Food powders classification using handheld Near-Infrared Spectroscopy and Support Vector Machine

Mohamed Yasser Mohamed\textsuperscript{1,3}, Mahmud Iwan Solihin\textsuperscript{1,4}, Winda Astuti\textsuperscript{2,5}, Chun Kit Ang\textsuperscript{1,6} and Wan Zailah\textsuperscript{1,7}

\textsuperscript{1}Faculty of Engineering, UCSI University, Jalan Puncak Menara Gading, Taman Connaught, 56000 Kuala Lumpur, Malaysia
\textsuperscript{2}Computer Engineering Department, Faculty of Engineering, Bina Nusantara University, Jakarta, Indonesia
E-mail: \textsuperscript{3}mohamedyasser63@gmail.com, \textsuperscript{4}mahmudis@ucsiuniversity.edu.my, \textsuperscript{5}winda1977@gmail.com, \textsuperscript{6}angck@ucsiuniversity.edu.my, \textsuperscript{7}zailah@ucsiuniwerst.edu.my

Abstract. The aim of this study is to build a classifier model based on spectra data collected using handheld spectrometer that can classify between different types of food powders (flour and starch). A total of 70 samples were prepared from three different types of flour (whole wheat, organic wheat, and rice flour) and two different types of starch (corn and tapioca starch). Hand-palm size handheld spectrometer is used to record the spectrum of each sample, the spectrometer has wavelength range of 900nm to 1700nm. The spectra data is pre-processed using gaussian smoothing to filter the data from noise and unrelated information. Multivariable data analysis method as principle component analysis (PCA) is used to eliminate irrelevant data and reduce the number of variables to three principle components for easier analysis and visualization. Support vector machine (SVM) is used to build a classification model. The training/calibration of the model was done by using 80\% of the dataset while the remaining 20\% was for testing the model. The results show that with proper pre-processing and PCA, classification of 100\% accuracy can be achieved. This study indicates the potential future application of this approach for rapid detection in food powders fraud and adulteration.

1. Introduction
Food authenticity has become a concern of the world. Authenticity is important as it is what keeps the demand and the sustainability of the industry. There are two types of food authenticity first is related to the production and secondly is to the description. With respect to production means for example organic or non-organic while with respect to the description mean adulteration [1]. Food authentication means verification of the food with its labeling [2]. Food adulteration became a common practice which comes from two reasons first is the profit as the demand on food products is high and the other reason is that it is hard to detect these adulterants [3]. Food fraud is the deliberate adulteration of the product, according to UK’s food standards agency there are two main types of fraud first is the harmful food sale while the second type is food misdescription [2].

Near infrared spectroscopy (NIRS) has developed rapidly and now became a mature technology. It has proven remarkable improvement in analytical applications in comparison with other analytical
method [4], [5]. Combining NIRS with chemometrics allows it to analyze different types of samples rapidly [5]. Greatest advantage of NIRS is can virtually study any sample [6]. NIRS has been widely used in food sector as is provides lower cost for recording the spectra in an nondestructive way without the need for sample preparation [7]. NIRS can be used to do qualitative and quantitative analysis of the sample. NIRS is more suitable for quantitative analysis than it is qualitative analysis of food because of the overlapping of the appearance [6].

NIRS is used to classify between different types of food powder as flour and starch. The food powders used are organic wheat flour, whole wheat flour, rice flour, tapioca starch and corn starch. NIRS was combined with PCA to extract the most relevant data variation and as dimension reduction tool [8].

2. Experimental Setup

2.1. Spectra data collection

For this study, five different types food powder were selected which are organic wheat flour, whole wheat flour, rice flour, tapioca starch and corn starch. For each powder five samples were taken and placed in plastic packets of size 7cm x 11.2cm regardless of the weight. total of 25 samples were prepared to be scanned by the spectrometer.

Samples were scanned by using handheld NIR spectrometer with wavelength range of 900nm to 1700nm. The device as shown in Figure 1 is developed using Texas Instruments electronic board. The samples were scanned in reflectance mode. Each sample spectrum was measured for 2 seconds in reflectance mode. Each sample was scanned three times in different environment, i.e. under different room lighting conditions, of total 75 spectra of 25 samples. Spectral data was transferred in a form of datasheet to Orange software for analysis (https://orange.biolab.si/).

The data measured by the spectrometer contains irrelevant data and affected by noise and scattering [9]. Per-processing spectral data is essential to get efficient and accurate model [9], [10]. For this paper, Savitzky-Golay (SG) filter technique is used to pre-process the spectra of window size of 15 and polynomial order of 2. It is effective for eliminating noise from the spectra [11].

![Figure 1](image)

**Figure 1.** The handheld (hand palm size) spectrometer scanning the powder sample

2.2. Principle component analysis

Principle component analysis (PCA) is most commonly used method to extract the most relevant data from the spectra. PCA reduces the dimensionality of the data by linear fitting [10]. PCA makes new variables which represent most of the variation of the original data while reducing dimensionality these variables are called principle components (PCs) [5], [12]. PCs are orthogonal linear combination of the original data [8]. The First PC is the one with have the minimum projection distance from the original data, this line will go through the maximum variation of the projection of the data. The second PC
(Subsequent) are selected in a similar way with more requirements as there must be no correlation with the previous, this means PCs that could possibly exist are equal to the number of the features or the samples [13]

2.3. Support vector machine (SVM)
Support vector machine is one of the supervised learning machine learning algorithms. SVM was firstly introduced by Vapnik (1995). It is used for classification and regression applications. SVM constructs a hyperplane in a higher dimensional space. Linear support vector machine looks for a hyperplane that can separate the data accurately. The best hyperplane has the maximum shortest distance from the closest training samples of each class \(d_+ \) for class \( \{+1\} \), \(d_- \) for class \( \{-1\} \). The Margins are defined by the maximum distance \((d_+ - d_-)\). Classification performance increases by maximizing the margins. [14]. Kernel is the mathematical functions that are being used by Support Vector Machine. Some data are non-linearly separable so SVM can be extended for nonlinear classification. There are different types of kernel functions that can be used with SVM for example Linear, non-Linear, polynomial, radial basis function (RBF). The most commonly used is radial basis function (RBF) (Dataflair Team 2017).

\[
\text{linear: } K(x_i,x_j) = x_i^T x_j .
\]
\[
\text{Polynomial: } K(x_i,x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0.
\]
\[
\text{Radial basis function (RBF): } K(x_i,x_j) = \exp(-\gamma|x_i - x_j|^2), \gamma > 0.
\]

where, \(\gamma\), \(r\), and \(d\) are kernel parameters.

3. Results and Discussions
All the recorded spectra as shown in Figure 2 is the raw absorption spectra data of 75 samples of five different food powder types acquired from the spectrometer of wavelength ranging from 900nm to 1700nm. This spectral data is then pre-processed by Savitzky-Golay (SG) derivative filter to eliminate the noise from the spectra data of window size 15 and polynomial of order 2. As can be seen from Figure 3, the spectra of each class are more defined in second half of the wavelength.

Figure 4 shows the plot that two PCs of PCA applied covering 96.8% of the total variance of the original data, PC1 covering 84.2% of the total variance while PC2 covering 12.6%. Figure 5 shows the scatter plot of PC1 and PC2. It is clearly indicated that the data from five different food powder types (whole wheat flour, organic wheat flour, rice flour, corn starch and tapioca starch) are separated efficiently. These processes are performed in Orange data mining free software (https://orange.biolab.si/).

![Figure 2. Raw spectral data](image-url)
Figure 3. Pre-processed spectra data with gaussian smoothing and SG derivative order two

Figure 4. PCA variance graph
The pre-processed data is then divided into two subsets where 80% of the data is used to train the SVM classifier model and the remaining 20% of the data are used for testing. Different $C$ (Cost) values are tested with different kernels to get the best classification accuracy. Table.1 shows the results of each kernel in classification. From Table.1, RBF and polynomial kernels can give 100% classification accuracy. On the other hand, linear kernel showing higher accuracy with $C=1$ and lower accuracy with higher $C$ values.

Table 1. SVM classifier results with different kernels

| SVM         | $C=1$  | $C=10$ | $C=100$ |
|-------------|--------|--------|---------|
| Linear      | 93.3%  | 86.7%  | 80%     |
| Polynomial  | 100%   | 100%   | 100%    |
| RBF (G=auto)| 100%   | 100%   | 100%    |

4. Conclusions
This study demonstrates the qualitative analysis (classification) of food powders using handheld NIR spectroscopy device. A total of 100 spectra data from five different food powder types (whole wheat flour, organic wheat flour, rice flour, corn starch and tapioca starch) are used to build a classification model using support vector machine. Gaussian smoothing and Savitzky-Golay derivative filter was efficiently eliminating the noise from the raw spectra. PCA was combined with machine learning algorithm to improve the classification accuracy by reducing the variables to two PCs. The results from this study shows the capability of SVM to build efficient classification model. This study also indicates the future potential application of this approach for rapid detection in food powders fraud and adulteration.
Acknowledgement
The authors would like to express the gratitude to Centre of Excellence for Research, Value Innovation and Entrepreneurship (CERVIE) – UCSI University for supporting this project under PSIF research grant scheme.

References
[1] D Cozzolino, “Near Infrared Spectroscopy and Food Authenticity,” in Advances in Food Traceability Techniques and Technologies: Improving Quality Throughout the Food Chain, Elsevier, 2016, pp. 119–136.
[2] C Alamprese, “Food Authentication,” in Spectroscopic Methods in Food Analysis, Boca Raton, FL: CRC Press, Taylor & Francis Group, 2017. CRC Press, 2017, pp. 327–352.
[3] J Spink and D C Moyer, “Backgrounder: Defining the Public Health Threat of Food Fraud [In partial fulfillment of a NCFPD grant],” pp. 1–7, 2011.
[4] M Cocchi, C Durante, G Foca, A Marchetti, L Tassi, and A Ulrici, “Durum wheat adulteration detection by NIR spectroscopy multivariate calibration,” Talanta, vol. 68, no. 5, pp. 1505–1511, Feb. 2006.
[5] C Zhang and J Su, “Application of near infrared spectroscopy to the analysis and fast quality assessment of traditional Chinese medicinal products,” Acta Pharm. Sin. B, vol. 4, no. 3, pp. 182–192, Jun. 2014.
[6] B Stuart and S Preface, Infrared Spectroscopy: Fundamentals and Applications, vol. 8. 2004.
[7] H Yu, Y Zhou, X Fu, L Xie, and Y Ying, “Discrimination between Chinese rice wines of different geographical origins by NIRS and AAS,” Eur. Food Res. Technol., vol. 225, no. 3–4, pp. 313–320, Jun. 2007.
[8] N O Mahony, T Murphy, K Panduru, D Riordan, and J Walsh, “Machine learning algorithms for estimating powder blend composition using near infrared spectroscopy,” in 2018 2nd International Symposium on Small-scale Intelligent Manufacturing Systems (SIMS), 2018, pp. 1–6.
[9] H Cen and Y He, “Theory and application of near infrared reflectance spectroscopy in determination of food quality,” Trends Food Sci. Technol., vol. 18, no. 2, pp. 72–83, Feb. 2007.
[10] X Wang, Near-infrared spectroscopy for food quality evaluation. Elsevier Inc., 2019.
[11] X Wang and G Zhou, “Study on pretreatment algorithm of near infrared spectroscopy,” in IFIP Advances in Information and Communication Technology, 2011, vol. 345 AICT, no. PART 2, pp. 623–632.
[12] H Zhang, Q Yang, and J Lu, “Classification of washing powder brands using near-infrared spectroscopy combined with chemometric calibrations,” Spectrochim. Acta - Part A Mol. Biomol. Spectrosc., vol. 120, pp. 625–629, 2014.
[13] V Powell and L Lehe, “Principal component analysis explained visually,” Disq. Available http://setosa.io/ev/principal-component-analysis/. [Accessed 11 July 2017], 2015.
[14] O Devos, C Ruckebusch, A Durand, L Duponchel, and J-P Huvenne, “Support vector machines (SVM) in near infrared spectroscopy (NIR) scope: Focus on parameters optimization and model interpretation,” Chemom. Intell. Lab. Syst., vol. 96, no. 1, pp. 27–33, Mar. 2009.
[15] Dataflair Team, “Kernel Functions-Introduction to SVM Kernel & Examples - DataFlair,” 2017.