Journal of Oil Palm Research
DOI: https://doi.org/10.21894/jopr.2023.0038

N-ALKANE PROFILES OF LARD AND VEGETABLE OILS, AND THEIR CHEMOMETRICS DIFFERENTIATION

NUR AIN SYAQIRAH SAPIAN[1,2]; MUHAMAD AIDILFITRI MOHAMAD ROSLAN[1,2,3]; AMALIA MOHD HASHIM[1,4]; YANTY NOORZIANNA ABDUL MANAF[5]; MOHD NASIR MOHD DESA[1]; MURNI HALIM[2]; MUHAMAD SHIRWAN ABDULLAH SANI[6]; MOHD TERMIZI YUSOF[4]; MOHD SABRI PAK DEK[7] and HELMI WASOH[1,2]*

ABSTRACT

This research aims to examine fat from various vegetable oils using n-alkane profiles, as well as chemometrics and machine learning. Unsaponifiable vegetable oils (coconut, peanut, palm and soybean oils) were separated and analysed using gas chromatography-mass spectrometry (GC-MS) to investigate the n-alkane profiles of each fat. The authenticity of the detected n-alkane profiles was determined by comparing to the retention time of C_7-C_{40} n-alkane standards. The test designs were developed using Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), Partial Least Squares-Discriminant Analysis (PLS-DA), and Random Forest (RF). Both PCA and HCA appeared to provide a clear distinction between each of the vegetable oil tests. Based on the PLS-DA and RF determination, tetracosane (C_{24}) and octadecane (C_{18}) are proposed as the key n-alkane markers for separating lard from vegetable oils. These findings suggest that additional work may be required to achieve and determine the different characteristics across oils in numerous statistical applications, notably chemometrics and machine learning.

Keywords: chemometrics, lard, n-alkane, principal component analysis, random forest.

Received: 1 January 2023**; Accepted:** 5 July 2023**; Published online:** 22 August 2023.

[1] Halal Products Research Institute, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.

[2] Department of Bioprocess Technology, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.

[3] Institute of Systems Biology (INBIOSIS), Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia.

[4] Department of Microbiology, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.

[5] Halal Research Group, Faculty of Food Science and Nutrition, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia.

[6] International Institute for Halal Research and Training, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia.

[7] Department of Food Science, Faculty of Food Science and Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.

* *Corresponding author e-mail: helmi_wmi@upm.edu.my*

INTRODUCTION

Lard is one of the cheapest semisolid fats widely used in food production to give food a rich flavour and creamy texture. Due to this reason, it has been subjected as a potential source for adulteration to other fats and oils. Since the use of lard and its by-products are prohibitive under the Kosher and Halal regulations (Al-Kahtani *et al.*, 2017), it is important to explore techniques for clear discrimination between lard and other permissible ingredients (Azir *et al.*, 2017). Each fat and oil exhibit specific components and their presence should be considered as an important detection tool. Mainly, fats and oils are composed of triglycerides (TAGs), while the remaining consists of minor components such as hydrocarbons, fatty acids (FAs), natural compounds, unsaponifiable and polar lipids (Ferreira *et al.*, 2021).

Hydrocarbons can be found in the unsaponifiable fraction and they are the least polar compounds, made up of only carbon (C) and hydrogen (H) (Giuffrè and Capocasale, 2016).

Single-chain hydrocarbons are known as normal alkanes (n-alkanes), with the formula CnH_2n_{+2} (Zenkevich, 2006). To date, several methods are employed to detect lard adulterants such as electronic nose, polymerase chain reaction (PCR), Fourier transform infrared spectroscopy (FTIR) (Salleh *et al.*, 2018), differential scanning calorimetry and other different chromatographic-based techniques (such as HPLC, GLC, LC-MS, GC-FID and GC-TOF-MS) (Azir *et al.*, 2017). The HPLC and GC-based detection techniques could, however, become more difficult if the TAG or FAs composition of the adulterant exhibit close similarity to that of the tested oils. Several analytical approaches have been developed to analyse the n-alkane quantitatively (Troya *et al.*, 2015). Previous research reported that the n-alkane composition could potentially be used to discriminate against a group of vegetable oils (Mihailova *et al.*, 2015; Troya *et al.*, 2015). Quantitative data considering the whole n-alkanes profile might serve better as a characteristic 'fingerprint' of the oil. If such profiles are to be used to study the effect caused by the adulteration, powerful statistical methods to deal with the multivariate data are recommended (Yousefinejad *et al.*, 2018) to consider the differences taking place in the whole chromatogram. It has now been realised that the qualitative analysis using mainly single-component species may be inadequate (Sharin *et al.*, 2021).

The objective of PCA and PLS-DA is to achieve a linear transformation that converts data to a very low dimensional condition with an error as minimum as possible. For PCA, in its first principal component (PC), the transformation preserves as much variance as possible, however, PLS-DA preserves as much covariance as possible between the original and labelling data. Both can be described as iterative processes as the rest of the variance and error are preserved to define the next PC (Ruiz-Perez *et al.*, 2020). Besides its use for dimensionality reduction, it can be adapted to be used for feature selection (Christin *et al.*, 2013) as well as for classification (Botella *et al.*, 2009). Yousefinejad *et al.* (2018) reported that the m-dimensional space of a matrix D (from the PCA result) can be reduced to a significantly lower-dimensional space of principal components (PCs): D = TP' + E. Where, T is the score matrix (containing PCs in the direction of the samples), P is the loading matrix (carrying in the direction of the wavenumbers) and E is the residual matrix. The superscript (') denotes the matrix transpose. The calculated PCs represent the compressed information, demonstrating the fine difference in the instrumental signal relating to various samples for discrimination purposes. PLS-DA can be assumed as a "supervised" version of PCA in the sense that it achieves dimensionality reduction but with full awareness of the class labels (Ruiz-Perez *et al.*, 2020). Previously, PCA and PLS-DA were successfully used for the blood-volatile organic compound in an animal study (Ataabadi *et al.*, 2023) and the solvent-electrochemical interactions of anthraquinones (Aq) in different organic media (Yousefinejad *et al.*, 2015).

The PLS-DA showed a good performance in the calibration step (Yousefinejad *et al.*, 2018) as a further consideration for classification. The other multivariate techniques such as hierarchical cluster analysis (HCA) can be further applied due to their high sensitivity. Variable Importance in Projection (VIP) scores estimate the importance of each variable in the projection used in the PLS model. In any proposed model, a variable with VIP Score close to one can be considered important. Variables with a score of less than one can be considered as less important and potential for exclusion (Chong and Jun, 2005). In PLSDA, Cocchi *et al.* (2018) introduced a filter method which is independent of the modelling step and does not require additional validation nor increase the adjustable model parameters (by referring to the VIP variables ranking). If R^2 and Q^2 are acceptable and the model is interpretable, thus, the suggested model can be improved by deleting unimportant terms, such as the variables with low VIP values. Thereafter a final model is developed, interpreted and thus, better predictions can be concluded (Wold *et al.*, 2001). In this case, less complex statistical applications can be applied due to the advantage to exclude unimportant items (Cocchi *et al.*, 2018). Therefore, this criterion is reasonable with good potential to discard irrelevant variables. To date there is no report has been published on the application of GC-MS in combination with multivariate analysis. Therefore, sophisticated methods need to be given due consideration and the present study aimed at developing an advanced and effective methodology for the study of lard adulteration by using the GC-MS. As such, the objective was coupled with multivariate data analysis, as a means for distinguishing lard from other vegetable oils.

MATERIALS AND METHODS

Reagents

Chemicals for column chromatography *i.e.*, potassium hydroxide, methanol, ethanol, anhydrous sodium sulphate, silica gel 60 and n-hexane were purchased from Merck (Germany). Saturated alkanes standard of C_7-C_{40} (1000 µg/mL, in hexane) for every component were purchased from Supelco (Bellefonte, PA, USA). All the reagents were of analytical grade.

Sample Preparation

Lard and vegetable oils including coconut, peanut and soybean oils were obtained from the Laboratory of Halal Science Research (Halal Products Research Institute, Universiti Putra Malaysia), while palm oil was obtained from the Malaysian Palm Oil Board (MPOB). Lard was made by rendering pig adipose tissue at 100ºC for 2 hr. To remove contaminants, the oils were filtered using a muslin cloth and then stored at -20ºC until use.

Extraction of Unsaponifiable Fraction

The n-alkane fractions of vegetable oils were created with slight adjustments according to Troya *et al.* (2015). The standard saponification process was used to extract the unsaponifiable fraction of each fat. The preferred fat samples (5 g) were saponified using a reflux condenser and heating mantle for 20 min at 70ºC with an ethanolic potassium hydroxide solution (10%, 19 mL). The solution was allowed to cool before being added to a separating funnel with 25 mL of distilled water. Using 25 mL of n-hexane, the non-saponifiable materials were extracted twice. The hexane extracts were mixed and added to a new separating funnel, which were then washed three times with an ethanol-water mixture in a similar ratio (12.5 mL). The extracts were dried over an anhydrous sodium sulphate and evaporated at 30ºC under vacuum using a rotary evaporator (EYELA, Japan). For separation, the residues were dissolved in 2 mL n-hexane.

Separation procedure using solvents and column chromatography was used to separate the hydrocarbon fraction and impurities from the unsaponifiable materials obtained. The stationary phase was created using a silica gel-filled glass column with an internal diameter of 1.5 cm and a length of 40 cm (15 g in n-hexane). A chromatographic elution technique was used with 40 mL of n-hexane as the mobile phase, and the eluate (20 mL) was regarded as an n-alkane fraction and subjected to further investigation. The eluate was evaporated under vacuum in a rotary evaporator (EYELA, Japan) at 30ºC to remove the solvent. The resulting sample was immediately diluted in 0.5 mL of n-hexane and stored in a chiller at 4ºC prior further analysis with gas chromatography-mass spectrometry (GC-MS).

Gas Chromatography-Mass Spectrometry Analysis

The n-alkane composition of lard and other animal fats was analysed using the Agilent 7890A gas chromatography paired with the Agilent 5975C mass spectrometry detector in a selected ion monitoring (SIM) mode (Agilent, USA). The HP-5MS (30 m × 0.25 μm × i.d., 0.25 μm film thickness) was used as the analytical capillary column. The oven temperature was set to 35ºC for 2 min, then increased to 250ºC (10º/min), and eventually to 300ºC for 23 min (20º/min). The detector temperature was 300ºC. The MS interface was kept at a constant temperature of 300ºC. The injection volume was 1 μL and operated in the split mode (at a split ratio of 1:10). Helium was used as the carrier gas at a flow rate of 1 mL/min. The compounds were identified by comparing the data with NIST 11 mass spectral library and C_7-C_{40} saturated alkanes standard (supplied by Supelco, Bellefonte, PA, USA). The composition of each identified n-alkane was determined by calculating the peak area of each alkane to the total peak area of all alkanes in the sample (Hassan *et al.*, 2010).

Chemometric Analysis and Machine Learning

Data analysis was performed using statistical online software packages; MetaboAnalyst 5.0 (Parasitology Building, 21111 Lakeshore Road Ste. McGill University, Anne de Bellevue, QC, Canada; http://www.metaboanalyst.ca) software. A total of 12 samples were analysed in triplicates. Prior to chemometric analysis, data were pre-tested using Kruskal-Wallis *post-hoc* test at $p<0.05$ to determine the significant differences in the means of peak area of each fat sample (Chong *et al.*, 2019). The data matrix was then normalised to a constant sum and scaled by Pareto scaling to modify the variance of spectral data so that the peaks are equally weighted in order to build multivariate models (Höjer Holmgren *et al.*, 2018). Clustering of oils' n-alkanes was accomplished using unsupervised methods, PCA and Hierarchical Clustering Analysis (HCA). Supervised approaches were used to classify oil samples, including the partial least squares-discriminant analysis (PLS-DA) and the Random Forest (RF) of the machine learning algorithm. The performance of the PLS-DA model was evaluated using a 10-fold cross-validation (Van Ruth *et al.*, 2010).

During the initial stage, it is unclear which value of the diagnostic statistics was really corresponding to discrimination between groups (Westerhuis *et al.*, 2008). Permutation tests assumed that there is no difference between two random groups (Westerhuis *et al.*, 2008). Furthermore, the labels of all samples were randomly permuted followed by the calculation of a new classification model (Lindgren *et al.*, 1996). By repeating the procedure N times, a null distribution of H_0 was obtained (Preece, 1990). The significance of the PLS-DA model was then assessed by relating the values of the calculated model to the H_0 distribution of the permuted data sets. From these classifications, H_0 distributions for Q^2 and regression coefficients and others were obtained. As the groups were formed in a random way, therefore, no difference between group was used as an assumption.

PLS-DA classifies lard from other animal fats by highlighting the important variables in term of VIP scores, and according to Cocchi *et al.* (2018) the score can be potentially applied especially for variable selection**.** The RF classification was performed using Metaboanalyst 5.0 software, and the model's performance was confirmed by using the RF package in the RStudio. In this study, the normalised data were randomly divided between training (90%) and test (10%) groups. A confusion matrix was generated to determine the RF model's prediction accuracy.

RESULTS AND DISCUSSION

Profiles of n-Alkane

The n-alkane fractions of lard and all tested vegetable oils were subjected to GC-MS analysis to determine their individual composition. A total of 17 n-alkanes ranging from C_8 to C_{27} were identified (*Figure 1*).

Some vegetable oils showed a similar pattern of n-alkane composition, where alkanes with chain lengths of C_{12}, C_{14}, C_{15} and C_{16} were the most abundant, except for peanut oil (C_{10}, C_{12}, C_{14} and C_{15}). For coconut, palm and soybean oils, the C_{14} was found with the highest abundance, composed of 21.46%, 19.78% and 21.06%, respectively. Meanwhile, for peanut oil, C_{12} was the most significant, with 23.49% of the total composition. Compared to lard, the most predominant alkanes are C_{24}, followed by C_{16}, C_{15}, C_{14} and C_{18}, composed of 15.72%, 14.84%, 14.05%, 13.60% and 9.94%, respectively. Among the long chain alkanes (C_{14}-C_{18}), the C_{18} (9.94%) contributed to the highest increase (0.44-fold) when compared to the nearest competitor (palm oil, 5.58%). The C_{16} and C_{15} only contributed to 0.18 and 0.15-fold, respectively (compared to the nearest competitor). The n-alkane C_{24} (15.72%) was notable for lard since this alkane was absent (or present at low percentage) in coconut, palm, peanut and soybean oils.

Clustering Analysis of Lard and Vegetable Oils

Only 14 significant n-alkanes (p<0.05) were chosen from the 17 discovered (*Figure 2a*). The clustering of lard from vegetable oils was done using unsupervised PCA and HCA methods. PCA is the most often used unsupervised technique in the chemometrics research, and it analyses the pattern identification by using the correlation between the examined data (Ruiz-Perez *et al.*, 2020). PCA score plot of lard and vegetable oil samples (coconut, palm, peanut and soybean oils) involving the first (PC1) and second (PC2) components, with 83.0% variations were explained. PC1 explained 68.5% of the variance, while PC2 explained 14.5% of the variance.

Figure 2b shows the PCA biplot of the samples as a combination of the score plot and loading plot. Variables with weights (loading) close to 1 or -1 have a strong influence on sample clustering (Azizan *et al.*, 2021), while those with almost no weight have a minor impact (Schaettler *et al.*, 2022). Based on the PCA biplot, C_{24} had a positive contribution towards lard clustering with the loading of -0.27277 and -0.48868 toward PC1 and PC2, respectively. The alkanes that contributed to the clustering of all vegetable oils at a positive score of PC2 are C_{10}, C_{12}, C_{14} and C_{15}. The influence of the variables in the biplot was supported in the heatmap of HCA (*Figure 2c*).

Consistent with the PCA result, HCA also clustered the samples into two groups, *i.e.*, lard and vegetable oils (coconut, palm, peanut and soybean oils). This clustering result was achieved through the combination of Euclidean distance measure and Ward's linkage algorithm. Based on the heatmap,

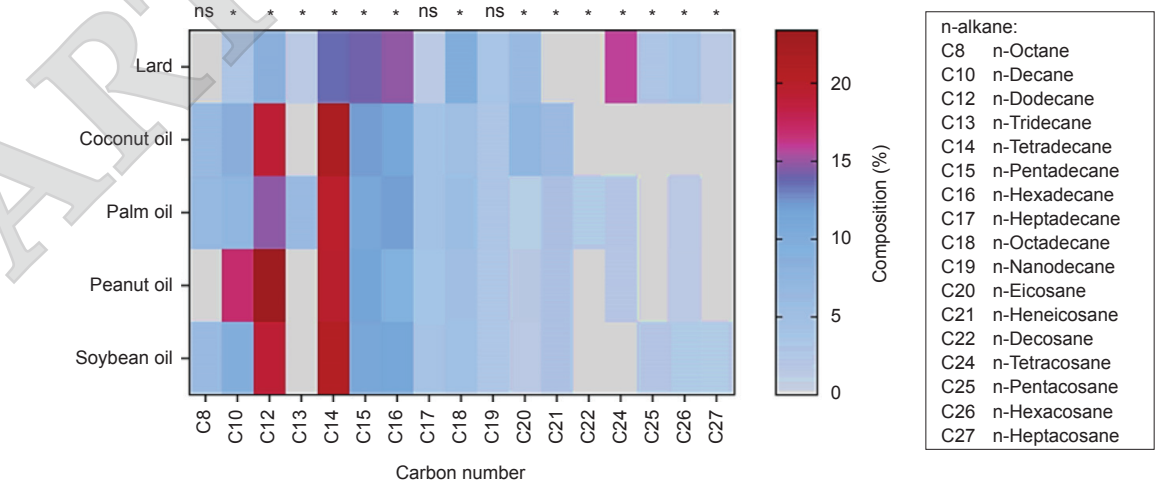
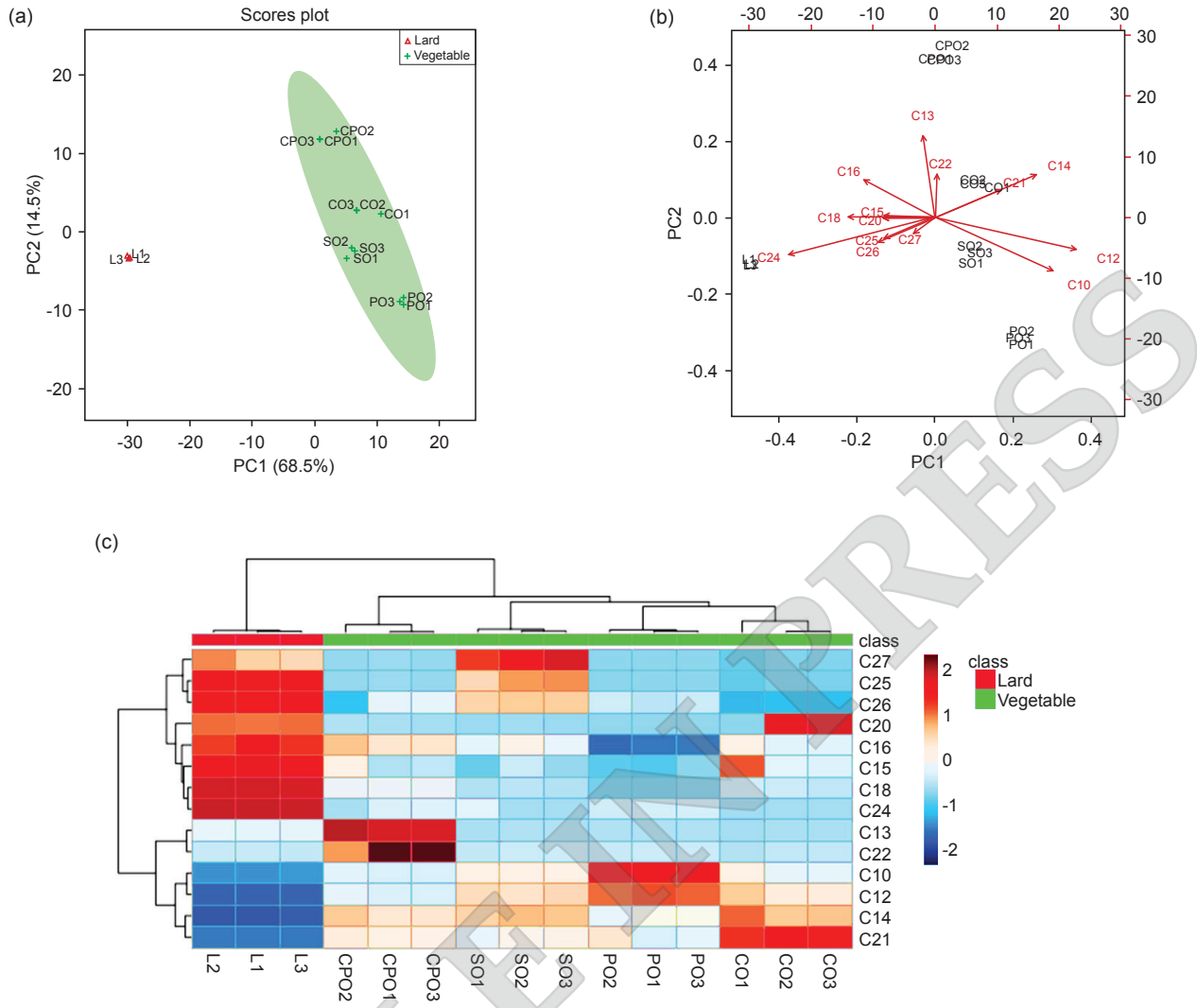


*Figure 1. Heatmap of n-alkane composition (%) of lard and vegetable oils supported by Kruskal-Wallis Test; * - significant difference (p<0.05); ns - non-significant difference (p>0.05).*



Note: CO - Coconut oil; CPO - Palm oil; PO - Peanut oil; SO - Soybean oil.

Figure 2. (a) PCA score plot, (b) PCA biplot, and (c) dendrogram heatmap of lard and other vegetable oils based on 14 significant n-alkanes profile.

all vegetable oil samples were grouped together in a cluster but separated distinctly into 4 sub-clusters based on the type of oils. The heatmap presents individual values contained in a matrix in the form of colours (Zhao *et al.*, 2014). The correlation of C_{18} and C_{24} was detected in high levels in lard but absent in the vegetable oil samples. The distribution of n-alkanes in vegetable oils was more closely related to each other. The heatmap shows that C_{10}, C_{12}, C_{14} and C_{21} contribute to the clustering of all vegetable oils distinctly from the lard. Meanwhile, C_{13} and C_{22} were notable and unique to palm oil as both alkanes were found in low amounts or absent in lard and other vegetable oils.

Classification Analysis of Lard and Vegetable Oils

The same dataset of 14 significant alkanes was used for the classification analysis using supervised methods of PLS-DA model and RF machine learning. Lard and vegetable oil samples were classified using the PLS-DA model with variance explanation of 68.3% (component 1) and 14.3% (component 2) (*Figure 3a*).

Because this model is prone to overfitting, cross validation (CV) was used to assess its performance. In this work, 10-fold cross-validation was used to establish the appropriate number of components needed to construct the PLS-DA model (*Figure 3b*). Based on three popular performance criteria, such as prediction accuracy, R^2 and Q^2 values, the three-component model was deemed to as the best classifier. The PLS-DA model's predictability can be rated excellent if $Q^2 \geq 0.90$ (Idris *et al.*, 2022). Using the 14 significant alkanes, a 10-fold cross-validation found an accuracy of 1.0, variance repeated in cross-validation (Q^2 value) of 0.99, and endpoint variation incorporated in the regression model (R^2 value) of 1.0. Based on the result, there were no outlier and overfitting occurred for the PLS-DA model as indicated by the value difference of R^2 and Q^2 (<0.01).

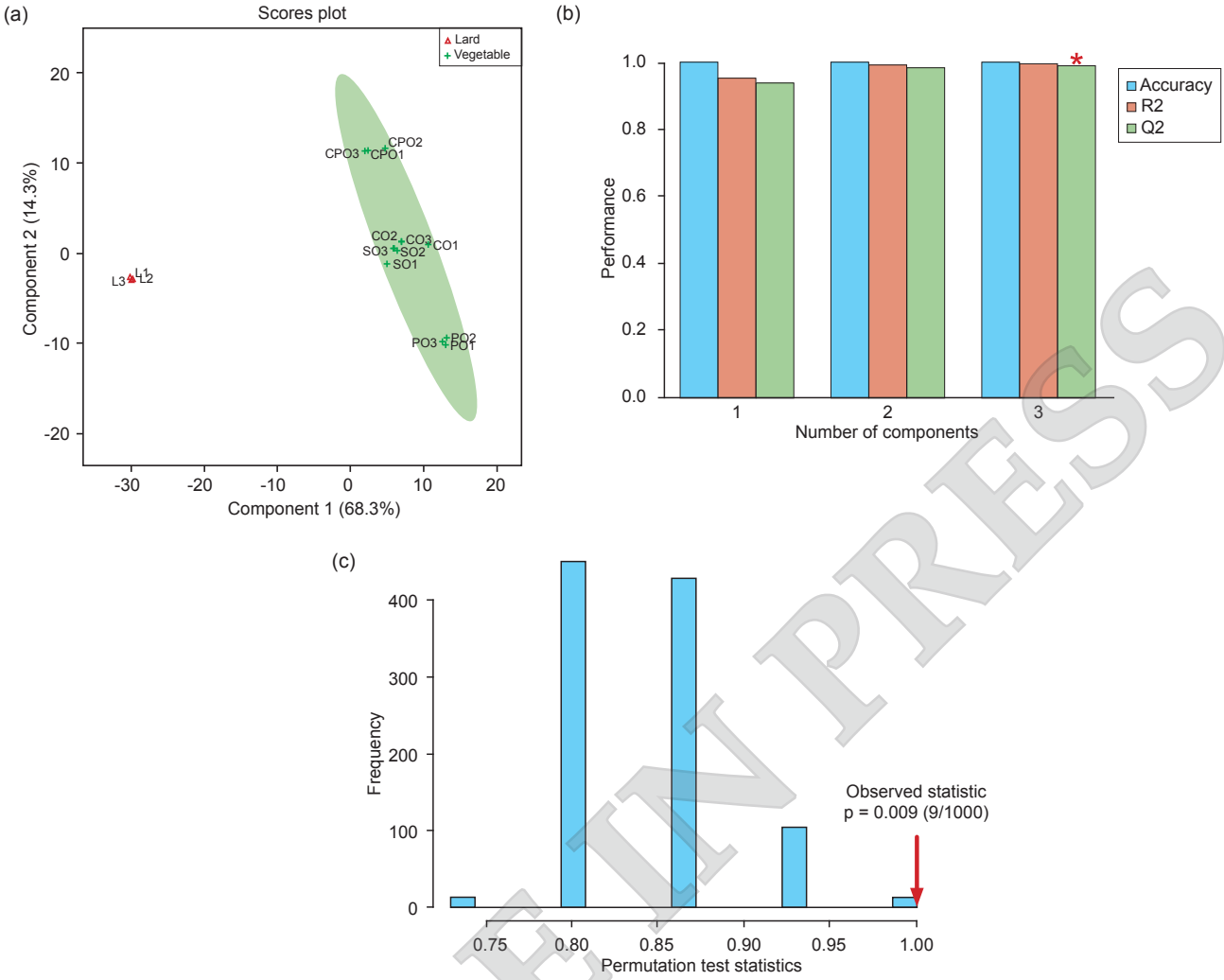


*Figure 3. 2-D PLS-DA model (a) of lard and other vegetable oil samples based on n-alkane profile with 82.6% of explained variance, 10-fold cross-validation (b) determined the best classifier for the PLS-DA model, marked with an asterisk (*), and permutation test of PLS-DA model (c) with 1000 random permutations (*p<0.009).*

The performance of the developed PLS-DA model was further evaluated. A permutation test using 1000 random permutations was conducted to determine the overfitting possibility. Despite the common use of PLS-DA for assessing the discriminatory and predictive ability of oils and alkanes, it seems that the developed models are too closely fitted to the current data. Such argument can be referred to the value of C_{24} (0.049) which is close to 0.050 by using Kruskal-Wallis test (*Figure 1*). In this study, the permutation test indicated a possibility of such accuracy (*Figure 3c*) due to potential overfitting since the *p*-value obtained was 0.009. Therefore, the permutation test can be considered as a very sensitive method for classification. Golland *et al.* (2005) reported that the number of permutations needs to be "large enough" to sample both tails of the available distribution dataset. A permutation test can be used to evaluate whether the specific classification of the individual group (within a pair of groups) is better than any other random classification group (Mielke and Berry, 2007). Hemmateenejad *et al.* (2011) also reported that the cross validation and permutation test are very promising statistical tools to evaluate the prediction ability of the proposed model.

Another classification approach using RF machine learning was performed using Metaboanalyst 5.0 and then validated by the R package in the R Studio. The same 14 pre-processed datasets were utilised in this classification analysis. The samples were grouped correctly into their respective classes, therefore producing an "out-of-bag" (OOB) error of 0 for all classes (*Table 1*). The RF model revealed the significant variables in terms of mean decrease accuracy values, where the importance of the variables was indicated in descending order (*Figure 4*). Consequently, random forest highlighted C_{18}, C_{24} and C_{16} as the top three significant alkanes in the classification of lard from the selected vegetable oils, with the values of 0.033743, 0.031867 and 0.030652, respectively.

According to Hengl *et al.* (2018), an RF model does not require cross validation since this

algorithm already makes an unbiased estimation internally through the OOB error estimation. In this study, an additional 10-fold cross-validation was implemented to calculate the prediction accuracy of the developed RF model. Classification accuracy was measured based on how many times the predictions accuracy made by the model are correct. This additional cross-validation step indicated a prediction accuracy of 1.0 that is acceptable criteria for model predictability.

TABLE 1. CONFUSION MATRIX OF SAMPLE CLASSIFICATION FOR LARD AND VEGETABLE OILS BY RANDOM FOREST

Item	Lard	Vegetable oil	Class error
Lard	3.00	0.00	0.00
Vegetable oil	0.00	12.00	0.00

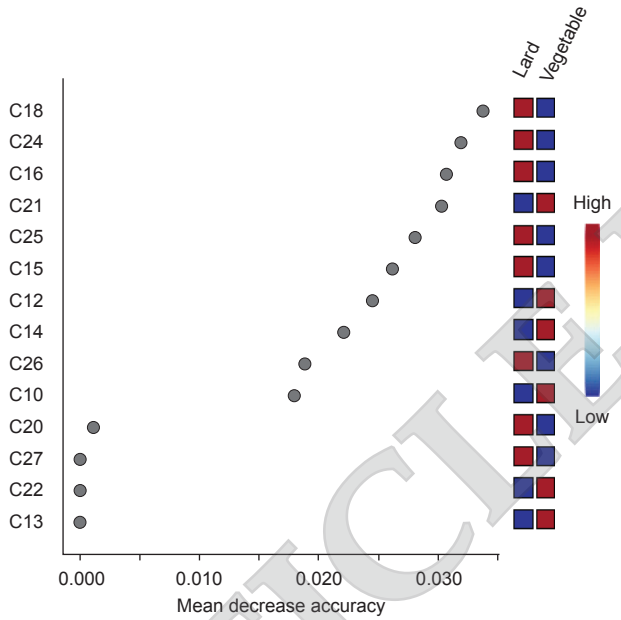


Figure 4. Random forest highlighted significant variables in sample classification in decreasing order of mean decrease accuracy.

Identification of Potential n-alkane Markers

Potential n-alkane marker of lard was identified after analysing the outcome and performance of the developed models to each other. *Table 2* shows the summary of loading scores of the PCA model, the mean decrease accuracy value determined by the RF, and the total abundance of selected significant n-alkanes in the differentiation of lard from vegetable oils. RF highlighted C_{18}, C_{24} and C_{16} as the top three significant alkanes in the differentiation of lard from the selected vegetable oils with mean decrease accuracy of 0.0337, 0.0319 and 0.0307, respectively. In the PCA model, C_{18} and C_{16} were located on the negative side of PC1 and the positive side of PC2, while C_{24} located on the negative sides of both PCs to indicate the distinctive clustering of lard samples (*Figure 2a*).

Even though, at first, the model for classification result was considered invalid since the PLS-DA developed in this study was overfit, however, further classification by the RF provides an acceptable prediction accuracy of 1.0. Consequently, tetracosane (C_{24}) and octadecane (C_{18}) were proposed as the potential lard markers based on the RF feature selection and their contribution to lard clustering in the PCA model. The C_{24} alkane was detected to be the most abundant in lard (15.72%) and the C_{18} was considered to be given the highest increase if compared to the close highest alkane value (0.44-fold). PLS-DA and RF are popular machine learning tools with useful selector and classifier features (Ruiz-Perez *et al.*, 2020). Since the PLS-DA was able to strengthen the PCA finding, the multitude tools available and the use of different methods depending on the dataset is an advantage. It is important to note that the PLS-DA role in the discriminant analysis can be easily misinterpreted (Brereton *et al.*, 2014). As the PLS-DA is prone to overfitting, it is important to further verify its function by using CV (Kjeldahl and Bro, 2010). To refine the result, the RF was trained and validated using a separate unrelated group data (Broughton-Neiswanger *et al.*, 2020; Sharin *et al.*, 2021). The developed methodology can be helpful in early routine laboratories practice (Felipe Bachion *et al.*, 2018) supposedly for the screening analysis of different oils.

TABLE 2. SUMMARY OF PCA LOADING SCORES, MEAN DECREASE ACCURACY VALUE AND TOTAL ABUNDANCE OF SELECTED SIGNIFICANT N-ALKANES FOR LARD CLASSIFICATION FROM OTHER FATS

n-Alkanes	PCA loading scores (PC1, PC2)	Mean decrease accuracy[1]	Decrease abundance of alkane in lard[2] (%)
C_{18}	-0.26319, 0.34170	0.0337	9.94
C_{24}	-0.27277, -0.48868	0.0319	15.72
C_{16}	-0.06548, 0.21531	0.0307	14.84

Note: [1] Mean decrease accuracy determined by random forest.
[2] Abundance of n-alkane in percentage calculated by the following formula: (Peak area of single alkane / Total peak area of n-alkanes) × 100.

CONCLUSION

Through GC-MS and multivariate analysis, lard and vegetable oil samples (coconut, palm, peanut, and soybean oils) were differentiated based on n-alkane profiles using unsupervised method (PCA and HCA) and verified by supervised method (PLS-DA and RF). Only 14 significant n-alkanes (p<0.05) were chosen from the 17 original n-alkanes

for detail classification. The PCA score plot (PC1 and PC2) was indicated by 83% variation with C_{24} had a major positive contribution towards lard clustering. Similar to PCA, the HCA separated lard and vegetable oils into two different clusters. Due to overfitting, PLS-DA and RF were further used to verify the PCA result and 82.6% of variance was obtained for both components 1 and 2. With a prediction accuracy of 1.0, the generated RF model was considered as acceptable categorisation. A cross-validation (accuracy of 1.0), variance repeated in cross-validation (Q^2 of 0.99) and endpoint variation incorporated in the regression model (R^2 of 1.0) concluded no outlier and overfitting occurred for the PLS-DA model (R^2 and Q^2<0.01). For distinguishing lard from other vegetable oils, the tetracosane (C_{24}) and octadecane (C_{18}) were chosen as the potential indicators because of their high influence on the sample clustering in the PCA and PLS-DA models, high abundance in lard, and prominence as the first and second most significant variables in the RF model.

ACKNOWLEDGEMENT

The authors would like to acknowledge the Universiti Putra Malaysia for financial support under the Putra Grant Scheme [Grant number 9646800] and Ministry of Higher Education, Malaysia (KPT reference code: FRGS/1/2020/STG01/UPM/02/14) under the Fundamental Research Grant Scheme [Grant number 5540357].

REFERENCES

Al-Kahtani, H A; Ahmed M A; Abou-Arab, A A and Hayat, K (2017). Identification of lard in vegetable oil binary mixtures and commercial food products by FTIR. *Qual. Assur. Saf. Crops Foods*, *9(1)*: 11-22. DOI: 10.3920/QAS2015.0692.

Ataabadi, M S; Bahmanpour, S; Yousefinejad, S and Alaee, S (2023). Blood volatile organic compounds as potential biomarkers for poly cystic ovarian syndrome (PCOS): An animal study in the PCOS rat model. *J. Steroid Biochem. Molecul. Biol.*, *226*: 106215. DOI: 10.1016/j.jsbmb.2022.106215.

Azir, M; Abbasiliasi, S; Tengku Ibrahim, T; Manaf, Y; Sazili, A and Mustafa, S (2017). Detection of lard in cocoa butter - Its fatty acid composition, triacylglycerol profiles and thermal characteristics. *Foods*, *6(11)*: 98. DOI: 10.3390/foods6110098.

Azizan, N I; Mokhtar, N F K; Arshad, S; Sharin, S N; Mohamad, N; Mustafa, S and Hashim, A M (2021). Detection of lard adulteration in wheat biscuits using chemometrics-assisted GCMS and random forest. *Food Analyt. Met.*, *14(11)*: 2276-2287. DOI: 10.1007/s12161-021-02046-9.

Botella, C; Ferré, J and Boqué, R (2009). Classification from microarray data using probabilistic discriminant partial least squares with reject option. *Talanta*, *80(1)*: 321-328. DOI: 10.1016/j.talanta.2009.06.072.

Brereton, R G and Lloyd, G R (2014). Partial least squares discriminant analysis: Taking the magic away. *J. Chemom.*, *28(4)*: 213-225. DOI: 10.1002/cem.2609.

Broughton-Neiswanger, L E; Rivera-Velez, S M; Suarez, M A; Slovak, J E; Hwang, J K and Villarino, N F (2020). Pharmacometabolomics with a combination of PLS-DA and random forest algorithm analyses reveal meloxicam alters feline plasma metabolite profiles. *J. Vet. Pharmacol. Ther.*, *43(6)*: 591-601. DOI: 10.1111/jvp.12884.

Chong, I-G and Jun, C-H (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometr. Intell. Lab. Syst.*, *78(1-2)*: 103-112. DOI: 10.1016/J.CHEMOLAB.2004.12.011.

Chong, J; Yamamoto, M and Xia, J (2019). MetaboAnalystR 2.0: From raw spectra to biological insights. *Metabolites*, *9(3)*: 57. DOI: 10.3390/metabo9030057.

Christin, C; Hoefsloot, H C; Smilde, A K; Hoekman, B; Suits, F; Bischoff, R and Horvatovich, P A (2013). Critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Mol. Cell Proteomics*, *12(1)*: 263-276. DOI: 10.1074/mcp.M112.022566.

Cocchi, M; Biancolillo, A and Marini, F (2018). Chemometric methods for classification and feature selection. *Compr. Anal. Chem.*, *82*: 265-295. DOI: 10.1016/bs.coac.2018.08.006.

Felipe Bachion, S; Sarmento, M; Lucas, G; Waldomiro, N and Ronei, P (2018). Rapid discrimination between authentic and adulterated andiroba oil using FTIR-HATR spectroscopy and random forest. *Food Analyt. Met.*, *11*: 1927-1935. DOI:10.1007/s12161-017-1142-5.

Ferreira, G F; Ríos Pinto, L F; Carvalho, P O; Coelho, M B; Eberlin, M N; Maciel Filho, R and Fregolente, L V (2021). Biomass and lipid characterization of microalgae genera *Botryococcus, Chlorella*, and *Desmodesmus* aiming

high-value fatty acid production. *Biomass Convers. Biorefin.*, *11(5)*: 1675-1689. DOI:10.1007/s13399-019-00566-3.

Giuffrè, A M and Capocasale, M (2016). N-Alkanes in tomato (*Solanum lycopersicum* L.) seed oil: The cultivar effect. *Int. Food Res. J.*, *23(3)*: 979-985.

Golland, P; Liang, F; Mukherjee, S and Panchenko, D (2005). Permutation tests for classification. *COLT*, *5*: 501-515.

Hassan, J; Farahani, A; Shamsipur, M and Damerchili, F (2010). Rapid and simple low density miniaturized homogeneous liquid-liquid extraction and gas chromatography/mass spectrometric determination of pesticide residues in sediment. *J. Hazard. Mater.*, *184(1-3)*: 869-871. DOI: 10.1016/j.jhazmat.2010.08.008.

Hengl, T; Nussbaum, M; Wright, M N; Heuvelink, G B and Gräler, B (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*: p.e5518. DOI: 10.7717/peerj.5518.

Hemmateenejad, B; Yousefinejad, S and Mehdipour, A R (2011). Novel amino acids indices based on quantum topological molecular similarity and their application to QSAR study of peptides. *Amino Acids*, *40*: 1169-1183. DOI: 10.1007/s00726-010-0741-x.

Höjer Holmgren, K; Hok, S; Magnusson, R; Larsson, A; Åstot, C; Koester, C; Mew, D; Vu, A K; Alcaraz, A; Williams, A M; Norlin, R and Wiktelius, D (2018). Synthesis route attribution of sulfur mustard by multivariate data analysis of chemical signatures. *Talanta*, *186*: 615-621. DOI: 10.1016/j.talanta.2018.02.100.

Idris, M H H; Abdullah Sani, M S; Mohd Hashim, A; Mohd Zaki, N N; Abdul Manaf, Y N; Mohd Desa, M N; Arshad, S; Yuswan, M H; Hassan, M S; Yusof, Y A; Kamaruddin, M S and Mustafa, S (2022). Forensic feed strategy: Incorporation of multivariate and instrumental analyses for authentication of fish feed sources. *J. Halal Industry and Services*, *5(1)*: a0000293. DOI: 10.36877/jhis.a0000293

Kjeldahl, K and Bro, R (2010). Some common misunderstandings in chemometrics. *J. Chemom.*, *24(7-8)*: 558-564. DOI:10.1002/cem.1346.

Lindgren, F; Hansen, B; Karcher, W; Sjöström, M and Eriksson, L (1996). Model validation by permutation tests: Applications to variable selection. *J. Chemom.*, *10(5-6)*: 521-532.

Mielke, P W and Berry, K J (2007). *Permutation Methods: A Distance Function Approach*. New York: Springer.

Mihailova, A; Abbado, D and Pedentchouk, N (2015). Differences in n-alkane profiles between olives and olive leaves as potential indicators for the assessment of olive leaf presence in virgin olive oils. *Eur. J. Lipid Sci. Technol.*, *117(9)*: 1480-1485. DOI: 10.1002/ejlt.201400406.

Preece, D (1990). RA Fisher and experimental design: A review. *Biometrics*: 925-935. DOI: 10.2307/2532438.

Ruiz-Perez, D; Guan, H; Madhivanan, P and Narasimhan, G (2020). So you think you can PLS-DA?. *BMC Bioinform.*, *21(2)*: 1-10. DOI: 10.1186/s12859-019-3310-7.

Salleh, N A M; Hassan, M S; Jumal, J; Harun, F W and Jaafar, M Z (2018). Differentiation of edible fats from selected sources after heating treatments using fourier transform infrared spectroscopy (FTIR) and multivariate analysis. *AIP Conf. Proc.*, *1972*(*1*): 030015. DOI: 10.1063/1.5041236.

Schaettler, M O; Richters, M M; Wang, A Z; Skidmore, Z L; Fisk, B; Miller, K E; Vickery, T L; Kim, A H; Chicoine, M R; Osbun, J W; Leuthardt, E C; Dowling, J L; Zipfel, G J; Dacey, R G; Lu, H-C; Johanns, T M; Griffith, O L; Mardis, E R; Griffith, M and Dunn, G P (2022). Characterization of the genomic and immunologic diversity of malignant brain tumors through multisector analysis. *Cancer Discov.*, *12(1)*: 154-171. DOI: 10.1158/2159-8290.CD-21-0291.

Sharin, S N; Sani, M S A; Jaafar, M A; Yuswan, M H; Kassim, N K; Manaf, Y N; Wasoh, H; Zaki, N N M and Hashim, A M (2021). Discrimination of Malaysian stingless bee honey from different entomological origins based on physicochemical properties and volatile compound profiles using chemometrics and machine learning. *Food Chem.*, *346*: 128654. DOI: 10.1016/j.foodchem.2020.128654.

Troya, F; Lerma-García, M J; Herrero-Martínez, J M and Simó-Alfonso, E F (2015). Classification of vegetable oils according to their botanical origin using n-alkane profiles established by GC-MS. *Food Chem.*, *167*: 36-39. DOI: 10.1016/j.foodchem.2014.06.116.

Van Ruth, S M; Villegas, B; Akkermans, W; Rozijn, M; van der Kamp, H and Koot, A (2010). Prediction of the identity of fats and oils by their fatty acid, triacylglycerol and volatile compositions using

PLS-DA. *Food Chem.*, *118(4)*: 948-955. DOI: 10.1016/j.foodchem.2008.10.047.

Westerhuis, J A; Hoefsloot, H C; Smit, S; Vis, D J; Smilde, A K; van Velzen, E J; van Duijnhoven, J P and van Dorsten, F A (2008). Assessment of PLSDA cross validation. *Metabolomics*, *4*: 81-89. DOI: 10.1007/s11306-007-0099-6.

Wold, S; Sjöström, M and Eriksson, L (2001). PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, *58*: 109-130.

Yousefinejad, S; Bahram, M and Baheri, T (2018). Classification of methamphetamine seized in different regions of Iran using GC-MS and chemometrics. *J. Iran Chem. Soc.*, *15*: 163-170. DOI: 10.1007/s13738-017-1219-5.

Yousefinejad, S; Honarasa, F; Fararouei, M and Moosavi-Movahedi, A A (2019). Structure-electrochemistry relationship for monovalent alkaline metals in non-aqueous solutions. *Phys. Chem. Liquids, 57(5)*: 600-620. DOI: 10.1080/00319104.2018.1507031.

Zenkevich, I G (2006). Use of recurrence relations for approximating properties of any homologs of organic compounds. *Russ. J. Gen. Chem.*, *76*: 1742-1752. DOI: 10.1134/S1070363206110120.

Zhao, S; Guo, Y; Sheng, Q and Shyr, Y (2014). Heatmap3: An improved heatmap package with more powerful and convenient features. *BMC Bioinform.*, *15(S10)*: P16. DOI: 10.1186/1471-2105-15-S10-P16.