying palm oil". However, to confirm and validate the performance of classifier it is important to monitor their %CC of test for every classifier. Variable selection method or *t*-statistic method also plays an important role in this study.

Table 2 shows PLS-DA is a superior classifier because it gave higher % CC among other classifier. The performance of PLS-DA, SVM and LVQ keeps increasing after applying *t*-statistic method. The higher numbers of % CC, the more samples were assigned to their correct group.

Table 2: Difference of %CC of test set without and with variable selection method

Percentage of Correctly Classified (% CC)		
Classifier	**Test set without *t*-statistic**	**Test set with *t*-statistic**
LVQ	50	88.88
SVM	50	88.88
PLS-DA	38.88	94.44

CONCLUSION

In this study, the discrimination of pure palm oil and frying palm oil was carried out using FT-MIR spectra. The best data pre-processing method is column standardisation data. While for classification, PLS-DA shows superior classifier. The application *t*-statistic would increase the performance of classification model since it only chooses the significant variable which is relevant to develop the model.

ACKNOWLEDGMENT

The author would like to thank members of Faculty Applied Science, UiTM Negeri Sembilan for their help in this study. Also special thanks to the research group members for their guidance, to Universiti Teknologi MARA for granting an Excellence Fund (600-RMU/ST/DANA 5/3/Dst (8/2012)) and to the Ministry of Education Malaysia (MOE) for supporting this research through Exploratory Research Grant Scheme (600-RMI/ERGS 5/3 (3/2012)).

REFERENCES

[1] Luna, A.S., A.P. da Silva, J.S.A. Pinho, J. Ferré, and R. Boqué, Rapid characterization of transgenic and non-transgenic soybean oils by chemometric methods using NIR spectroscopy. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. 100,(2013). 115-119.

[2] Kuligowski, J., D. Carrión, G. Quintás, S. Garrigues, and M. de la Guardia, Direct determination of polymerised triacylglycerides in deep-frying vegetable oil by near infrared spectroscopy using Partial Least Squares regression. Food Chemistry. 131,(2012). 353-359.

[3] Bazlul Mobin Siddiquea, A.A., Mahamad Hakimi Ibrahima and Mohd Omar A. Ka, Thermal Effect on the Physico-chemical Properties of Blends of Palm Olein with other Vegetable Oils. Journal of Industrial Research & Technology. 1,(2011). 127-134.

[4] Zhang, Q., C. Liu, Z. Sun, X. Hu, Q. Shen, and J. Wu, Authentication of edible vegetable oils adulterated with used frying oil by Fourier Transform Infrared Spectroscopy. Food Chemistry. 132,(2012). 1607-1613.

[5] Harian, B., Minyak masak terpakai punca darah tinggi,(30 December 2003).

[6] *Minyak masak terpakai punca darah tinggi.*, in Berita Harian 2003.

[7] Soriguer, F., G. Rojo-Martínez, M.C. Dobarganes, J.M.G. Almeida, I. Esteva, M. Beltrán, . . . E. García-Fuentes, Hypertension is related to the degradation of dietary frying oils. The American journal of clinical nutrition. 78,(2003). 1092-1097.

[8] Oliveri, P., V. Di Egidio, T. Woodcock, and G. Downey, Application of class-modelling techniques to near infrared data for food authentication purposes. Food Chemistry. 125,(2011). 1450-1456.

[9] Sim, S.F. and W. Ting, An automated approach for analysis of Fourier Transform Infrared (FTIR) spectra of edible oils. Talanta. 88,(2012). 537-543.

[10] Rohman, A. and Y.B.C. Man, Fourier transform infrared (FTIR) spectroscopy for analysis of extra virgin olive oil adulterated with palm oil. Food Research International. 43,(2010). 886-892.

[11] Che Man, Y.B. and G. Setiowaty, Multivariate calibration of Fourier transform infrared spectra in determining iodine value of palm oil products. Food Chemistry. 67,(1999). 193-198.

[12] Brereton, R.G., Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. TrAC Trends in Analytical Chemistry. 25,(2006). 1103-1111.

[13] Luna, A.S., A.P. da Silva, J. Ferré, and R. Boqué, 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. 100,(2013). 109-114.

[14] Alsberg, B.K., D.B. Kell, and R. Goodacre, Variable Selection in Discriminant Partial Least-Squares Analysis. Analytical Chemistry. 70,(1998). 4126-4133.

[15] Ng, C.L., R.L. Wehling, and S.L. Cuppett, Near-Infrared Spectroscopic Determination of Degradation in Vegetable Oils Used To Fry Various Foods. Journal of Agricultural and Food Chemistry. 59,(2011). 12286-12290.

[16] Wongravee, K., G.R. Lloyd, J. Hall, M.E. Holmboe, M.L. Schaefer, R.R. Reed, . . . R.G. Brereton, Monte-Carlo methods for determining optimal number of significant variables. Application to mouse urinary profiles. Metabolomics. 5,(2009). 387-406.

[17] Xu, Y., S. Zomer, and R.G. Brereton, Support vector machines: A recent method for classification in chemometrics. Critical Reviews in Analytical Chemistry. 36,(2006). 177-188.

[18] Quintás, G., N. Portillo, J.C. García-Cañaveras, J.V. Castell, A. Ferrer, and A. Lahoz, Chemometric approaches to improve PLSDA model outcome for predicting human non-alcoholic fatty liver disease using UPLC-MS as a metabolic profiling tool. Metabolomics. 8,(2012). 86-98.