Differentiation of Organic and Non-organic Apples Using Near Infrared Reflectance Spectroscopy – A Pattern Recognition Approach

Weiran Song[1], Hui Wang[1], Paul Maguire[2], Omar Nibouche[1]

[1] School of Computing and Mathematics, [2] School of Engineering, Ulster University
BT37 0QB, Newtownabbey, County Antrim - UK

***Abstract*—With the organic food market on the rise, organic food fraud has become an issue to consumers, producers and the market. Traditional methods of food quality determination are time consuming and require expert laboratory analysis. Recent studies based on spectroscopic analysis have shown its potential effectiveness in non-destructive food analysis. This paper explores the use of low cost Near Infrared Spectroscopy (NIRS) combined with a pattern recognition approach for the differentiation of organic and non-organic apples. The spectra of organic and non-organic Gala apples are measured using a low cost and portable NIR Spectrometer. A pattern recognition pipeline is proposed, where spectra data are pre-processed and then classified into organic and non-organic. Baseline correction and normalization are used in pre-processing, and Partial Least Squares Discriminant Analysis (PLS-DA) is used for classification. The experimental results show that the apple samples can be classified into organic and non-organic ones with accuracies of over 96%. The results and the fact the NIR spectrometer used was low cost and portable suggest this is potentially a cost effective solution to the detection of organic food fraud.**

Keywords*—NIR *spectroscopy; Pattern recognition; PLS-DA; Organic; Apple

I. INTRODUCTION

Organic farming uses only natural pesticides and fertilisers, and adheres to a high standard of animal welfare and crop rotation. Generally, organic products have the benefit of lower pesticide or hormone exposure, are fresher and richer in certain nutrients and are generally better for the environment. Organic farming has been growing for years, with almost 0.5 million additional hectares organically managed in 2014 compared to the previous year and its global market value is currently at $80bn [1]. Among organic foods, fruits and vegetables have a 43.3% share of the market in US and 23% in UK [2, 3]. Organic foods are generally more expensive than their non-organic counterparts. This, coupled with the fact that routine distinction of organic from non-organic is often not possible, has led to fraudulent marketing (organic food fraud) including mislabelling and mixing [4, 5].

Non-organic food is differentiated from the organic by the presence of unwanted contaminating substances. Unfortunately detection of these unwanted substances is challenging and requires expensive laboratory facilities, such as Liquid Chromatography-Mass Spectrometry (HPLC-MS) [6], which are not readily available in the food distribution chain or near point of sale. Consumer confidence is a very important aspect of the food market; thus the availability of a low cost analysis system, suitable for field use within the food industry and which can reliably detect organic from non-organic products would be very valuable. However, standard NIR spectroscopy on its own requires expensive and laboratory-based equipment to obtain the required detection accuracy. In this work, we investigate the use of pattern recognition techniques coupled with low cost and hence lower quality spectra in order to develop a portable detection system. Organic apples are among the more popular fruits but the non-organic variety is vulnerable to high levels of pesticide contamination [7] and the need to protect against mislabelling is high. In this paper we report on the evaluation of this portable detection system in distinguishing organic from non-organic apples.

Spectroscopic analysis involves the measurement and analysis of optical intensity versus wavelength (or frequency) spectra produced when matter interacts with electromagnetic radiation. Chemical compounds reflect, absorb and/or transmit-light at different wavelengths and are often characterised by a wavelength fingerprint. However, low signal to noise ratios, interferences from other chemicals and the background matrix can lead to a requirement for complex instrumentation and careful sample preparation. With its instantaneous and non-destructive nature, spectroscopy is an attractive approach for the quantitative and qualitative analysis of food and is often combined with chemical pattern recognition (termed chemometrics) methods. For quantitative analysis, techniques such as Partial Least Squares (PLS), Support Vector Regression (SVR) and Principle Component Analysis (PCA) have been used to determine the level of specific pesticide residuals with data obtained from Surface-Enhanced Raman Spectroscopy (SERS), Laser-Induced Breakdown Spectroscopy (LIBS) and Fourier Transform Infrared (FTIR) spectroscopy measurements [8-10]. For qualitative analysis, apples of different variety or quality acquired by NIRS can be classified up to 98% accuracies with PLS-DA [11] and fuzzy discriminant c-means [12]. The algorithm of Soft Independent Modelling of Class Analogy (SIMCA) is used in [13] to determine fungicides residual on apples permitting over 99% of all the samples to be correctly classified. With regard to the definition of organic apples, the residual pesticide level is the most important component. However attempts at individual residual pesticide determination via specific chemical detection have shown the limitations of this approach due to the range of pesticide varieties and their combinations [14]. Recently the

separation of organic from non-organic tomatoes has been investigated using a combination of spectroscopy, PCA for dimensionality reduction, Linear Discriminant Analysis (LDA) and PLS-DA for classification. Such a combination can achieve differentiating results ranging between 95% and 100% [15]. For all of the above examples, the measurements were laboratory based and therefore not suitable for low cost portable field operation.

In this paper, we use a low cost portable NIR spectrometer to collect spectra from sets of Gala apples. The spectra contain 512 wavelengths i.e. variables and thus are considered a high dimensionality problem. To differentiate organic and non-organic categories pattern recognition techniques based on PLS-DA with appropriate pre-processing techniques are used. The classification accuracies which are up to 98% show that organic and non-organic Gala apples can be efficiently differentiated.

II. A PATTERN RECOGNITION FRAMEWORK

The pattern recognition framework in our work aims to establish a model for detecting one from two categories, precisely, to differentiate organic and non-organic Gala apples. A pattern recognition framework on the differentiation of NIR spectral data is shown in Fig.1.

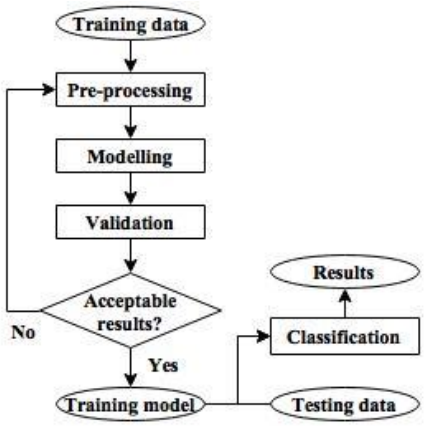

Fig. 1. A pattern recognition framework for the differentiation of NIR spectral data

A. Pre-processing

Pre-processing can be regarded as a series of techniques applied prior to data analysis. These techniques are aimed at removing physical phenomena in spectra in order to improve the performance of a classification or regression model [16]. Dimensionality reduction techniques are often included in the pre-processing phase. Typical pre-processing techniques used on NIR spectral data can include:

- Smoothing: a general path to reduce the effect of noise and capture the important trend in spectral data analysis, such as Moving Average (MA) and Savitzky-Golay (SG).
- Baseline correction: another de-noising method that removes background noise or unnecessary peaks caused by different environmental factors.
- Normalization: adjusting values measured on different scales to a notionally common scale. Standard Normal Variate (SNV) and Multiplicative Signal Correction (MSC) are two typical methods applied in NIR spectral data.

In this paper, we select baseline correction [17] and SNV normalization for pre-processing. The two main parameters in baseline correction, namely window size and step size, are determined in the validation phase, as detailed in the experimental section.

B. Modelling

Modelling is a learning stage in which patterns are established in training data and are later extended, generalized and sought in testing data. PLS is a statistical regression approach which has shown great effectiveness for the analysis of high-dimensional as well as multicollinearity spectral data. PLS modelling uses a Latent Variables (LVs) approach that seeks to maximize the covariance between predictor variables and response whilst reducing the dimensionality at the same time. It naturally lends itself to be extended for classification purposes, yielding techniques such as PLS-DA [18] in which a dummy matrix to represent the categorical response is used. PLS-DA is a chemometric method and has been widely applied to classify data obtained from laboratory-based NIRS [19]. There are various algorithms that implement PLS, among which NIPALS [20] and SIMPLS [21] are widely used. As SIMPLS is a non-iterative algorithm that can provide a fast and efficient computation, we implement PLS-DA based on SIMPLS approach in this paper.

C. Validation

The model constructed as above has various parameters for the pre-processing and modelling parts. For example, the baseline correction window size used in pre-processing and the number of LVs used in PLS-DA in modelling. We use 10-fold cross validation to select the parameters which lead to the optimal performance, and then use the resulting model on test data, which will produce the evaluation results.

III. EXPERIMENTS

A. NIRS measurement and data acquistion

Apples were scanned in the reflectance mode using a NIR spectrometer (NIRQuest512 spectrometer, Ocean Optics, Inc., United States) equipped with an InGaAs detector and having a wavelength range of 901.06-1721.242 nm with a 1.65 nm interval [22]. NIR spectra, each with a dimensionality of 512 were collected with the OceanView software [23].

A total of 60 organic and non-organic Gala apples were obtained from two local supermarkets on the same day. The ratio of organic to non-organic was 20:18 and 10:12, from the 1st and 2nd supermarkets respectively. All of the apples were defect free and no surface preparation was carried out prior to NIRS analysis. The experiment consisted of a measurement / data collection stage followed by a data selection stage:

- Stage-1: For each apple, four areas are selected for scanning. In each area, three spectra are taken and averaged to represent a sample corresponding to the

area. Therefore, 240 samples are acquired for classification.

- Stage-2: From the data collected in Stage-1, one area is randomly selected out of four in each apple. As a result, there are 60 samples to be classified.

B. *Pattern recognition procedure*

Given Stage-1 and Stage-2, a 10-fold cross validation strategy is implemented to classify every spectrum. This strategy randomly partitions spectra into 10 equal sized groups. A single group is retained as the validation set while the remaining 9 groups are used as a training set. The cross-validation process is then repeated 10 times with each of the 10 groups used exactly once as the validation set. The experimental steps along with a pattern recognition framework are outlined below:

- 60 apples are randomly divided into 10 groups of 6 apples each. Each group is alternately selected as testing set while the remaining groups are used for training following above criteria.
- For the classification in each group, a 10-fold cross validation is undertaken to select the appropriate pre-processing techniques and their corresponding parameters as well as the PLS-DA latent variables within the training sets. The average validation accuracies against the number of LVs are then obtained. It is noted that the same indices are used in the comparison of pre-processed and raw data accuracy during validation.
- The optimal parameters identified in validation are directly implemented in the classification phase. Ten final classification results are returned, corresponding to the 10 groups.

A baseline correction function [17] and SNV normalization were chosen to pre-process raw data and the effect of pre-processing is shown in Fig. 2 by comparison with raw spectra. In the latter, there is a large variation in intensity at different wavelengths for both types of apple. However, after pre-processing the intensity variation is greatly reduced.

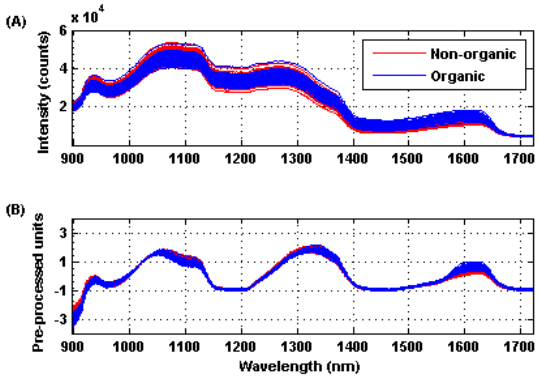

Fig. 2. NIR spectra of Gala apples (240 spectra): (A) Raw spectra; (B) Spectra are pre-processed by baseline correction and SNV.

IV. Results

In the training phase, suitable combinations of pre-processing along with LV parameters are identified initially. Pre-processing via baseline correction [17] and SNV normalization improve the validation results compared with unprocessed data, as shown in Fig.3. The baseline correction window size is 200 and 150 in Stage-1 and Stage-2, respectively, while the step size is 50 in both stages. The average accuracy after pre-processing is higher for PLS-DA using up to three LVs. For a greater number of LVs, up to 10, the accuracy is comparable to or slightly lower than raw data while above 10, the accuracy is greater for pre-processed data. The scope for varying pre-processing conditions and the selection of LV's is considerable and hence determination of the optimal set was not feasible. Nevertheless, in this work, the chosen pre-processing techniques and PLS-DA outperform the results obtained from raw unprocessed spectral data.

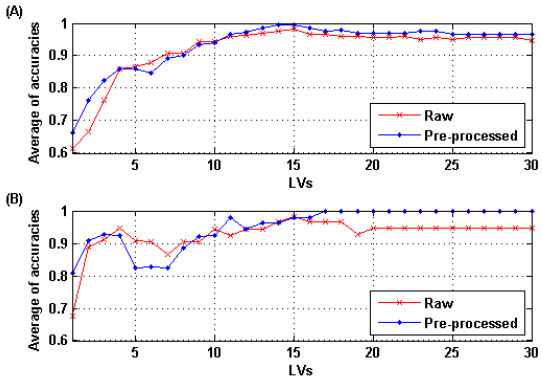

Fig. 3. PLS modelling results of raw and pre-processed spectral. The overall accuracies are averaged in validation phase with LVs ranging from 1 to 30: (A) 240-sample of Stage-1; (B) 60-sample of Stage-2.

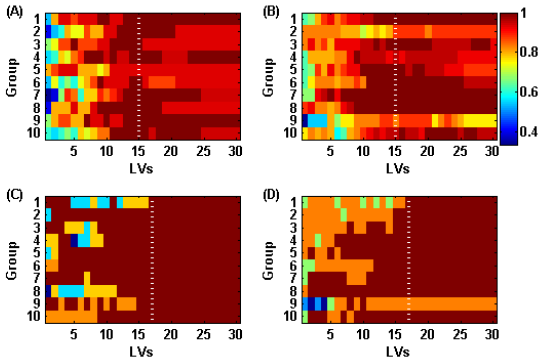

Fig. 4. Colour map of average cross validation results on 'leave each group out' and classification results on each group with LVs ranging from 1 to 30: Validation (A) and Classification (B) in Stage-1; Validation (C) and Classification (D) in Stage-2. The colour bar indicates the level of accuracy. The dashed line is the optimal results achieved in validation and applied in classification.

The results of the 'leave one group out' validation are presented in Fig.4 A and C with a colour map displaying the level of accuracy obtained in each validation. Generally, the best accuracies (red hue) are obtained for LV numbers above 10. By comparing the average accuracies of 10 groups, the number of LVs for testing is set to 15 and 17 for Stage-1 and

Stage-2 respectively with accuracies of 99.5% and 100%. Using these parameters the average accuracy for 10-fold cross classification is 96.25% in Stage-1 and 98.33% in Stage-2 (Fig.5). For Stage-1 the accuracy peaks at 96.67% for 16 LVs before gradually decreasing to 93%. In Stage-2, the accuracy reaches its maximum of 98.33% at 17 LVs and then remains constant. These results demonstrate that 231 out of 240 samples in Stage-1 and 59 out of 60 apples in Stage-2 are correctly differentiated.

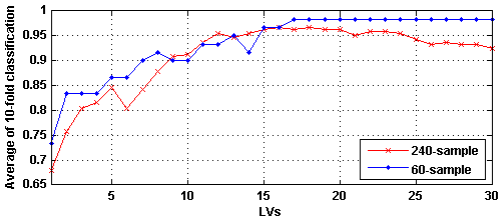

Fig. 5. The average of 10-fold cross classification results in Stage-1: 240-sample and Stage-2: 60-sample.

V. CONCLUSIONS

This paper explores the combination of low cost portable NIRS and pattern recognition to differentiate organic and non-organic Gala apples. With pre-processed spectral data classified by PLS-DA, the results obtained in two sections demonstrate a significant classification rate which is over 96% in the 240-sample stage and 98% in the 60-sample stage. The results clearly demonstrate the potential for low cost field analysis in the food industry at different stages from preparation through distribution to near point of sale and can be applied to many food systems to protect against fraud, contamination or spoilage. This approach once integrated to a sensor form, can effectively detect organic food fraud if prior models are well established. Our future work will take into consideration of separate organic and non-organic apples of different species. This task may require a 'divide and conquer' strategy to pre-assign samples to sub-groups of apple species, and then classification decisions are made within each sub-group.

REFERENCES

[1] H. Willer and J. Lernoud, *The World of Organic Agriculture Statistics and Emerging Trends 2016*, 1st ed. Research Institute of Organic Agriculture (FiBL) and IFOAM – Organics International, 2016.

[2] "Share of organic food sales in the United States by product category, 2014 | Statistic", *Statista*, 2016. [Online]. Available: http://www.statista.com/statistics/244388/share-of-organic-food-sales-in-the-united-states-by-product/. [Accessed: 25- Apr- 2016].

[3] "Organic food product sales share in the UK 2014 | Statistic", *Statista*, 2016. [Online]. Available: http://www.statista.com/statistics/300129/organic-food-product-sales-share-in-the-united-kingdom-uk/. [Accessed: 25- Apr- 2016].

[4] "Big food corporations committing massive organic fraud - investigation", *NaturalNews*, 2016. [Online]. Available: http://www.naturalnews.com/048024_organic_industry_factory_farms_food_fraud.html. [Accessed: 25- Apr- 2016].

[5] H. Blake, "One sixth of 'fresh', 'organic' or 'handmade' food is fake", *Telegraph.co.uk*, 2011. [Online]. Available: http://www.telegraph.co.uk/foodanddrink/foodanddrinknews/8364731/One-sixth-of-fresh-organic-or-handmade-food-is-fake.html. [Accessed: 25- Apr- 2016].

[6] A. Vanzo, M. Jenko, U. Vrhovsek, and M. Stopar, "Metabolomic profiling and Sensorial quality of 'Golden Delicious', 'liberty', 'Santana', and 'topaz' apples grown using organic and integrated production systems," *Journal of Agricultural and Food Chemistry*, vol. 61, no. 26, pp. 6580–6587, Jul. 2013.

[7] "What's On My Food: Pesticides on Apples", *Whatsonmyfood.org*, 2016. [Online]. Available: http://www.whatsonmyfood.org/food.jsp?food=AP. [Accessed: 25- Apr- 2016].

[8] Y. Fan, K. Lai, B. A. Rasco, and Y. Huang, "Determination of carbaryl pesticide in Fuji apples using surface-enhanced Raman spectroscopy coupled with multivariate analysis," *LWT - Food Science and Technology*, vol. 60, no. 1, pp. 352–357, Jan. 2015.

[9] F. Ma and D. Dong, "A measurement method on pesticide residues of apple surface based on laser-induced breakdown spectroscopy," *Food Analytical Methods*, vol. 7, no. 9, pp. 1858–1865, Mar. 2014.

[10] G. Xiao et al., "Detection of pesticide (Chlorpyrifos) residues on fruit peels through spectra of Volatiles by FTIR," *Food Analytical Methods*, vol. 8, no. 5, pp. 1341–1346, Oct. 2014.

[11] W. Luo et al., "Preliminary study on the application of near infrared spectroscopy and pattern recognition methods to classify different types of apple samples," *Food Chemistry*, vol. 128, no. 2, pp. 555–561, Sep. 2011.

[12] X. Wu, B. Wu, J. Sun, and N. Yang, "Classification of apple varieties using near infrared reflectance spectroscopy and fuzzy Discriminant c-means clustering model," *Journal of Food Process Engineering*, p. n/a–n/a, Feb. 2016.

[13] N. Arias, S. Arazuri, and C. Jarén, "Ability of NIRS technology to determine pesticides in liquid samples at maximum residue levels," *Pest Management Science*, vol. 69, no. 4, pp. 471–477, Sep. 2012.

[14] M. Alamgir Zaman Chowdhury, A. N. M. Fakhruddin, M. Nazrul Islam, M. Moniruzzaman, S. H. Gan, and M. Khorshed Alam, "Detection of the residues of nineteen pesticides in fresh vegetable samples using gas chromatography–mass spectrometry," *Food Control*, vol. 34, no. 2, pp. 457–465, Dec. 2013.

[15] M. Hohmann, Y. Monakhova, S. Erich, N. Christoph, H. Wachter, and U. Holzgrabe, "Differentiation of organically and conventionally grown tomatoes by Chemometric analysis of combined data from proton nuclear magnetic resonance and mid-infrared spectroscopy and stable Isotope analysis," *Journal of Agricultural and Food Chemistry*, vol. 63, no. 43, pp. 9666–9675, Nov. 2015.

[16] A. Rinnan, F. van den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, Nov. 2009.

[17] "Correct baseline of signal with peaks - MATLAB msbackadj", *Uk.mathworks.com*, 2016. [Online]. Available: http://uk.mathworks.com/help/bioinfo/ref/msbackadj.html. [Accessed: 25- Apr- 2016].

[18] M. Barker and W. Rayens, "Partial least squares for discrimination," *Journal of Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.

[19] B. M. Nicolaï et al., "Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review," *Postharvest Biology and Technology*, vol. 46, no. 2, pp. 99–118, Nov. 2007.

[20] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, Aug. 1987.

[21] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression,"*Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251–263, Mar. 1993.

[22] "NIRQuest512 - Ocean Optics", *Ocean Optics*, 2016. [Online]. Available: http://oceanoptics.com/product/nirquest512/. [Accessed: 25- Apr- 2016].

[23] "OceanView 1.5.2 Now Available! - Ocean Optics", *Ocean Optics*, 2016. [Online]. Available: http://oceanoptics.com/product/oceanview/. [Accessed: 25- Apr- 2016].