Evaluation of machine learning methods for organic apple authentication based on diffraction grating and image processing

Weiran Song [a], Nanfeng Jiang [b, *], Hui Wang [a], Gongde Guo [b]

[a] *School of Computing, Ulster University, BT37 0QB, Newtownabbey, Co. Antrim, UK*

[b] *School of Mathematics and Informatics, Fujian Normal University, 350007, Fuzhou, China*

* Corresponding author.

E-mail address: jiangbbplayer@163.com (N. Jiang).

Abstract

Optical measuring technologies coupled with machine learning algorithms can be used to build a home-made sensor system. We built such a sensor system using a smartphone and a diffraction grating sheet. Diffraction images were captured under white light illumination and converted into a data matrix for data analysis. In this paper we present a systematic evaluation of this sensor system on the task of differentiating organic apples from conventional ones. We used the sensor system to measure 150 organic and conventional apples as rainbow images. We processed the rainbow images using computer vision techniques, built machine learning and chemometrics models, and used the resultant models to classify testing samples. Moreover, a comparative study was conducted where the same set of apples were scanned by a commercial spectrometer resulting in spectral data of the apple samples and classification was undertaken using partial least squares discriminant analysis (PLS-DA). Experimental results show that state of the art machine learning algorithms such as support vector machine (SVM) and locally weighted partial least squares classifier (LW-PLSC) are effective in handling low-quality image data with classification accuracies of 93−100%. These results suggest that the sensor system is convenient and low-cost, and provides a fast, effective, non-destructive and viable solution for in-line food authentication.

Keywords: Food authentication, Organic apple, Diffraction grating, Machine learning, Chemometrics

1. Introduction

Apple, which plays an important role in healthy diet, is one of the most cultivated and consumed fruits in the world. According to FAOSTAT, the total apple production reached 83.1 million tons worldwide in 2017 (http://www.fao.org/faostat). Meanwhile, apple quality is gaining increasing attention due to the rising concerns about food safety and quality. Some issues related to the external and internal quality of apple, such as bruise degree, diseases, pesticide residue contamination and organic fraud, pose a serious threat to consumer health and damage the fair trade-off between quality and price. Traditional testing techniques based on sensory and chemical analysis are laborious and time-consuming, so they do not meet the growing demand for large-scale and real-time apple quality testing in industrial process and consumer market.

Optical sensors coupled with chemometrics have become an effective approach to predict the quantitative and qualitative attributes of apple (Ignat et al., 2014; Moscetti et al., 2018). They are also effective for rapid and non-invasive food evaluation which requires minimal sample preparations. For quantitative research, many studies investigate the internal contents of apples by using near-infrared (NIR) spectroscopy and NIR hyperspectral imaging techniques. These studies predict the internal contents, including sweetness (soluble solids content, SSC) (Ma et al., 2018; Tang et al., 2018; Yuan et al., 2016), sourness (acidity or pH value) (Ignat et al., 2014; Jha and Ruchi, 2010), firmness (Ignat et al., 2014) and moisture (Dong and Guo, 2015), which directly influence flavours and textures of apples. Other studies are related to food safety and health issues, i.e., the pesticide contamination of apples. It has been reported that surface-enhanced Raman spectroscopy (SERS), Fourier transform infrared (FTIR) spectroscopy and laser-induced breakdown spectroscopy (LIBS) techniques can effectively measure the level of pesticide residuals on apple surface, such as chlorpyrifos (Dhakal et al., 2014; Ma and Dong, 2014; Xiao et al., 2015), carbaryl (Fan et al., 2015), phosmet and thiabendazole (Luo et al., 2016). Spectroscopy and hyperspectral imaging are also two state of the art techniques for determining apple qualities, for example, identifying varieties, grades, geographical origins (Luo et al., 2011), detecting apple diseases (Jarolmasjed et al., 2017) and bruising degree (Tan et al., 2018; Vetrekar et al., 2015). Furthermore, computer vision system (CVS) is recently used for the evaluation of ripening stages (Cárdenas-Pérez et al., 2017), surface gloss (Sun et al., 2017), diseases (Dubey and Jalal, 2016) and defectiveness (Zhang et al., 2015).

The use of portable spectrometer for real-time apple quality assessment is in a strong uptrend which has met the requirement for practical use (Gao et al., 2016; Yuan et al., 2016). Our previous studies have demonstrated that the use of portable NIR spectrometer coupled with chemometrics provides a feasible approach for authenticating organic apples with an accuracy of over 90% (Song et al., 2018a, 2016). This approach is simple, quick and non-destructive compared to conventional analytical techniques for authenticating organic foods such as compound-specific isotope analysis (CSIA), inductively coupled plasma-mass spectrometry (ICP-MS), isotope ratio mass spectrometry (IRMS) (de Lima and Barbosa, 2019). However, the miniaturisation and field portability of spectrometers will normally degrade the fingerprint data quality due to the variable sampling conditions, posing challenges to linear chemometric algorithms (Liu et al., 2018; Song et al., 2018a). One of the most standard chemometric method for data classification is partial least squares discriminant analysis (PLS-DA), which effectively handles high dimensionality, high collinearity and small sample size problems. However, it sometimes yields unsatisfied performance due to the high degree of nonlinearity (Song et al., 2018b; Zou et al., 2010). To tackle this issue, many machine learning algorithms have been investigated and become an indispensable part in chemometrics field. For example, nonlinear classifiers such as support vector machine (SVM) and random forest (RF) are often well-performing in classifying spectral data due to the good generalization performance (Devos et al., 2009; Zhang et al., 2016). Artificial neural networks (ANN) and extreme learning machine (ELM) can also efficiently capture the nonlinear relationship between observations and classes, gaining an advantage over PLS-DA in classification accuracy and robustness (Moncayo et al., 2015; Zheng et al., 2014).

Despite the fact that nonlinear algorithms can improve the classification capability of low-quality spectral data to some extent, the price of portable spectrometer far exceeds expectation in consumer market. Recent studies attempt to use CVS for food quality evaluation based on mobile devices such as smartphone and tablet (Cruz-Fernández et al., 2017; Cubero et al., 2018), which demonstrate potential in on-line application. CVS simulates human visual system using artificial sensor and automatically gains high-level understanding from digital images via image acquisition, processing and data analysis. Our recent study proposes a low-cost CVS based on diffraction grating and demonstrates its feasibility in organic apple identification (Jiang et al., 2018). This study uses locally weighted partial least squares classifier (LW-PLSC) to handle nonlinear

food data which significantly improves the classification performance compared to PLS-DA. Nevertheless, an up-to-date comparative study is essential for three reasons: first, LW-PLSC requires to be evaluated by comparing its capability with baseline classifiers; second, the empirical reference in selecting the most appropriate classifiers for the new type of image data has yet been studied; third, the performance comparison between the new sensor system and the other optical sensors such as spectrometer remains to be investigated.

Organic apples are generally more expensive than conventional ones. This, coupled with the fact that visual differentiation between organic and conventional apples is often not possible, has led to organic food fraud involving apples. In this study, we aim to distinguish organic and conventional labelled apples using a combination of low-cost sensor system and supervised machine learning methods. We evaluate the classification performance of ten methods on rainbow image data, including PLS-DA, kernel PLS-DA (KPLS-DA), LW-PLSC, soft independent modelling of class analogies (SIMCA), *k*-nearest neighbours (*k*-NN), logistic regression (LR), SVM, least squares SVM (LS-SVM), decision tree (C4.5) and RF, and choose the best-performing ones under different sample distributions. Then a benchmark instrument, a commercial high-resolution spectrometer, is used to measure the same samples and compared with the sensor system in identifying organic and conventional apples. This study reveals that the low-quality image data obtained from low-cost measurement is effective for apple fruit authentication with the aid of high performance machine learning methods.

2. Materials and methods

2.1. Sample preparation

A total of 150 apples were collected from local super markets in Fuzhou during a week. There were three apple varieties (50 Braeburn, 50 Gala and 50 Pink lady, respectively) and each variety contained two classes (25 organic and 25 conventional, respectively). All apples were defect-free, similar in size and maturity, and no surface preparation was carried out prior to data collection. We conducted the imaging and spectral measurements at room temperature (22 ± 2 °C).

2.2. Imaging system based on diffraction grating

A recently proposed sensor system is used to obtain image data from apple samples, which includes diffraction image acquisition, rainbow image extraction and feature vector representation (Jiang et al., 2018).

2.2.1. Diffraction image acquisition

An apple sample was placed 20 cm in front of a flashlight, two diffraction grating sheet (60 × 40 nm) was set on both sides of the apple and a smartphone camera was fixed by 1 cm above the flashlight. We use the flashlight to illuminate the apple surface, so reflected polychromatic light can pass through a diffraction grating and then disperse into several beams travelling in different directions. Each beam has a single rainbow of colours under white light illumination. Then the rainbows are photographed by smartphone with 1080 × 720 pixels spatial resolution and stored in JPG format with file size of approximately 1.14MB. To eliminate the influence of ambient light and generate comparably high-quality data, the image acquisition was conducted in a dark environment. Fig. 1a shows a central part of the diffraction image with rainbow colour spectra.

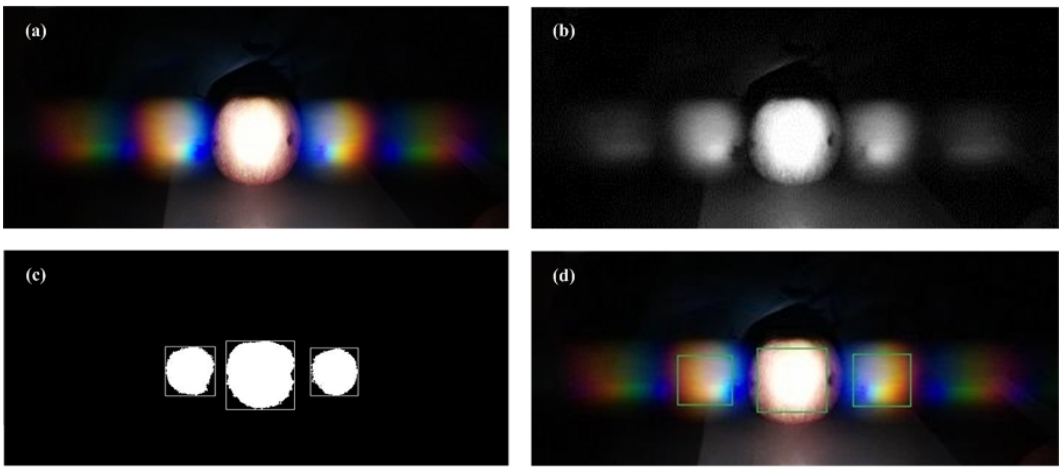


Fig. 1. (a) The original diffraction image of rainbow colour spectra; (b) grayscale processed image; (c) mathematical morphology processed image; (d) the extracted rainbows in image.

2.2.2. Rainbow image extraction

We use a combination of image processing techniques to extract a single rainbow from the obtained diffraction image, including pre-processing, denoising and segmentation. Grayscale processing firstly coverts colour pixels into grey ones (see Fig. 1b) which only carry intensity information. This step enables mathematical morphology to capture the most essential shape features of target rainbows, as shown in Fig. 1c. Then denoising adopts median filter to replace the value of a point in the digital image with the median of neighbouring points. Finally, the OSTU method (Otsu, 1979) calculates the foreground and background class probability, so a single rainbow image can be derived from the raw diffraction image (see Fig. 1d). A resized rainbow image (50 × 100 pixels) to be converted into numerical values is shown Fig. 2.

Fig. 2. The resized rainbow image (50 × 100 pixels) of an apple sample.

2.2.3. Feature vector representation

We map the rainbow image into 3-dimentional Cartesian coordinate system of red (R), green (G) and blue (B) colour. Each colour channel has 256 colour levels varying from 0 to 255. It is noted that the row pixels usually contain more hues than the column pixels, so we calculate the mean of each column and use the obtained feature vector to represent the spectral line of sample. The raw image data is shown in Fig. 3.

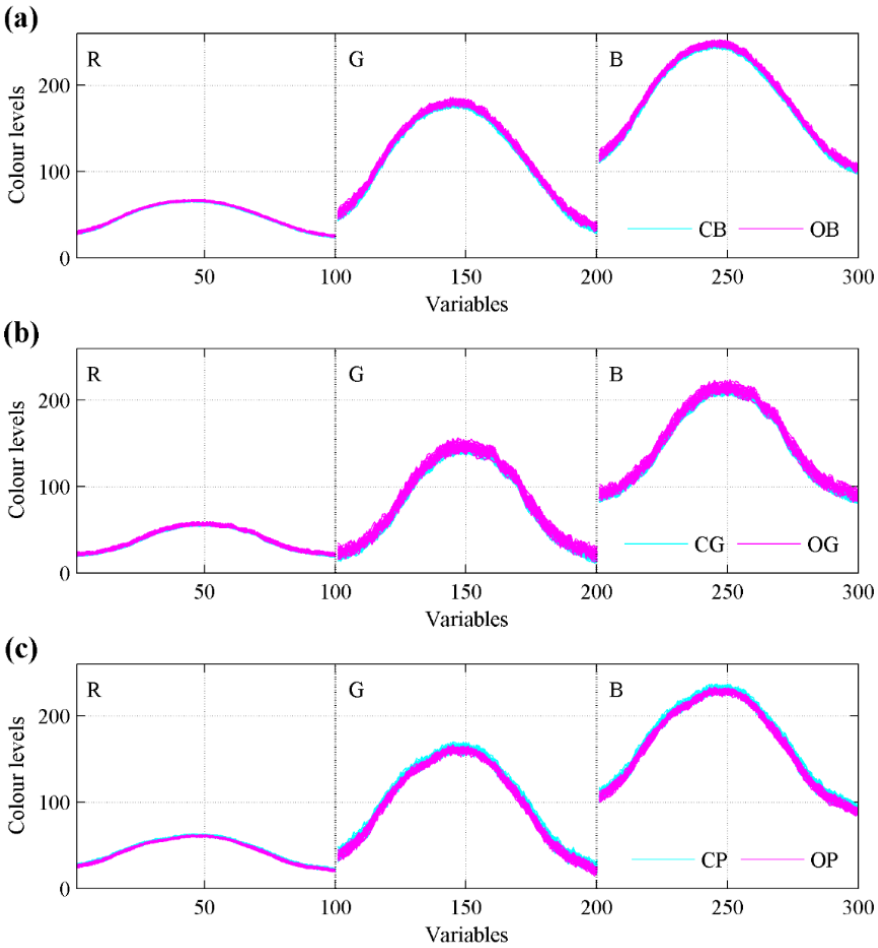


Fig. 3. The raw image data of Braeburn (B), Gala (G) and Pink lady (P) apples. Data in cyan and magenta colour represents conventional (C) and organic (O) samples, respectively. Variables 1-100, 101-200 and 201-300 belongs to red, green and blue channels, respectively.

2.3. Spectroscopy

Apples spectral reflectance data were collected with a high-resolution spectrometer (USB4000-FL spectrometer, Ocean Optics, Inc., USA) equipped with an optical fiber probe and having a wavelength range of 200-1100 nm with an interval of 0.9879 nm. The experiments were conducted under ambient light conditions in a colour assessment cabinet which can provide a standard visible light source. Apple spectra of three varieties were collected with the Ocean-View software containing 912 variables, as shown in Fig. 4.

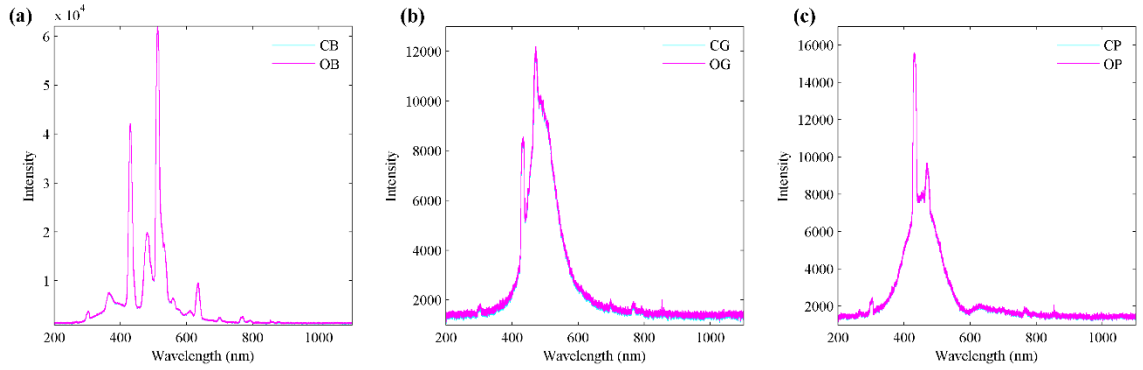


Fig. 4. Conventional and organic apple spectra (148 samples and 912 variables) of three varieties (Braeburn, Gala and Pink lady). Two outliners in Gala and Pink lady varieties have been removed for better visualization.

2.4. Data analysis

2.4.1. Sample division and pre-processing

As the investigated apples contain six groups, namely, conventional Braeburn (CB), organic Braeburn (OB), conventional Gala (CG), organic Gala (OG), conventional Pink lady (CP) and organic Pink lady (OP), we use DUPLEX algorithm (Snee, 1977) to partition apples belonging to the same group into training and testing samples according to the ratio of 2:1. To explore the differentiation between conventional and organic apples within each variety, training samples belongs to CB and OB groups are merged as a whole training set for Braeburn variety meanwhile the corresponding testing samples are combined as a testing set. The same procedure is also applied on Gala and Pink lady varieties. We also attempt to build models based on the overall varieties and classify testing samples of different varieties. Therefore, we integrate the three training sets and mix the corresponding testing sets. Such way of sample division maintains the same diversity in both sets and keeps the balance between two classes.

We only apply Savitzky-Golay smoothing (fitted by a polynomial of degree two and a 33-point moving window) to pre-process raw image data, because our previous study reports such pre-processing can effectively reduce the noise and improves the modelling performance (Jiang et al., 2018). For data obtained from spectroscopy, we use raw spectra without pre-processing due to its high quality and the high performance of PLS-DA. This will be demonstrated in Section 3.

2.4.2. Classification methods

This work implements ten commonly used algorithms in chemometrics and machine learning fields to classify apple image data. PLS-DA relies on the assumption that the investigated system or process is driven by a set of latent variables (LVs) in low dimensional space. It transforms the categorical vector into numerical responses and searches for latent variables with maximum covariance with the responses (Barker and Rayens, 2003). KPLS-DA maps the original data into Hilbert feature space via kernel function and then constructs a PLS-DA model for classification. The nonlinear relationship among variables in the original sample space becomes linear after mapping, so data nonlinearity can be effectively captured. LW-PLSC is an extension of LW-PLS (Kim et al., 2011), which uses weighting schemes for queries which respectively enlarges and lessens the influence of neighbouring and remote samples towards a PLS-DA model. Thus, the degree of global nonlinearity is reduced by using local linear models. SIMCA performs PCA on each class and constructs principal component models with the optimal numbers of PCs identified by cross-validation. A query is then attributed to the class which yields the least residue during prediction.

The *k*-NN predicts a query according to the *k* closest samples of the query and assign it to the class which has the largest category probability. The Euclidean distance is the most commonly used distance function in *k*-NN, however, it can barely provide sufficient distinctions between different samples in high-dimensional case (Aggarwal et al., 2001). The LR is a widely used statistical model which aims to solve binary classification problems. It extends ordinary least squares to model the logistic relationship between the probability of class membership and the input variables. The SVM classification searches for a hyperplane to correctly separate samples of different classes meanwhile maximizing the shortest distances from the hyperplane to the nearest samples for each class. It can be extended to non-linear classification by projecting data from low dimensional input space to high dimensional feature space via kernel functions. The LS-SVM is an advanced version of SVM for binary classification which applies the linear least squares criteria to the loss function instead of inequality constraints (Suykens and Vandewalle, 1999). Decision tree uses a flowchart-like structure to present the various outcomes from a series of decisions. It mainly consists of a root node, branches and leaf nodes. The root node represents a query to be assigned, the branch represents the flow from question to answer and the leaf node represents a class label. The RF is an ensemble method

191 which generates multiple decision trees and predict a query based on a simple majority voting of the single
192 classification tree.

193 In our experiments, the decision tree classifier (C4.5) was from the WEKA learning environment using
194 J48 function while other classifiers were from MATLAB (Mathworks, 2011a). Some baseline classifiers
195 were from MATLAB external toolboxes, including LIBSVM toolbox (SVM) (Chang and Lin, 2011),
196 Classification toolbox (SIMCA) (Ballabio and Consonni, 2013) and LS-SVM toolbox (LS-SVM) (De
197 Brabanter et al., 2011).

198 *2.4.3. Parameter setting*

199 We use leave-one-out cross validation on training set to optimize the parameters of different algorithms.
200 The range of LVs in PLS-DA, KPLS-DA and LW-PLSC is set from 1 and 10 to prevent overfitting. The
201 number of components for each class model in SIMCA is no more than 10. The number of nearest neighbours
202 in *k*-NN is selected from 1 to 15. The regularization parameter λ in LR is varied from 10^{-7} to 10^{3} on a
203 logarithmic scale. This work adopts radial basis function for three kernel methods and implements a grid
204 search approach for kernel methods and LW-PLSC. The width of the RBF σ in KPLS-DA are 1, 5, 10, 50,
205 100, 500 and 1000. The regularization parameter C (C = 1, 10, 100, 1000) and RBF kernel parameter γ (γ =
206 10^{-5}, 5×10^{-5}, 10^{-4}, 5×10^{-4}, 10^{-3}, 5×10^{-3}, 10^{-2}) totally construct 28 SVM models, while the parameters of LS-
207 SVM build 20 models (C = 1, 10, 100, 1000 and σ = 1, 10, 100, 1000, 10000). The localization parameter φ
208 in LW-PLSC is adjusted to the values of 0.1, 0.5, 1, 5, 10, 15 and 20. The depth of tree in C4.5 are 2, 4, 6, 8,
209 10, 12 and 14. We use the default setting of RF parameters (*ntree* = 500, *mtry* = $\sqrt{p}$ and *nodesize* = 1, where
210 p is the number of variables) for validation and classification, which has been reported well-performing in
211 most cases (Strobl et al., 2009).

212 **3. Results and discussion**

213 *3.1. PCA of image and spectral data*

214 The data of each apple variety plotted according to the PCA scores is shown in Fig. 5. Training and testing
215 data are represented as empty and filled points, respectively. The first two principal components (PCs)

accumulate over 93% of the total variance in each variety. Samples of Braeburn variety have a good separation between organic and conventional classes, while samples of Gala or Pink lady varieties present an overlap between two classes. Reasonable classification models may exist in Gala and Pink lady varieties despite the visualized two dimensions does not display a clear separation.

The PCA charts of the overall samples obtained from imaging and spectroscopy are shown in Fig. 6a and b, respectively. The distinction of samples based on apple varieties is quite clear compared to that of samples based on classes. Data obtained from low-cost imaging technique fails to provide a linear separation between organic and conventional classes by using two PCs, while data collected by high-resolution spectroscopy presents a good class separation. Typically, spectral data of individual variety are linearly separable based on organic and conventional classes. Two outliers from CG and CP groups can be identified from the PCA scatter plot of spectral data, which may be induced by mislabelling or measuring distortion.

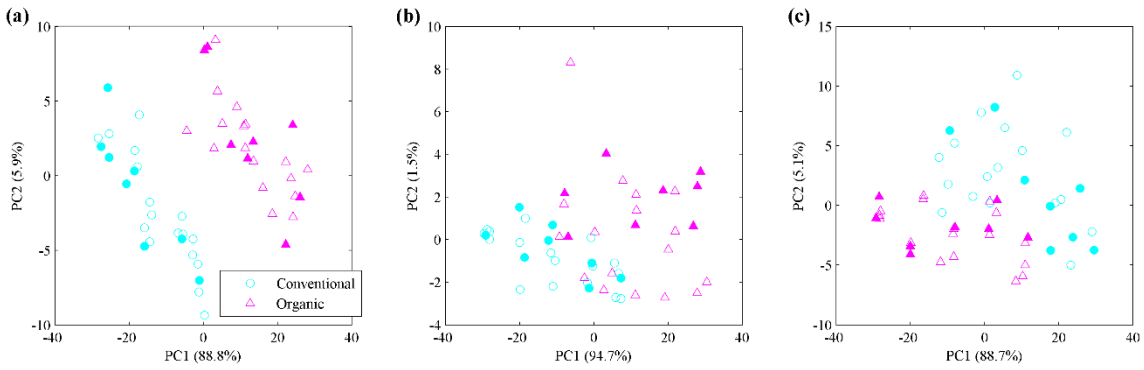


Fig. 5. PCA scatter plot of rainbow image data: Braeburn (a), Gala (b) and Pink lady (c) apples. Training and testing data are represented as empty and filled points, respectively.

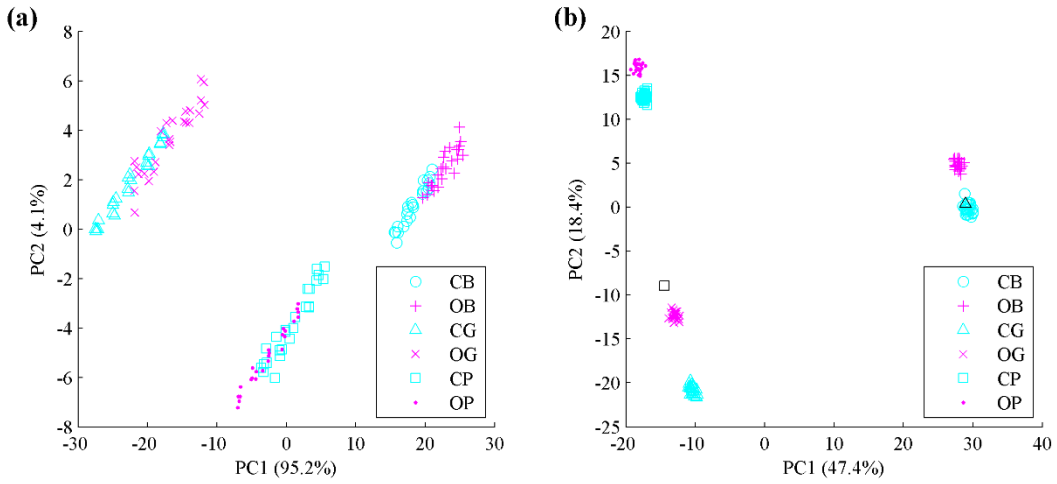


Fig. 6. PCA scatter plots of the apple data from imaging (a) and spectroscopy (b). CB: conventional Braeburn; OB: organic Braeburn; CG: conventional Gala; OG: organic Gala; CP: conventional Pink lady; OP: organic Pink lady. Two outliers from CG and CP group are marked as black in spectral data.

3.2. Classification of image and spectral data

The classification of apple image datasets, including each variety and the overall varieties, are presented in Table 1 and

Table 2. For Braeburn variety, despite several algorithms (LR, *k*-NN, C4.5 and RF) misclassify one sample in training phase, all algorithms can successfully identify testing samples. For Gala variety, both KPLS-DA and SVM reach the highest results of 94.1% and 100%, respectively in training and testing phase. Moreover, LW-PLSC and LS-SVM have the same results of 100% in classification. While other classifiers yield lower validation and classification results ranging from 76.5% (SIMCA and *k*-NN) to 91.2% (LS-SVM) and 75% (SIMCA) to 93.8% (PLS-DA, LR and *k*-NN), respectively. The number of wrongly classified organic apples does not exceed 1 for most of the algorithms. For Pink lady variety, the highest validation accuracy is 91.2% obtained by KPLS-DA, while that of classification is 100% achieved by both KPLS-DA and *k*-NN. Other classifiers yield lower validation and classification results which are respectively below 90% and 95%. The conventional samples are easily identified by all algorithms, while the organic ones can only be correctly recognised by KPLS-DA and *k*-NN. The highest accuracy of the overall dataset is lower than that of each dataset. LW-PLSC and RF models attain the highest validation results of 90.2% while LW-PLSC and *k*-NN give the best performance in classification with 97.9% accuracy. Classifiers such as KPLS-

DA, SVM and LS-SVM provide comparable results in both phases. The highest accuracy of organic class (95.8%) is obtained by LW-PLSC, SIMCA and *k*-NN, of which 23 out of 24 samples are correctly identified.

Table 1

The accuracy (%) of different algorithms for the classification of organic and conventional apples (Braeburn and Gala varieties) based on data obtained from diffraction images.

Braeburn	Training	Testing	Organic	Conventional
PLS-DA	**100**	**100**	100	100
KPLS-DA	**100**	**100**	100	100
LW-PLSC	**100**	**100**	100	100
SVM	**100**	**100**	100	100
LS-SVM	**100**	**100**	100	100
SIMCA	**100**	**100**	100	100
LR	97.1	**100**	100	100
k-NN	97.1	**100**	100	100
C4.5	97.1	**100**	100	100
RF	97.1	**100**	100	100

Gala	Training	Testing	Organic	Conventional
PLS-DA	85.3	93.8	87.5	100
KPLS-DA	**94.1**	**100**	100	100
LW-PLSC	88.2	**100**	100	100
SVM	**94.1**	**100**	100	100
LS-SVM	91.2	**100**	100	100
SIMCA	76.5	75	100	50
LR	82.4	93.8	87.5	100
k-NN	76.5	93.8	87.5	100
C4.5	85.3	87.5	87.5	87.5
RF	85.3	81.3	75	87.5

Table 2

The accuracy (%) of different algorithms for the classification of organic and conventional apples (Pink lady and the overall varieties) based on data obtained from diffraction images.

Pink lady	Training	Testing	Organic	Conventional
PLS-DA	85.3	93.8	87.5	100
KPLS-DA	**91.2**	**100**	100	100
LW-PLSC	88.2	93.8	87.5	100
SVM	88.2	93.8	87.5	100
LS-SVM	88.2	93.8	87.5	100
SIMCA	85.3	87.5	75	100
LR	85.3	93.8	87.5	100
k-NN	85.3	**100**	100	100
C4.5	76.5	93.8	87.5	100
RF	82.4	93.8	87.5	100

Overall	Training	Testing	Organic	Conventional
PLS-DA	55.9	58.3	58.3	58.3
KPLS-DA	89.2	95.8	91.7	100
LW-PLSC	**90.2**	**97.9**	95.8	100
SVM	88.2	95.8	91.7	100
LS-SVM	89.2	95.8	91.7	100
SIMCA	70.6	79.2	95.8	62.5
LR	58.8	60.4	54.2	66.7
k-NN	86.3	**97.9**	95.8	100
C4.5	84.3	87.5	75	100
RF	**90.2**	93.8	91.7	95.8

Among the ten algorithms, KPLS-DA achieves the top validation and classification results on the first three datasets. SVM based algorithms also present good classification performance on four datasets. Nevertheless, kernel method is more prone to overfitting than its non-kernel counterpart if data has a limited number of samples (Despagne et al., 2000). By adjusting the contribution of training samples in a local model for a query, LW-PLS classification improves the performance of PLS-DA on the classification of nonlinear data. Linear classifiers i.e., PLS-DA and LR, provide acceptable results for differentiating organic apples

from conventional ones within each variety. However, such results will drastically degrade by over 30% when classifying testing samples from the overall varieties. The *k*-NN provides the highest classification accuracies in three datasets by selecting one nearest neighbour. However, the validation results of *k*-NN are usually lower than that of kernel algorithms.

We also provide the validation and classification results of PLS-DA on apple spectral datasets, as in **Error! Not a valid bookmark self-reference.**. PLS-DA model can effectively identify organic samples in Braeburn, Gala and Pink lady varieties with validation results of 100%, 97.1% and 97.1%, respectively. Two outliers from CG and CP groups are the only misclassified samples. However, the outlier from CG group can be correctly attributed to the conventional class as it is close to CB samples in PCA scatter plot (see Fig. 6b), yielding an overall accuracy of 99% in validation. PLS-DA selects additional numbers of LVs across the overall dataset, showing an increased degree of nonlinearity. Nevertheless, the optimal number of LVs identified by leave-one-out cross validation does not exceed 3 for each dataset. The corresponding PLS-DA model has low simplicity but still correctly distinguish organic apples from conventional ones on the four datasets due to the high quality of spectral data.

Table 3

The accuracy (%) of PLS-DA for the classification of organic and conventional apples (Braeburn, Gala, Pink lady and the overall varieties) based on data obtained from spectroscopy.

Datasets	Training	LVs	Testing	Organic	Conventional
Braeburn	100	1	100	100	100
Gala	97.1	2	100	100	100
Pink lady	97.1	2	100	100	100
Overall	99	3	100	100	100

By comparing the above results, many classifiers on image data achieves the same level of accuracies compared to PLS-DA on spectral data when classifying apples from Braeburn and Gala variety. If we merge the apples of different varieties, the best classification result of image data will be lower than that of spectral data by 6.7%. Such degradation in performance indicates that the image data has lower quality compared to the spectral one. However, the sensor system is still feasible for organic apple authentication ($\geq$ 90% accuracy) with the aid of state of the art machine learning methods.

4. Conclusion

A prototype sensor system and ten classification methods were evaluated as a solution for fast and non-destructive detection of apple quality, more specifically, to determine if an apple is organic or conventional. The rainbow image data obtained from the sensor system was lower in resolution and higher in degree of nonlinearity compared to the spectral data generated by a commercial spectrometer. It was found that the classification results of image data were comparable to that of spectral data when equipped with the of state of the art classifiers, such as SVM and LW-PLSC. Such results demonstrate the effectiveness and significance of the sensor system for differentiating organic apples from conventional ones based on the colour level. Moreover, the sensor system has extremely lower price in hardware compared to commercial spectrometer, which is practically suitable for low-cost food quality detection. However, due to the instrumental restrictions (size of diffraction grating sheet, dispersion of flashlight and resolution of smartphone camera), the food produce used for experiments currently requires having a proper size and shape to ensure that a complete rainbow image is clearly presented and effectively captured. Our future work will optimize the experimental settings and improve the detection performance by selecting variables of class distinction.

Acknowledgments

This work was supported by Fujian science and technology department project (No. JK2017007), Natural Science Foundation of Fujian Province, China (No. 2018J01776), Natural Science Foundation of Fujian Province, China (No. 2018J01775) and the National Natural Science Foundation of China under Grant (No. 61672157).

References

Aggarwal, C.C., Hinneburg, A., Keim, D.A., 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Space BT - Database theory, in: Database Theory. pp. 420–434.

Ballabio, D., Consonni, V., 2013. Classification tools in chemistry. Part 1: Linear models. PLS-DA. Anal. Methods 5, 3790–3798. https://doi.org/10.1039/c3ay40582f

Barker, M., Rayens, W., 2003. Partial least squares for discrimination. J. Chemom. 17, 166–173. https://doi.org/10.1002/cem.785

313 Cárdenas-Pérez, S., Chanona-Pérez, J., Méndez-Méndez, J. V., Calderón-Domínguez, G., López-Santiago, R., Perea-Flores, M.J.,
314 Arzate-Vázquez, I., 2017. Evaluation of the ripening stages of apple (Golden Delicious) by means of computer vision system.
315 Biosyst. Eng. 159, 46–58. https://doi.org/10.1016/j.biosystemseng.2017.04.009

316 Chang, C.C., Lin, C.J., 2011. LIBSVM: A Library for support vector machines. ACM Trans. Intell. Syst. Technol.
317 https://doi.org/10.1145/1961189.1961199

318 Cruz-Fernández, M., Luque-Cobija, M.J., Cervera, M.L., Morales-Rubio, A., de la Guardia, M., 2017. Smartphone determination of fat
319 in cured meat products. Microchem. J. https://doi.org/10.1016/j.microc.2016.12.020

320 Cubero, S., Albert, F., Prats-Moltalbán, J.M., Fernández-Pacheco, D.G., Blasco, J., Aleixos, N., 2018. Application for the estimation of
321 the standard citrus colour index (CCI) using image processing in mobile devices. Biosyst. Eng. 167, 63–74.
322 https://doi.org/10.1016/j.biosystemseng.2017.12.012

323 De Brabanter, K., Karsmakers, P., Ojeda, F., Alzate, C., De Brabanter, J., Pelckmans, K., De Moor, B., Vandewalle, J., Suykens, J.,
324 2011. LS-SVMlab Toolbox User ' s Guide Version 1.8. ESAT-SISTA Tech. Rep.

325 de Lima, M.D., Barbosa, R., 2019. Methods of Authentication of Food Grown in Organic and Conventional Systems Using
326 Chemometrics and Data Mining Algorithms: a Review. Food Anal. Methods 887–901. https://doi.org/10.1007/s12161-018-
327 01413-3

328 Despagne, F., Luc Massart, D., Chabot, P., 2000. Development of a robust calibration model for nonlinear in-line process data. Anal.
329 Chem. 72, 1657–1665. https://doi.org/10.1021/ac991076k

330 Devos, O., Ruckebusch, C., Durand, A., Duponchel, L., Huvenne, J.P., 2009. Support vector machines (SVM) in near infrared (NIR)
331 spectroscopy: Focus on parameters optimization and model interpretation. Chemom. Intell. Lab. Syst. 96, 27–33.
332 https://doi.org/10.1016/j.chemolab.2008.11.005

333 Dhakal, S., Li, Y., Peng, Y., Chao, K., Qin, J., Guo, L., 2014. Prototype instrument development for non-destructive detection of
334 pesticide residue in apple surface using Raman technology. J. Food Eng. 123, 94–103.
335 https://doi.org/10.1016/j.jfoodeng.2013.09.025

336 Dong, J., Guo, W., 2015. Nondestructive Determination of Apple Internal Qualities Using Near-Infrared Hyperspectral Reflectance
337 Imaging. Food Anal. Methods 8, 2635–2646. https://doi.org/10.1007/s12161-015-0169-8

338 Dubey, S.R., Jalal, A.S., 2016. Apple disease classification using color, texture and shape features from images. Signal, Image Video
339 Process. 10, 819–826. https://doi.org/10.1007/s11760-015-0821-1

340 Fan, Y., Lai, K., Rasco, B.A., Huang, Y., 2015. Determination of carbaryl pesticide in Fuji apples using surface-enhanced Raman
341 spectroscopy coupled with multivariate analysis. LWT - Food Sci. Technol. 60, 352–357.
342 https://doi.org/10.1016/j.lwt.2014.08.011

343 Gao, F., Dong, Y., Xiao, W., Yin, B., Yan, C., He, S., 2016. LED-induced fluorescence spectroscopy technique for apple freshness and

quality detection. Postharvest Biol. Technol. 119, 27–32. https://doi.org/10.1016/j.postharvbio.2016.04.020

Ignat, T., Lurie, S., Nyasordzi, J., Ostrovsky, V., Egozi, H., Hoffman, A., Friedman, H., Weksler, A., Schmilovitch, Z., 2014. Forecast of Apple Internal Quality Indices at Harvest and During Storage by VIS-NIR Spectroscopy. Food Bioprocess Technol. 7, 2951–2961. https://doi.org/10.1007/s11947-014-1297-7

Jarolmasjed, S., Zúñiga Espinoza, C., Sankaran, S., 2017. Near infrared spectroscopy to predict bitter pit development in different varieties of apples. J. Food Meas. Charact. 11, 987–993. https://doi.org/10.1007/s11694-017-9473-x

Jha, S.N., Ruchi, G., 2010. Non-destructive prediction of quality of intact apple using near infrared spectroscopy. J. Food Sci. Technol. 47, 207–213. https://doi.org/10.1007/s13197-010-0033-1

Jiang, N., Song, W., Wang, H., Guo, G., Liu, Y., 2018. Differentiation between organic and non-organic apples using diffraction grating and image processing—A cost-effective approach. Sensors (Switzerland) 18, 1667. https://doi.org/10.3390/s18061667

Kim, S., Kano, M., Nakagawa, H., Hasebe, S., 2011. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. Int. J. Pharm. 421, 269–274. https://doi.org/10.1016/j.ijpharm.2011.10.007

Liu, N., Parra, H.A., Pustjens, A., Hettinga, K., Mongondry, P., van Ruth, S.M., 2018. Evaluation of portable near-infrared spectroscopy for organic milk authentication. Talanta 184, 128–135. https://doi.org/10.1016/j.talanta.2018.02.097

Luo, H., Huang, Y., Lai, K., Rasco, B.A., Fan, Y., 2016. Surface-enhanced Raman spectroscopy coupled with gold nanoparticles for rapid detection of phosmet and thiabendazole residues in apples. Food Control 68, 229–235. https://doi.org/10.1016/j.foodcont.2016.04.003

Luo, W., Huan, S., Fu, H., Wen, G., Cheng, H., Zhou, J., Wu, H., Shen, G., Yu, R., 2011. Preliminary study on the application of near infrared spectroscopy and pattern recognition methods to classify different types of apple samples. Food Chem. 128, 555–561. https://doi.org/10.1016/j.foodchem.2011.03.065

Ma, F., Dong, D., 2014. A Measurement Method on Pesticide Residues of Apple Surface Based on Laser-Induced Breakdown Spectroscopy. Food Anal. Methods 7, 1858–1865. https://doi.org/10.1007/s12161-014-9828-4

Ma, T., Li, X., Inagaki, T., Yang, H., Tsuchikawa, S., 2018. Noncontact evaluation of soluble solids content in apples by near-infrared hyperspectral imaging. J. Food Eng. 224, 53–61. https://doi.org/10.1016/j.jfoodeng.2017.12.028

Moncayo, S., Manzoor, S., Navarro-Villoslada, F., Caceres, J.O., 2015. Evaluation of supervised chemometric methods for sample classification by Laser Induced Breakdown Spectroscopy. Chemom. Intell. Lab. Syst. 146, 354–364. https://doi.org/10.1016/j.chemolab.2015.06.004

Moscetti, R., Raponi, F., Ferri, S., Colantoni, A., Monarca, D., Massantini, R., 2018. Real-time monitoring of organic apple (var. Gala) during hot-air drying using near-infrared spectroscopy. J. Food Eng. 222, 139–150. https://doi.org/10.1016/j.jfoodeng.2017.11.023

Otsu, N., 1979. A Threshold Selection Method from Gray-Level Histograms. IEEE Trans. Syst. Man. Cybern. 9, 62–66. https://doi.org/10.1109/TSMC.1979.4310076

Snee, R.D., 1977. Validation of Regression Models: Methods and Examples. Technometrics 19, 415–428. https://doi.org/10.1080/00401706.1977.10489581

Song, W., Wang, H., Maguire, P., Nibouche, O., 2018a. Nearest clusters based partial least squares discriminant analysis for the classification of spectral data. Anal. Chim. Acta 1009, 27–38. https://doi.org/10.1016/j.aca.2018.01.023

Song, W., Wang, H., Maguire, P., Nibouche, O., 2018b. Collaborative representation based classifier with partial least squares regression for the classification of spectral data. Chemom. Intell. Lab. Syst. 182, 79–86. https://doi.org/10.1016/j.chemolab.2018.08.011

Song, W., Wang, H., Maguire, P., Nibouche, O., 2016. Differentiation of organic and non-organic apples using near infrared reflectance spectroscopy — A pattern recognition approach. 2016 IEEE Sensors 1–3. https://doi.org/10.1109/ICSENS.2016.7808530

Strobl, C., Malley, J., Tutz, G., 2009. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. Psychol. Methods 14, 323–348. https://doi.org/10.1037/a0016973

Sun, K., Li, Y., Peng, J., Tu, K., Pan, L., 2017. Surface Gloss Evaluation of Apples Based on Computer Vision and Support Vector Machine Method. Food Anal. Methods 10, 2800–2806. https://doi.org/10.1007/s12161-017-0849-7

Suykens, J.A.K., Vandewalle, J., 1999. Least Squares Support Vector Machine Classifiers. Neural Process. Lett. 9, 293–300. https://doi.org/10.1023/A:1018628609742

Tan, W., Sun, L., Yang, F., Che, W., Ye, D., Zhang, D., Zou, B., 2018. Study on bruising degree classification of apples using hyperspectral imaging and GS-SVM. Optik (Stuttg). 154, 581–592. https://doi.org/10.1016/j.ijleo.2017.10.090

Tang, C., He, H., Li, E., Li, H., 2018. Multispectral imaging for predicting sugar content of 'Fuji' apples. Opt. Laser Technol. 106, 280–285. https://doi.org/10.1016/j.optlastec.2018.04.017

Vetrekar, N.T., Gad, R.S., Fernandes, I., Parab, J.S., Desai, A.R., Pawar, J.D., Naik, G.M., Umapathy, S., 2015. Non-invasive hyperspectral imaging approach for fruit quality control application and classification: case study of apple, chikoo, guava fruits. J. Food Sci. Technol. 52, 6978–6989. https://doi.org/10.1007/s13197-015-1838-8

Xiao, G., Dong, D., Liao, T., Li, Y., Zheng, L., Zhang, D., Zhao, C., 2015. Detection of Pesticide (Chlorpyrifos) Residues on Fruit Peels Through Spectra of Volatiles by FTIR. Food Anal. Methods 8, 1341–1346. https://doi.org/10.1007/s12161-014-0015-4

Yuan, L. ming, Cai, J. rong, Sun, L., Han, E., Ernest, T., 2016. Nondestructive Measurement of Soluble Solids Content in Apples by a Portable Fruit Analyzer. Food Anal. Methods 9, 785–794. https://doi.org/10.1007/s12161-015-0251-2

Zhang, B., Huang, W., Gong, L., Li, J., Zhao, C., Liu, C., Huang, D., 2015. Computer vision detection of defective apples using automatic lightness correction and weighted RVM classifier. J. Food Eng. 146, 143–151.

https://doi.org/10.1016/j.jfoodeng.2014.08.024

Zhang, T., Xia, D., Tang, H., Yang, X., Li, H., 2016. Classification of steel samples by laser-induced breakdown spectroscopy and random forest. Chemom. Intell. Lab. Syst. 157, 196–201. https://doi.org/10.1016/j.chemolab.2016.07.001

Zheng, W., Fu, X., Ying, Y., 2014. Spectroscopy-based food classification with extreme learning machine. Chemom. Intell. Lab. Syst. 139, 42–47. https://doi.org/10.1016/j.chemolab.2014.09.015

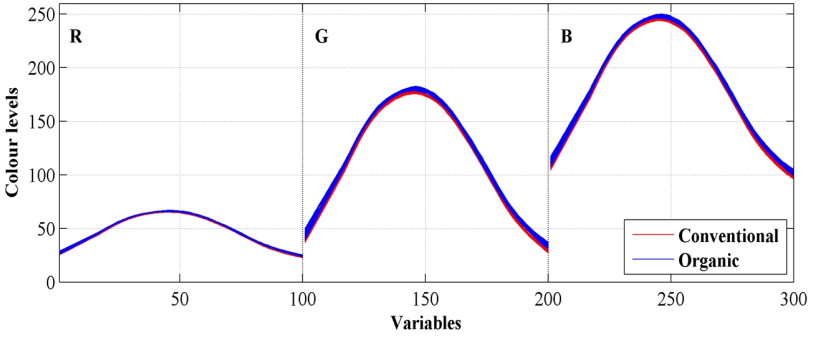
Zou, H.Y., Wu, H.L., Fu, H.Y., Tang, L.J., Xu, L., Nie, J.F., Yu, R.Q., 2010. Variable-weighted least-squares support vector machine for multivariate spectral analysis. Talanta 80, 1698–1701. https://doi.org/10.1016/j.talanta.2009.10.009

HIGHLIGHTS

- A low-cost sensor system was used to differentiate organic apples from conventional ones.
- Ten machine learning algorithms were evaluated using rainbow image data from the sensor system.
- The classification results of rainbow image data were comparable to that of spectral data.

Conventional and organic apples

Rainbow image data



Classification