CHAPTER 2

LITERATURE REVIEW

2.1 Chemometrics Technique

According to the International Chemometrics Society (ICS), chemometrics is a new chemical discipline that employs theory and methods from mathematics, statistics, computer science, and other related fields to optimise chemical measurement procedures. It aims to extract as much chemical information as possible from chemical data. For example, chemometrics techniques have been used to analyse edible fats and oils. Typically, researchers have used univariate analysis such as analysis of variance (ANOVA) or *t-tests* by selecting a few analytes of oils or fats for comparison. However, many modern analytical tools could generate a massive volume of data from oil and fat material that are difficult to process or interpret literally and require multivariate analysis for more meaningful information.

Chemometrics utilised in multivariate data analysis could involve calibration, process modelling, pattern identification, classification, signal correction, and statistical process control. Chemometrics is applied in the experimental biological sciences, particularly chemistry, to address descriptive and predictive problems. The chemical property is modelled in descriptive solutions to understand the system's underlying relationships and structure (Artrith & Urban, 2016; Ward & Wolverton, 2017). In other words, descriptive modelling aims to understand the model of a chemical system or system identification (Kumar & Sharma 2018). In predictive applications,

established models are projected outside samples identified for classification and authentication purposes (Bevilacqua et al., 2017).

2.1.1 Principal Component Analysis (PCA)

In terms of descriptive solutions, PCA is an explorative technique without any prior knowledge or being natural. Therefore, it is also an unsupervised clustering and classification technique (Biancolillo & Marini, 2018). It is similar to Factor Analysis in statistics (De Winter & Dodou, 2016). PCA reduces the dependence of variables in multivariate data by computing eigenvalues and eigenvectors, which can then be translated into visual forms. The graphic forms of the data analysis are presented through a scores plot and loadings plot. In brief, the scores plot describes the identification of clusters of the same or different from the others class. Loading plots are variables that map scores plot based on the similarities and differences of the samples.

Scores plot is a vector arranged straight by the most variance, known as the Principal Component (PC) and identified as the first PC (PC-1). Each PC is not the same and opposed to straight lines or orthogonal, as the second most variance is called the second PC (PC-2). Each PC determines the plots' position and variances from the data tables as per Equation 2.1 below;

$$X = TP' + E \tag{2.1}$$

which X is a combination m x n data matrix of spectral data or column. T is $m \times k$ matrix of score values for all the spectra, and P is k x n matrix of PCs. The E matrix contains the spectral residuals not fit by the optimal PCA and has the same dimension as X; m is the number of samples, n is the

number of data points, i.e., wavenumbers, chemical shift, and retention time. k is the number of PC used to reconstruct the X, which is $k < m$. The superscript (') denotes matrix transpose.

PCA is commonly used in the chemometric technique as a correlation analysis of different types of edible fats and oils as scores plot distribution and loadings plot displayed as FAs or TAGs mapping the pattern of the scores vector. The generalisability of the first 2 PCs as an indicator of the abundant FAs in GC has been widely published in oils and fats research (Caridi et al., 2021; Tian et al., 2019; Xing et al., 2019). For instance, the first 2 PCs always have been an indicator of the most variance contributing to the C16:0, C18:1, and C18:2 in edible fats and oil measured by GC-FID and GC-MS (Muthai et al., 2019).

The FAs of 540 Tunisian virgin olive oil hybrids PC-1 contribute 36.84%, mainly attributed to C18:1 using GC-FID (Dabbou et al., 2012). Thermally raw and processed lipids from Mangalitza pig demonstrated that the explained variance for PC-1 is 53.28 % related to the linoleic and vaccenic/elaidic FA (Petroman et al., 2021). The study of olive oil and sunflower oil classification blends at 40 %, 50 %, and 65 % showed linolenic, oleic, arachidic, and margaroleic acids located at the most variance PC-1 (Monfreda et al., 2012). Heidari et al., (2020) performed a combination GC-MS and PC-1 at 92 % variances to identify FAMEs and quantify lard adulteration in olive oil.

It is possible to find the main variable for each most variance of PC. The first two PCs of PCA with mathematical meaning represent the highest variances. Thus, the reduction original of data into the particular PC for the %

variance accounting of the total variance for better visualisation of samples. Moreover, the higher the PCs contribute the % variance, then the PCs could represent more information on the original variables.

2.1.1.1 Hotelling's T-Squared (T^2)

Various studies have combined PCA with other chemometric methods, such as multivariate calibration and discriminant for pattern recognition (Agussabti et al., 2020; Amante et al., 2019; Mohammadi et al., 2019). The future prediction validates this technique on new samples with selected analytical platforms for the developed method. In the oils and fats field, Lörchner et al., (2022) have claimed the novelty of sample rapeseed oil in quality control evaluation. Their work demonstrated that FTIR-PCA data was integrated with Hotelling's T-squared distribution and Q-residuals.

Hotelling's T^2 statistics measure how extreme the observation in the model space and distance to the model centre as spanned by the PCs. Thus, Hotelling's T^2 statistics can track systematic changes in the setting and detect situations where a system is outside normal conditions (Catelani et al., 2018). Hotelling's T^2 critical limit is shown as a confidence ellipse in the 2D scores plot or as a line plot with an upper critical limit. Individual samples can be used to test the similarity of the models by Hotelling's T^2 value at a certain confidence level (CI).

2.1.1.2 PCA Projection

The prediction of new samples has been extended based on classes or categories such as PLS, PCR, and SVM. In comparison, another supervised

data reduction method is PCA projection, or Supervised PCA, which is not almost considered for the classification method (Badcock et al., 2004; Bair et al., 2006; Chao et al., 2019).

More recently, PCA projection has been applied in other fields. Straková et al., (2020) demonstrated that they could successfully group root samples into more universal root types by re-constructing the PCA combined with FTIR spectra of plant roots in peatlands. Privé et al., (2020) demonstrated that PCA projections could efficiently give unbiased population structure from genetic data of the UK Biobank and the 1000 Genomes project. Although both studies claimed the novelty of PCA projection in their respective fields, this application was found to have been introduced by Wold et al., (1987) over thirty years ago.

PCA projection is equivalent to the prediction in regression methods and provides a robust statistic. Furthermore, projecting the test set samples onto the existing calibration model (training set) and checking residual variances and leverages that allows to simultaneously determine model validation for the test set samples. In brief, test set x is projected into the A-dimensional space generated in training set by multiplying with loadings P of a trained set as in Equation 2.2.

$$t = xP \tag{2.2}$$

The residuals vector e computed from the trained data as in Equation 2.3. The identity matrix is symbolised as I from the vectors. Since the vectors are independent, this estimation of the new scores t, or loadings P, is analogous to linear regression (Eriksson et al., 2006 & 2014).

$$e = x - tp' \quad (2.3)$$

or

$$e = x(I - PP')$$

The possibilities of PCA projection as a multivariate calibration have not been fully established in many fields compared to other successive projection algorithms such as PLS. In another predictive solution, multivariate discriminant and multivariate regression analysis were employed using multivariate calibration. Multivariate calibration studies combine data from several sources to solve selectivity issues, obtain new insights, and instantly detect outliers (Brereton et al., 2018). Escandar et al., (2006) described modern multivariate calibration methods, which appear appealing when applied to fundamentally unselective spectroscopic or electrochemical signals.

2.1.2 Multivariate Classification

Multivariate classification is a multivariate dimensionality reduction technique that uses an optimal new axis to optimise the distance between classes and minimise the variation within categories. The goal is to reduce the number of dimensions while maintaining as much information as possible. Oliveri et al., (2019) simplified multivariate classification as discriminant analysis; partial least squares discriminant analysis (PLS-DA), linear discriminant analysis (LDA), *k*-nearest neighbours (*k*-NN) and quadratic discriminant analysis (QDA), class-modelling methods unequal class models (UNEQ), and soft independent modelling of class analogy (SIMCA).

PLS-DA and LDA are the most successive projections algorithm models which are probability-based. De Santana et al., (2016) proved that PLS-DA could effectively distinguish between legitimate rosehip oil and adulterated rosehip oil containing 5 %, 10 %, 15 %, and 20 % (w/w) containing soybean, corn and sunflower oils. Vegetable oils were compared using a PCA, PLS-DA, and LDA mixture by various parameter distinct principal components. Despite their apparent similarities, FAs are considered the most powerful in discriminating between olive and camellia oils, while sterols and squalene also play a useful role (Shen et al., 2021).

Overfitting occurs when a model performs exceptionally well on training data but poorly on test data (new data). Since PLS-DA has overfitting issues, many researchers have already found other alternatives. For example, Random Forest (RF), a method based on decision trees, was applied to authentic and adulterated andiroba oil-containing soybean and corn oils in different proportions using FTIR and gave better results than PLS-DA (De Santana et al., 2018). The other option to overcome overfitting issues is selecting the optimal number of PLS components (Lei et al., 2019) and selecting the best pre-processing first derivatives and standard normalisation variable (SNV) transformation before applying PLS-DA on the FAs peanuts dataset (Yu et al., 2020).

2.1.2.1 Linear Discriminant Analysis (LDA)

LDA is generalised for the case of more than two classes. The generalisation forms for the within-class and between-class covariance

matrices. Thus, to find the weight vector, W by maximising the eigenvectors that correspond to their largest eigenvalues or PCs as per Equation 2.4,

$$W = eig(s_{w^{-1}} s_B) \tag{2.4}$$

where the eig is eigenvectors that correspond to their most significant eigenvalues within-class covariance matrix S_W, estimating the between-class covariance S_B. Once the projection has been found, all data points, class means, and covariances can be transformed into the new axis system.

LDA was applied in chromatographic data of sesame oil and perilla oil and successfully detected adulterated samples event at as low as 2 % by mixing contaminated levels of perilla oil with cheaper oils like maize oil and soybean oil (Kim et al., 2020). Moreover, when the algorithm is trained, no internal parameters need to be altered, which is an advantage of the LDA learning approach (Chen et al., 2015).

The other types of discriminant analysis depend on the variability of each group that does not have the same structure (unequal covariance matrix). If the shape of the curve separating groups is not linear thus, quadratic discriminant analyses can provide a better classification model, as in Figure 2.1 (Hastie et al., 2009).

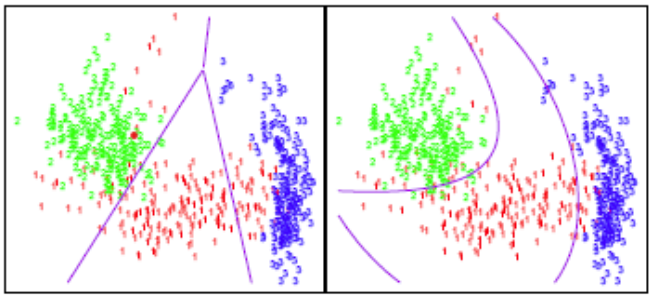

Figure 2.1: Illustration of Boundaries LDA (left) vs QDA (right).

The centre distances of the groups also can be measured using the Mahalanobis distance known as MDA as per Equation 2.5,

$$d_{kl} = \sqrt{(x_k - x_l).C^{-1}(x_k - x_l)'} \quad (2.5)$$

where x_k is the data vector of k in a matrix $(n \times 1)$, x_l is the mean data vector $(n \times 1)$, $C^{-1}$is the covariance matrix $(n \times n)$, $(x_k - x_l)'$ denotes the transposition of $(x_k - x_l)$, and n is the number of data points in x_k (Adams, 2004). QDA is another distance metric similar to the LDA metric. However, QDA calculates the distance to each class using the sample variance-covariance matrix rather than the overall pooled matrix.

2.1.2.2 Support Vector Machines (SVM)

SVM can be applied to regression by introducing an alternative loss function, and the results were favourable. The basic idea of SVM is to map the data X into a higher-dimensional feature space (F) via a nonlinear mapping and then to do linear regression in this space, as illustrated in Figure 2.2 (Wang et al., 2007).

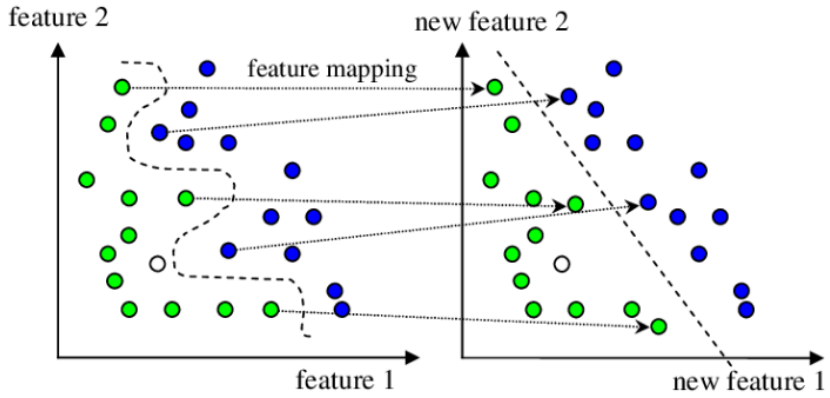

Figure 2.2: Basic Idea of the Kernel Function in SVM.

Thus, regression approximation addresses the problem of estimating a function established to a given data set $G = \{(x_i: d_i)\}_{i=1}^{l}$, where $(x_i: d_i)$ is the input vector, d_i is the chosen value. SVM approximates the function in Equation 2.6 as follows;

$$y = \sum_{i=1}^{l} w_i \, \Phi_i(w) + b \tag{2.6}$$

From Equation 2.6 the $\{\Phi_i(x)\}$ is the set of mappings of input features, and $\{(w)\}_{i=1}^{l}$, b are the coefficients. Hence, a margin of tolerance between the two types of kernel parameters, epsilon (ε) and *Nu* (υ), were compared to get the best calibration and prediction results. The difference between ε and υ is the ε parameter which is less sensitive than υ. The *C* hyperplanes parameter is regularised term in SVM that controls the penalty of the model for each misclassified point for a given curve. The linear discriminant function in The Unscrambler® X has the default option type is as ε-SVM and *C*-SVM.

2.1.2.3 Evaluation of Multivariate Classification

A comprehensive comparison of the best case for each classification model among the fat classes is according to class sensitivity and specificity. Class sensitivity describes the model's ability to correctly recognise samples belonging to one category namely sensitivity (*Sens*) as per Equation 2.7,

$$Sens = \frac{TP}{(TP + FN)} \tag{2.7}$$

Class specificity describes the model's ability to reject samples of all the other classes from the other one. In this regard, if all the representatives of

class groups are correctly identified (true assigned), then a sensitivity (*Spec*) was calculated as per Equation 2.8,

$$Spec = \frac{TN}{(TN + FP)} \tag{2.8}$$

where *TP* is the number of true positives, *FP* is false positives, *TN* is true negatives, and *FN* is false negatives. These parameters extract the actual class *vs* the predicted class from the confusion matrix.

The Matthews Correlation Coefficient (MCC) is a more reliable statistical rate that yields a high score only if the prediction performed well in all four confusion matrix categories (*TP*, *FP*, *TN*, and *FP*). The proportion of the size of positive and negative elements in the dataset produced a coefficient as a balanced metric that can be employed even if the classes are substantially different because it considers true and erroneous positives and negatives (Chicco & Jurman, 2020).

The MCC is a correlation coefficient that yields a number between +1 and -1 for observed and anticipated binary classifications (Equation 2.9.)

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FN)(TN + FN)}} \tag{2.9}$$

A coefficient of +1 denotes a perfect forecast, a coefficient of 0 indicates a random prediction, while a coefficient of -1 denotes a strong disagreement between prediction and observation.

2.1.3 Multivariate Regression

Multivariate regression is a strategy for determining the degree to which several independent variables (predictors) and multiple dependent

variables (responses) are connected linearly. The relationship between the acquired spectra and the quality features of interest is described in a multivariate framework. The model can be written as per Equation 2.10.

$$Y = f(X) \tag{2.10}$$

From a statistical standpoint, the model to determine the quality aspects of interest from the acquired spectra is defined as the search for a mathematical relationship (f) between the predictor or independent factors X and the predicted or dependent variables Y.

PCR and PLSR are the most used in multivariate regression algorithms to compare oil and fats analysis (Rohman et al., 2020). A combination of FTIR analysis with the PCR and PLSR was used for authentication of extra virgin olive oil and extra virgin coconut oil (Rohman et al., 2017); adulteration analysis of virgin coconut oil with grape seed oil and soybean oil (Rohman et al., 2019); mustard oil (Jamwal et al., 2020a) and virgin coconut oil from its adulterant paraffin oil (Jamwal et al., 2020b).

2.1.3.1 Principal Component Regression (PCR)

PCR is applied on the data matrix *X*, followed by multiple linear regression (MLR) steps. MLR was used between the scores obtained in the PCA step and the characteristic *Y* to be modelled in Equation 2.11. The first new variables or PCs are assumed to have the most variance of the original data. Meanwhile, the different numbers of PC sequences are less than the first PC. Consequently, only r are retained PCs and r < min (n, p) to data simplification. After performing PCA on *X*, the second step in PCR consists

of the linear regression of the scores and the y property of interest. The linear model between $(n\ x\ p)$ and $A\ (n\ x\ r\ )$ is as follows,

$$Y = Ab + e \tag{2.11}$$

where A is the weighted normalised matrix of order $n\ x\ p$ and will represent the new coordinates for the n objects in the new system, and p is the loading matrix, and the column vectors are called eigenvectors or PCs loading. Where b (R x k) are the coefficients and e is the error vector $(n\ x\ p)$. The objective of the response y is not directly correlated with X, but with its PC. The PC is obtained by decomposing X via PCA as per Equation 2.12. The prediction responses on X with $(\hat{b}_p)$ are given by:

$$\hat{y} = X\hat{b}_p \tag{2.12}$$

The coefficients of that combination are called regression coefficients or b-coefficients (b_p). If the X-variables weighted in this plot present the weighted regression coefficients, then the b_p are confounded by using the number of factors (or PCs) according to the models' interest scores.

2.1.3.2 Partial Least Squares (PLS)

The PLS algorithm finds principal components from spectral data relevant to analytic concentration (Brereton, 2009). PLS is a method for constructing regression models on the latent or hidden variable decomposition to relate two blocks, matrices X as per Equation 2.13,

$$X = \sum t_f \dot{p}_f + E \tag{2.13}$$

and Y, as per Equation 2.14, which contain the independent, x, and dependent, y, variables, respectively.

$$Y = \sum u_f \grave{q}_f + F \tag{2.14}$$

The objective of PLS is to find the best number of latent variables, typically performed using cross-validation (CV) based on the determination of minimum prediction error. In which T and U are the score matrices for the data set of X in a combination of Equation 2.13 and Y in Equation 2.14 respectively; p and q are the loading matrices for X and Y, respectively, E and F are the residual matrices. The two matrices are associated by the scores T and U, for each latent variable, as Equation 2.15,

$$u_f = b_f + t_f \tag{2.15}$$

from u_f, matrix Y can be calculated in Equation 2.15, and the constant of the new samples can be estimated from the new scores T, which are substituted from Equation 2.13, combined in Equation 2.14, leading to Equation 2.16.

$$Y_{new} = TBQT \tag{2.16}$$

Although the regression coefficients with PCs in PCR outperform MLR, there is no guarantee that PCs representing a significant proportion of variance in X would help predict Y. As a result, defining the new variables based on the relationship between X and Y could be more efficient.

The PLSR has been offered towards this goal, e.g., in the prediction of hyperspectral images of bulk samples of Canadian wheat by Mahesh et al., (2015) and the determination of informative wavelength band in NIR region for non-invasive blood glucose by Suryakala and Prince, (2019). The new parameters are presented as orthogonal (uncorrelated) linear combinations of the original variables that retain the covariance between X and Y to the maximum extent.

2.1.3.3 Orthogonal Signal Correction Partial Least Squares (OSC-PLS)

This research includes PLS models extended on orthogonal signal correction partial least squares (OSC-PLS). The most used in the literature are orthogonal partial least squares regression (OPLSR) or orthogonal partial least squares (OPLS). OPLS is a new version of PLS introduced by Trygg and Wold (2002), which possesses built-in OSC that filters out some variances in the *X*-matrix unrelated to *Y*.

The OPLS model separates systematic variation in the *X*-block into two parts, predictive and orthogonal. The predictive fit is the covariance between *X* and *Y*. The second part, orthogonal, contains systematic variation in *X* unrelated to *Y*. When *Y* is constructed using a dummy variable (0/1), PLS and OPLS are known as PLS-DA and OPLS-DA, respectively.

2.1.3.4 Multivariate Regression Coefficient as Important Features Assessment

The important variables can be observed by plotting the *X*-loadings weight for all the components *vs* the variable number. The variable has a significant positive or negative loading weight, indicating that the important variable is concerned with the component. A sample with a significant score value for this component will have a large positive value for a variable with a large positive loading weight. Regression coefficients summarise the relationship between all predictors and given responses. Using the following Equation 2.17,

$$B = W(P^T W)^{-1} Q^T \tag{2.17}$$

the regression coefficients b_k in a matrix as B calculated by the T scores.

For PLS, the regression coefficients can be computed for any number of components or factors. For example, the *B* for three factors summarises the relationship between the predictors and the response, as a model with three components approximates it. The weighted *B* provides information about the importance of the *X*-variables. The *X*-variables with a large *B* play an essential role in the regression model; a positive coefficient shows a positive link with the response, and a negative coefficient shows a negative relation. Conversely, predictors with a small coefficient are often negligible (CAMO, 2015).

2.1.3.5 Evaluation of Multivariate Regression Analysis

Model accuracy is evaluated by the coefficient of determination in calibration (*R* squared or R^2), coefficient of determination in prediction (*adjusted R* squared or *adj*. R^2), root mean square of error calibration (RMSEC) and prediction or validation (RMSEV), and means square of error in prediction (MSEP). The calculation of R^2 or *adj*. R^2 as shown in Equation 2.18,

$$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.18}$$

the y_i is the calibration value, $\hat{y}$ is predicted by the multivariate techniques model and $\bar{y}$ is computes the average of the calibration values in the training data set, and calculation of the RMSEC and RMSEV as in Equation 2.19,

$$\sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n_c}} \tag{2.19}$$

where n_c is the number of calibration or validation sets, $\hat{y}_i$ and y_i are the predicted and calibration values of the i^{th}observation samples. The model has prediction performance capability measured by the MSEP as shown in Equation 2.20,

$$\frac{1}{n_p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.20}$$

where n_p is the prediction set denoted by y_i and $\hat{y}_i$; are the measured and predicted values of the i^{th} observation.

The smallest RMSEC value was associated with the optimal calibration and RMSEV value for validation model. R^2 shows the percentage of the variance in the Y variables representing the variable X. The values of R^2 nearest to the one (1) is considered an excellent model performance in calibration. The RMSEC, RMSEV, and MSEP values are indicators of the reliability and predictive ability of the model and should be lower than the zero value.

Multivariate classification and multivariate regression are two of the most often used and beneficial approaches in applied statistics and chemometrics in the current era to demonstrate multivariate calibration (Baqueta et al., 2020; Kumar et al., 2014; Saeys et al., 2019). These approaches are involved in various fields and applications, from chemical spectroscopy to chromatography (Anzardi et al., 2021; Belal et al., 2020). The primary distinction between these two methods is that regression analysis uses a continuous dependent variable, whereas discriminant analysis requires a discrete dependent variable. Discriminant analysis is similar to a regression

analysis such as PLS-DA and PLSR. However, the output of the finding is different. For the discriminant analysis, the output is represented by a confusion matrix, whereas R^2 and RMSEC represent the output of multivariate regression.

Borghi et al., (2020) developed PLS model of the binary blends to identify and quantify adulterations in extra virgin olive oils (EVOO) with the construction of soybean, sunflower, corn, and canola oils; the outcomes were R^2 higher than 0.98 with RMSEV ≤ 5.0 wt %. The OPLS-DA and PLSR regression techniques using FTIR-UV–vis were successfully used to predict adulteration oils from the pure oils levels by Uncu and Ozen, (2019). The calibration of FTIR & UV; (R^2 = 0.94 & 0.98) and prediction (*adj.* R^2 = 0.91 & 0.97) models and low error values for calibration (4.22 % and 2.68 %). The best predictions were achieved using standard spectra determination of squalene content of EVOO in combination between fluorescence and infrared with the lowest RMSEC of 0.1065, RMSEV of 0.131, and MSEP of 0.150 in the spectral region 250–730 nm (Tarhan, 2020).

From a statistical or data analysis perspective, overcoming collinearity between spectral variables is the primary difficulty in such problems. Collinearity happens when a high correlation between two variables exists, making estimation of each variable by regression coefficient difficult or impossible. Furthermore, real-world applications frequently involve databases with a few known spectra and many spectral variables (Næs & Mevik, 2001). Therefore, data pre-processing, feature selection, and projection are required for any technique based on the original spectra

variables because these variables do not adequately represent the feature space and improve the prediction.

2.1.4 Data Pre-processing

Pre-processing, transformation or pretreatment of the obtained spectral data is the essential and initial stage in chemometrics modelling. Pre-processing aims to remove any noise or disturbances caused by the baseline shift from the spectra that compensated deviation of spectra to stay close as to the Beer-Lambert law. Thus, the most important variance to chemical features can be retained during the chemometrics analysis.

Engel et al., (2013) overviewed data artefacts for the most common analytical data types. The artefact is present for a NIR and FTIR data such as baseline, scatter, and noise. The baseline, misalignment, and noise are particular to the NMR, mass spectrometry, and GC spectra.

The Savitzky-Golay algorithm is reported to be frequently used for this purpose by calculating the derivatives of the fitted polynomials for the spectral baseline issues. The second-order derivative will also eliminate a baseline slope. The combination of data preprocessing, such as derivatisation with some degree of smoothing, for example, employing Savitzky-Golay filtering, can help improve the data (Suhandy & Yulia, 2017).

Pre-processing methods can correct light-scatter effects such as Standard Normal Variate (SNV) and Multiplicative Signal Correction (MSC). These pre-processing techniques estimate the scattering coefficient by regressing the spectrum to a reference. Rinnan et al., (2009) reviewed the

close relationship between SNV and MSC pre-processing on Near Infrared (NIR) data.

The misalignment in NMR is recommended to be corrected by Correlation Optimized Warping (COW) by splitting the signal into different segments. The optimal alignment should be reached by stretching or compressing the individual segments to match the reference segments. Local alignment is particularly relevant for the pre-processing of NMR spectra, where each peak may shift in both (spectral) directions (Skov et al., 2006). Also, the amount of shifting can vary between peaks corrected to a reference.

Normalisation equates each pattern vector's components to an arbitrary constant, such as chromatography. It has been reported as a proportion of the overall integrated peak area (Lee et al., 2018). Many approaches to normalisation and scaling can be implemented. The most popular spectra normalisation includes scaling to total response, scaling to individual metabolite (or peak), log pre-processing, scaling to unit variance (auto scale), Pareto scaling, derivatisation, mean centring, and vector normalisation (Tugizimana et al., 2016).

In a study of the quality of apple juice screening by UV and NIR, Włodarska et al., (2021) used chemometrics in their comparative study. The pre-processed data were subsequently used in chemometrics techniques such as PCA to input a classification or calibration model. Then, the RMSEC, or RMSEV, was utilised to assess the quality of the pre-processing method in this model. Further, the deep learning model attained the lowest RMSEV of 0.76 %, which was 13 % lower than the PLSR on the raw absorbance data. Finally, an analysis of apple juice by NIR and PLSR for data analysis was

optimised using pre-processing and variable selection, resulting in the *adj.* R^2 > 0.9 in the predicted total phenolic compound from the spectra of apple juice.

Interpreting the main effects and interactions enables the selection of an optimal pre-processing strategy (Gerretzen et al., 2015). Combining infrared spectroscopy with chemometrics has proved its ability to provide a fast and reliable indicator for soil diagnostics (Barra et al., 2021) and authenticate heparin for pharmaceutical control applications (Burmistrova et al., 2021). However, the pre-processing steps must be taken with caution in some cases. In general, spectral pre-processing was reported to have decreased the performance of both the PLS and deep learning models (Mishra et al., 2021).

2.2 Application of Chemometrics in Lard Profiles for *Halal* Authentication

According to Islamic dietary guidelines, *halal* foods are foods allowed to be consumed by Muslims. It is vice-versa of the foods that are not permitted or called *haram*, meaning forbidden. Both *halal* and *haram* are in Arabic, which is very important to Muslim lifestyles. The market for *halal* foods worldwide is estimated to be worth USD 1.2 trillion in 2022 and is expected to be increased by 5.6% throughout the forecasting period to reach USD 3.0 trillion by 2032 (Future Market Insight, 2022). Malaysia is a country that promotes International *Halal* Food Hub, such as the fat-based food industry. However, there is often a target of lard adulteration of fat-based foods since commercial pigs are breeding in this country. Therefore, the sources of fats in *halal* food products are concerned to be misused. Noted that

the lard is a part of the pig, stated as *haram* or forbidden to be consumed by Muslims according to the Islamic dietary guidelines.

Lard is rendered or refined from pig fat adipose tissues to be added as an ingredient in commercial food products such as bakery and confectionery. The general term "fats" refers to a naturally occurring oily fluid that can be found in the body of animals like cattle and poultry, particularly when it forms a layer under the skin or around particular organs. In chemical terms, fats are groups of glycerol esters combined with other fatty acids in solid form at room temperature, which are primary building blocks of animal and plant fat, e.g., hydrogenated palm kernel oil. Lard is similar to fats of other animals that are in solid form at room temperature, then the analysis of lard for *halal* foods authentication research can be broadly represented on FAs and TAGs (Marikkar et al., 2005; Nagai et al., 2020).

The term *halal* foods authentication can be included in public foods authentication. Food authentication ensures that food complies with the information on its label. Similarly, *halal* foods authentication is subject to the *halal* label on the packaging of food products. Danezis et al., (2016) reviewed that the authentication of foods might refer to the origin (species, geographical location, or genetic), the production process (conventional, organic, conventional methods, free range), or the processing techniques (irradiation, freezing, microwave heating). In another term of *halal* foods authentication on meat and meat products, Khadijah et al., (2012) emphasised the possible analytical methods to detect pork and lard.

Rusni et al., (2020) found that the practitioners of fat-based food were quite aware of the *halal* status in pastries and other confectioneries

ingredients. The main challenges are the lack of data on lard profiles because it is complicated to identify when mixing in a food matrix as fat-based food and thus could decrease *halal* status confidences. Nevertheless, it is possible to view lard profiles as an essential determinant of assuring *halal* food status.

In order to construct models for the authenticity and traceability of the lard, it is feasible to combine chemometrics with instrumental analytical techniques under this approach. Furthermore, Ng et al., (2022) emphasised the categorization of several approaches based on validity criteria that offer valuable data for future advancements in sophisticated technologies. Therefore, applying chemometrics using multiple analytical platforms on lard profiles significantly impacts *halal* authentication by a *halal* certification authority to create awareness for the public of *halal* foods and their benefits to Muslim consumers.

2.2.1 FTIR spectroscopy in Authenticity Studies

FTIR spectroscopy has been considered an effective tool in oils and fats authentication, and its application has recently been increased with developments in chemometrics techniques (Jamwal et al., 2021). FTIR offers attractive advantages as an analytical platform due to its consistency, rapidity and, more importantly, its availability and low-cost operation.

The difference between lard and other fats (animal fat; lard, beef, chicken, and mutton fats and fish; plant oil; cod liver, canola, corn, extra virgin olive, grape seed, palm, pumpkin seed, rice bran, sesame, soybean, sunflower) was studied by Che Man et al., (2011a). The scores plot of the PCA observation demonstrated in PC-1 accounts for 44.1 % of the variation

denoted by lard, cod liver oil, corn, soybean, and rice that contribute by 2852.8 cm^{-1}. In contrast, other than oils, fats located at PC-2 account for 30.2 % of the variances contributed by 1116.8 cm^{-1} and 1236.3 cm^{-1}. Cluster Analysis (CA), an additional chemometrics technique, supported PCA for the fats grouping, indicating that lard and other fats are highly correlated. These findings would have been helpful if the author explained why lard is anti-correlated with other animal fats instead of plant fats since it has been known that animal fat has a high similarity of FAs composition (Shorland, 2012).

Saputra et al., (2018) have identified pig contaminants in a mixture of fat samples and food products. First, lard was extracted from pig products as compared to chicken. Then raw data were preprocessed using smoothing, baseline correction, then normalisation. Differences between products containing the selected wavelength-classified lard and *halal* product at 1236 cm^{-1} and 3007 cm^{-1} in the FTIR-PCA were subsequently identified as biomarkers lard. The smoothing usually can eliminate part of the noisy data, but it can distort the original signal and reduce its resolution (Xu et al., 2008).

The combinations of PCA and PLSR are the most useful applied in the spectra data and chemometrics of lard detection. Erwanto et al., (2016) conducted binary mixtures (0, 10, 20, 30, 40, 50, 60, 70, 80, 90 & 100 %) of pigskin and cow skin for calibration of *rambak* crackers. The authors selected the performances of prediction at wavenumber 1200–1000 cm^{-1}. The relationship between the actual value of lard gave the values of R^2, RMSEC, and RMSEV to be 0.997, 1.38, and 2.77, respectively. Then commercial *rambak* crackers were projected into PCA and found in the average clusters between cow, buffalo, and pig skin. A similar study on adulteration buffalo

skin by Muttaqien et al., (2016) yielded R^2 and RMSEC values at 0.961 and 2.56, respectively. Then the 3D-PCA projected *rambak* crackers containing pigskin, buffalo skin, and commercial *rambak* crackers were well separated, indicating the successful determination using FTIR combined with chemometrics strategies.

Guntarti et al., (2019) distinguished between lard and beef sausages before applying them to commercial sausages. First, the FTIR-PLSR was developed at various concentrations reference (0 % – 100 %) of lard. The calibration identified the frequency region of 1200–1000 cm^{-1} yielded R^2 and RMSEC at 0.985 and 2.09, respectively. Then sausages from the local market were projected onto the FTIR-PCA model, and it found that collected commercial samples closer to the beef sausage could indicate leads for *halal* authentication.

Rohman et al., (2011b) quantitated lard, beef fat, and their mixtures in meatball formulation. The spiked pork to meatball in the concentration range of 1.0; 3.0; 5.0; 10.0; 25.0 ratios yielded R^2, and RMSEV obtained were 0.996 and 0.712, respectively. PLSR was successfully used to quantify pork's impurity at the selected fingerprint region (1200–1000 cm^{-1}). Ahda and Safitri (2016) conducted a series of binary mixtures of lard in crude palm oil using FTIR-PLS. The finding is characteristic of lard in crude palm oil could be detected using FTIR spectroscopy combined with PLS at wavenumber 1481.22-999.05 cm^{-1} and 1793.67-1650.95 cm^{-1} where R^2, RMSEC, and RMSEV value were reported to be 0.998, 1.291% (v/v), 0.838 % (v/v), respectively. The pre-processes normal spectra give much better accuracy prediction than the 1st and 2nd derivatives. Sa'ari and Che Man (2016)

investigated the addition of lard in the chocolate formulation. FTIR-PLSR were used on lard cocoa butter and their blends (ranging from 0 – 10 %) of lard in cocoa butter. The calibration model developed and reported the values of R^2 and RMSEC to be 0.989 and 0.450, respectively.

The analysis of functional groups responsible for FTIR absorption in fat and oil samples could be related to the FAs and thermal analysis. Rohman & Che Man (2010) demonstrated four types of animal fats, lard and body fats of lamb, cow, and chicken, in quaternary mixtures were quantitatively analysed using FTIR-PLS and FAs by GC-FID. The increase in peak intensities at 3007 cm^{-1} corresponded with the high concentration of monosaturated fatty acids (MUFA) in animal fats, in the sequence of Lard > Cow > Chicken > Lamb. The calibration model was developed by comparing the PLS model using different spectral frequencies to analyse oils in ternary mixtures. The frequency region between 1500–1000 cm^{-1} was found suitable for analysing four animal fats using PLS calibration and then preprocessed selection at normalised 1st, and 2nd derivatives. The 1st derivatives spectra at the frequency region of 1500–1000 cm^{-1} are the most accurate to calibrate lard. On the other hand, normalised data found accurately calibrated lamb, cow, and chicken fats.

Kurniawati et al., (2014) demonstrated FTIR-PLSR of lard in meatball broth for the quantitative determination using a wavenumber region 1018 - 1284 cm^{-1}. The R^2 and RMSEC were 0.9975 and 1.34 % (v/v), and *adj.* R^2 and RMSEV were 0.9944 and 2.89 % (v/v), respectively. The commercial samples were collected and then classified using the FTIR-PCA model at a selected wavenumber region of 1200-1000 cm^{-1} and correlated with beef fats

at PC-1. The additional information by GC-FID on the FAs analysis of lard indicated that lard contained more oleic and stearic. On the contrary, beef fat relatively contained more palmitic and oleic acids. Ahda et al., (2020) investigated wild boar meat in beef meatballs by PCA and PLSR. In this study, FTIR-PLSR was conducted to quantitatively measure the chemical profiles for authentication after transformation spectra at 1st derivative in the regions covering the frequency region of 999-1481 cm^{-1} combined with 1650-1793 cm^{-1}.

Witjaksono et al., (2017) combined FTIR and GC-TOF/MS techniques to study the fat of pig, cow, lamb, and chicken to find possible biomarkers for lard identification. Lard and chicken fat were found to have distinct peaks at wavenumbers 1159.6 cm^{-1}, 1743.4 cm^{-1}, 2853.1 cm^{-1}, and 2922.5 cm^{-1}. According to GC-TOF/MS data, the concentrations of 1,2,3-trimethyl-benzene, indane, and undecane are 250, 14.5, and 1.28 times more outstanding in lard than in chicken fat, and 91.4, 2.3, 1.24 times higher in cow fat, respectively.

Mansor et al. (2011) demonstrated binary admixtures of lard in virgin coconut oil in various percentage concentrations ranging from 1 % to 50 % (v/v). The mixtures were assayed using fast gas chromatography with a surface acoustic wave detector (GC-SAW) system. The authors found one peak in the fast GC-SAW system chromatogram to indicate the presence of lard in virgin coconut oil, with an R^2 value of 0.934, when fitted into the second-order polynomial curve.

Rohman et al., (2012a) developed standards consisting of concentrations ranging from 1 to 60 % (v/v) of lard in palm oil using FTIR-

PLSR analysis. Both the unadulterated oils samples have high oleic acid content measured by GC-FID. However, lard had double the linolenic acyl groups compared to palm oil. Therefore, the band at frequency 3006 cm^{-1} in FTIR spectra of lard could relate to the linolenic acyl groups. The optimized frequency region used by the author was 1480–1085 cm^{-1} where R^2, RMSEC and RMSEV values were reported to be 0.998., 1.69 % and 2.87 % (v/v), respectively.

Suparman et al., (2015), conducted a study for authentication on several variants of imported chocolate products circulating on the market using FTIR and GC-MS. Their findings indicated that the spectra of lard and chocolate show a typical lard-specific difference at wavenumber region 3006.8 cm^{-1}, 1118.84 cm^{-1}, and 1097.42 cm^{-1}. Chemometric analysis of PCA and PLS calibration models using the fingerprint region 999.05-1190.63 cm^{-1} indicated the lard identification in chocolate fat. The R^2 and RMSEC were found to be 0.997 % and 1.563 %, with a minimum detection limit at a concentration of 4 %. GC-MS results showed 11, 14-eikosadienoat acid, which has been suggested to be a marker of lard that can present in a mixture of lard concentration at ≥ 10 %. The collected samples from six imported chocolate variants were found to show lard profiles marked by the appearance of 11, 14-eikosadienoat acid.

The Mahalanobis distance principle of discriminant analysis (MDA) has effectively classified lard and other fats. A study demonstrated that fingerprint regions of 1500–1000 cm^{-1} were used for quantifying and classifying lard in the mixture with vegetable oils such as canola oil, corn oil, extra virgin olive oil, soybean oil, and sunflower oil by Rohman et al.,

(2011a). Che Man et al. (2011b) selected edible fats and oils and those in biscuit formulation by MDA. In their study, FTIR spectral frequency regions at 3050-2800 cm^{-1}, 1800-1600 cm^{-1}, and 1500-650 cm^{-1} were exploited for the classification of lard and other commercial vegetable oils and animal fats,

Che Man et al., (2014) detected the presence of lard in French fries pre-fried in palm oil adulterated with lard. Prediction of lard in a blended mixture of lard and palm oil was conducted. The R^2 and RMSEC (0.979, 2.45%) were obtained with 0.5 % detection limit in the study. MDA test could distinguish between fries samples adulterated with lard and samples pre-fried with palm oils.

Al-Kahtani et al., (2017) attempted to study the adulteration of lard in commercial food products (frozen French fries) and vegetable oils (corn oil, sunflower oil, palm oil, and olive oil). The binary mixtures were prepared from 1 to 20 % (i.e., 1, 5, 10, and 20 % of lard). FTIR analysis detected levels of lard as low as 1% in all mixtures where some important FTIR spectral regions were observed at 1405-1365 cm^{-1}, 1260-1198 cm^{-1}, 935-910 cm^{-1}, 877-857 cm^{-1} and 857-833 cm^{-1}.

Upadhyay et al. (2018) detected pig body fat in pure ghee. The author used pure mixed ghee that was spiked with pig body fat at 3, 4, 5, 10, and 15%. Some wavenumber ranges at 3030–2785 cm^{-1}, 1786-1680 cm^{-1}, and 1490-919 cm^{-1} were selected based on differences in the spectra obtained. Separate clusters of the samples were obtained by employing PCA at a 5 % significance level on the chosen wavenumber range. SIMCA classified 90 % of the samples into their respective class containing pure ghee and pig body fat. FTIR-PLSR could detect spiking of pig body fat in pure ghee even at a

level of 3 % by calibration, with a validation value of R^2 > 0.99. The application chemometrics and FTIR on profiles lard by various studies are tabulated in Table 2.1.

Table 2.1: Application Chemometrics and FTIR on Profiles Lard.

Application	Chemometrics	Data preprocessing	Wavenumbers	References
Discrimination lard and other fats (animal fat and edibles oils).	PCA	Mean center	2852.8 cm^{-1}	Che Man et al., (2011a)
Discrimination pig in a mixture of fat samples and food products.	PCA	Smoothing, Baseline Correction Normalise	1236 cm^{-1} & 3007 cm^{-1}	Saputra et al., 2018
Discrimination cow, buffalo pig and. commercial *rambak* crackers.	PCA, PLSR (R^2 , RMSEC, & RMSEV = 0.997, 1.38, & 2.77), PCA projection.	Normalise	1200–1000 cm^{-1}.	Erwanto et al., (2016)
Discrimination pigskin, buffalo skin, and commercial *rambak* crackers	PLSR (R^2, RMSEC = 0.961 & 2.56), PCA projection.	Normalise	1200–1000 cm^{-1}.	Muttaqien et al., (2016)
Quantification pork fat (PF), beef fat (BF), and their mixtures in meatball formulation.	PLSR (R^2, RMSEV = 0.996, 0.712)	Normalise	1200– 1000 cm^{-1}.	Rohman et al., (2011b)
Discrimination of the binary mixtures of lard in crude palm oil (CPO)	PLSR (R^2 , RMSEC, RMSEV = 0.998, 1.291, 0.838)	Normalise 1st and 2nd derivatives	1481.22-999.05 cm^{-1} & 1793.67-1650.95 cm^{-1}	Ahda & Safitri (2016)
Discrimination lard in the chocolate formulation. (0 – 10%)	PLSR (R^2, RMSEC = 0.989 & 0.450)	Normalise	1500–1000 cm^{-1}	Sa'ari & Che Man (2016)
Discrimination lard and body fats of lamb, cow, and chicken.	PLSR	Normalise 1st, & 2nd derivatives.	1500–1000 cm^{-1}	Rohman & Che Man (2010)
Discrimination lard in meatball broth, commercial samples.	PCA, PLSR (R^2, RMSEC 0.997 & 1.34; *adj.* R^2 & RMSEV 0.994 & 2.89)	Normalise	1200-1000 cm-1 1018-1284 cm^{-1}	Kurniawati et al., (2014)

Table 2.1, continued

Discrimination pig, cow, lamb, and chicken.	PCA	Normalise	1159.6 cm^{-1}, 1743.4 cm^{-1}, 2853.1 cm^{-1}, & 2922.5 cm^{-1}	Witjaksono et al., (2017)
Quantification lard in palm oil (1 to 60%).	PLSR (R^2, RMSEC & RMSEV = 0.998, 1.69, 2.87)	Normalise	1480 – 1085 cm^{-1},	Rohman et al., (2012a)
Discrimination variants of imported chocolate products.	PLSR (R^2, RMSEC = 0.997 & 1.563)	Normalise	1097.42 cm^{-1}, 999.05 - 1190.63 cm^{-1}	Suparman et al., (2015)
Discrimination lard in the mixture with vegetable oils.	MDA & FTIR-PLS	Normalise	1500 – 900 cm^{-1}	Rohman et al., (2011a)
Discrimination lard selected edible fats and oils in biscuit formulation.	PLSR	Normalise	3050-2800 cm^{-1}, 1800-1600 cm^{-1}, 1500-650 cm^{-1}	Che Man et al., (2011b)
Discrimination of lard in French fries pre-fried in palm oil adulterated with lard.	PLSR (R^2, RMSEC = 0.979, 2.45%)	Normalise	1500 – 900 cm -1	Che Man et al., (2014)
Discrimination lard in frozen French fries and vegetable oils 1 to 20% (1, 5, 10 and 20% of lard).	FTIR-PLSR	Normalise	1405-1365 cm^{-1}, 1260-1198 cm^{-1} , 935-910 cm^{-1}, 877-857 cm^{-1} & 857-833 cm^{-1}	Al-Kahtani et al., (2017)
Discrimination pig body fat in pure ghee. Pure mixed ghee was spiked with pig body fat (3, 4, 5, 10, and 15%).	FTIR-PLSR SIMCA was classified 90 %	Normalise	3030–2785 cm^{-1}, 1786–1680 cm^{-1}, 1490–919 cm^{-1}	Upadhyay et al., (2018)

2.2.2 NMR (^{1}H & ^{13}C)

High-resolution NMR spectroscopy techniques have been used to effectively analyze and structure unsaturated lipids (Alexandri et al., 2017). The limited spectrum width spanned by protons is the fundamental disadvantage of ^{1}H-NMR spectroscopy in oil analysis. As an alternative to ^{1}H-NMR, ^{13}C-NMR spectra have few advantages. The ^{13}C spectrum is

particularly informative since it provides a large number of signals covering a wide range of chemical shifts (Lankhorst & Chang, 2018).

Fang et al., (2013) recognised that the characterisation of oils and fats for classification, prediction, and adulteration detection is essential to consumers from commercial and health perspectives. ^{1}H-NMR, GC-MS fingerprinting, and chemometrics successfully differentiated oils and fats. PCA of both techniques showed group clustering of 14 types of oils and fats. PLS-DA and OPLS-DA using GC-MS data had excellent classification sensitivity and specificity compared to models using ^{1}H-NMR data. The study also demonstrated that PLS models were successfully established to detect as low as 5 % lard and beef tallow spiked into canola oil.

Fadzillah et al., (2015) developed multivariate calibration of PLSR to model the relationship between the actual value of lard and the predicted value using the data from ^{1}H-NMR spectroscopy. The model yielded R^2 of 0.998, RMSEC of 0.0091 %, and RMSEV of 0.0090 %, respectively. In addition, this study demonstrated that the PLS model was good, as the intercepts of R^2Y and Q^2Y were 0.0853 and −0.309, respectively.

Little study has been published on the identification of lard using NMR, although this approach has been successfully used to identify FAs in other oils for authentication reasons. For instance, Lia et al., (2020) demonstrated that ^{1}H and ^{13}C-NMR spectra successfully differentiated Maltese and non-Maltese and EVOO samples by exploiting PLS-DA and artificial neural networks (ANN).

2.2.3 Chromatography (GC-FID, GC-MS, & HPLC)

Several analytical methods, including liquid chromatography (LC) and gas chromatography (GC), especially when combined with mass spectrometry, have been presented in the recent decade for rapid screening or selective confirmation of the quality and authenticity of lard. In addition, a chemometric technique is frequently used in investigations on complex chromatograms and chemical signatures, thus could serve as providing reliable qualitative or classification and quantitative or calibration of the multivariate. However, GC-FID detection has traditionally been a preferable option for lard analysis since the FAs and TAGs in lard can be monitored and compared to reference standards.

Dahimi et al., (2013) attempted to differentiate lard adulteration at very low beef and chicken fats using GC-FID combined with PCA and *k*-mean CA application. The measurements were made from pure lard, beef fats, pure chicken fat, and fats which have been adulterated with different concentrations of lard (0.5 %, 1%, 2%, 3%, 4%, 5% & 10 %) in beef and chicken fats). The results showed that lard contains higher FAs of *cis* C18:2 and low C16:0 FA, but oppositely for beef tallow and chicken fat. Therefore, PCA could classify lard, beef fat, chicken fat, and the mixtures of lard and beef tallow, lard, and chicken fat, even at a lower concentration level of 0.5 % (w/w). On the other hand, the *k*-mean CA can only classify pure lard, pure chicken fat, and pure beef.

PCA is an important technique in FAMEs data analysis and plays a crucial role in FAs distribution. FAs data from GC-MS was used with a PCA model that assists in discriminating lard from lard olein, lard stearin, and

chicken fat (Naquiah et al., 2013); partial acylglycerols of lard with those of chicken fat, beef fat, and mutton fat (Naquiah et al., 2016); monoacylglycerols (MAG) and diacylglycerols (DAG) of lard with those of six commercial partial acylglycerols (Nasyrah et al., 2014). These three studies also show that C18:0, C18:1, and C18:2 are the most discriminating FAs in clustering lards and other fats. The findings supported that extracted lard from different processed products, such as stearin, and partial acylglycerols, have similar FAs composition. However, not all FAs in PCA models could well discriminated lard against other fats. Furthermore, the relationship between FAs and lard has not been firmly established since the most findings stated that lard and fat chicken clusters are correlated.

The OSC-PLS or orthogonal projections to latent structures analysis OPLS has been applied to the FAME data. Hanafy et al., (2021) studied *sn*-2 fatty acids (*sn*-2 FAs) to identify lard adulteration in the fish feeds. The identification of *sn*-2 C16:0 FAs that structured of TAG in middle position of glycerol backbone were conducted by OPLS-DA, OPLSR and ANOVA of the cross-validation or CV-ANOVA as diagnostic tools. Both models showed similar performances in machines learning, namely SVM, RF, and ANN. The OPLS-DA identified *sn*-2 of C16:0 FAs as markers of lard. Noted that the comparisons between two chemometrics models using CV-ANOVA were adopted from the pioners, Eriksson et al., (2008).

Recently, the development of omics data has influenced the data analysis of edible oils and fats. The term metabolite profiling is also used in the study of FAMEs on lard. Heidari et al., (2020) conducted a metabolite profiling approaches using GC-MS to identify and quantify lard adulteration

in plant oil. The authors demonstrated that the 3D-PCA scores plot could correctly identify and clusters olive oil, sunflower oil, sesame oil, lard, and adulterated samples through the changes in their FAMES profile. Methyl myristate, methyl palmitate, methyl oleate, and methyl stearate were selected as discriminant markers to identify and quantify lard adulteration even as low as 5% (w/w), with errors ≤ 2% in the comparison of the absolute concentration.

Although the PLS-DA was successfully used to study lard differences in most published studies, a recent study showed that the other algorithms outperformed PLS-DA. Azizan et al., (2021) attempted to detect lard alduteration in wheat biscuits using chemometrics and GC-MS. It was found that the RF which outperformed PLS-DA in sample classification. Features selection using RF proposed C18:3n6 as the potential biomarker in discriminating pure wheat biscuits and lard-adulterated biscuits. In addition, the author reported that PCA and hierarchical cluster analysis (HCA) could cluster lard, wheat biscuits and lard-adulterated samples from their FAs distribution. Therefore, from this study the trial of several algorithms becomes a necessity on lard data.

Rohman et al., (2012b) differentiated lard and other animal fats (beef, mutton, and chicken fats) and cod liver oil has using HPLC and PCA. Lard was found to be separated along the opposing side in the PC-1 and PC-2 of the otal variance 82 % have highly correlated with chicken fats. This lard indicated to be similar to chicken fat in terms of TAG composition. In addition, palmitooleoolein and palmitooleostearin of the lard and chicken fats contributed to differences in the PCA model.

Ahda et al., (2016) attempted to distinguish between beef and pork meatball using HPLC combined with PCA. However, the authors have highlighted that the application of HPLC has a problem with lard detection in meatballs because it could not control the hydrolysis of TAG. Noted that the hydrolysis of TAG can also be detected using HPLC-UV and giving the interference for authentication.