Development of non-destructive mango assessment using Handheld Spectroscopy and Machine Learning Regression

Dheya Galal Abdullah Al-Sanabani¹, Mahmud Iwan Solihin¹,⁴, Liew Phing Pui², Winda Astuti³, Chun Kit Ang¹, Lim Wei Hong¹

¹ Mechatronics Engineering, Faculty of Engineering, UCSI University, Malaysia
² Food Science Dept., Faculty of Applied Science, UCSI University, Malaysia
³ Computer Engineering Dept., Faculty of Engineering, Bina Nusantara University
⁴ E-mail: mahmudis@ucsiuniversity.edu.my

Abstract. Quality determines the shelf-life and selling prices of fresh mango, and therefore quality observation and control of fresh mango are of utmost significance in the processing and management of its supply chain. Mango fruit (mangifera indica) quality methods are mostly destructive in nature. Different mechanical, electromagnetic and non-destructive methods are increasingly important nowadays because of the ease of operation, speed, and reliability of the process. This project aims to develop a non-destructive assessment of mango quality using handheld micro NIR (near-infrared) spectroscopic device. NIR spectra data and Brix levels, which indicate the sugar content of the plant, i.e. indicating the sweetness of the mango, were collected from three different types of Mango (Chokanan, Rainbow, and Kai Te), resulting 80 samples (i.e. 60 samples for training and 20 samples for testing) in this project. NIR spectra can be converted mathematically to obtain quantitative information of chemical and physical nature by multivariate calibration. The spectra data is pre-processed using Gaussian smoothing and extended multiplicative signal correction (EMSC) for the elimination of uncontrollable path length or scattering effects. These samples were then used to develop a predictive model using both Support Vector Machine (SVM) regression and Partial Least Squares regression (PLS) methods. The coefficient of determination (R²) obtained from SVM for training/calibration and testing dataset are 0.96 and 0.95 respectively. Meanwhile, the coefficient of determination (R²) obtained from PLS for calibration/training and testing dataset are 0.89 and 0.86 respectively. The results obtained from this project indicate that the handheld NIR has potential use for non-destructive assessment of mango fruits quality.

1. Introduction
Mango (Mangifera indica) is one of the highest nutrients of carbs, proteins, fats, minerals, and vitamins contained in fruits. As mangoes mature, glucose, fructose, and sucrose levels rise, while vitamin C levels decrease [3]. Fruit quality can be classified as an external and internal component in two common groups. Interior quality may include texture, flavors, the value of nutrition and factor defects. Sizing, shape, and color are generally recognized as factors of external quality.

Several methods for measuring the internal quality of fruits have been tried, but most of these techniques still are invasive, requiring a lot of time and labor. Some research has examined an accelerated and simpler method of evaluating fruit quality using technology for optical spectroscopy. Near-infrared spectroscopy [NIR, (750-2500 nm)] shows the best results compared to other
electromagnetic spectrum wavelength ranges [4]. NIR spectroscopy is used to analyze ingredients, intermediates, and final products compositionally, functionally and sensitively. It is used in food and feed, farm, dairy, pharmaceutical, and chemical industries which are constantly under pressure to produce products which satisfy customer demands while increasing plant production and profitability [4].

NIR can be used in quantitative analysis (substance concentrations determination), qualitative testing and process control (raw material identification, intermediate and finished product identification). It can provide moisture, protein, fat, and starch content information. In each industry, NIR applications vary and are tailored to suit different companies and their products and needs [1, 2, 5]. In assessing the fruit maturation of mango and as a guide to final food quality, short wave near-infrared spectroscopy (NIR) (900-1700 nm) has been investigated. To obtain a predictive model using spectroscopy data, real data needed to be collected so that it can be used to calibrate and validate the accuracy of the prediction model. Refractometer – A device used to measure the refractive index of plant juices in order to determine the mineral/sugar ratio of the plant cell protoplasm. The refractometer measured in units called Brix. NIRS is used to predict the Brix values in mango fruit. The mango fruits used as samples are of three different types namely: Chokonan, Rainbow, and Kai Te.

2. Experimental Setup

2.1. Spectra data collection

For this project, three different types of mango fruit were selected namely Chokonan, Rainbow, and Kai Te. Total of 60 samples was prepared to be scanned by the spectrometer. Samples were scanned by using handheld NIR spectrometer with a wavelength range of 900nm to 1700nm. Figure 1 is the picture of the handheld NIR spectrometer used in this research develop with Texas Instruments electronics board.

The samples were scanned in reflectance mode to record the absorbance spectra data. Each sample spectrum was measured for 3 seconds in reflectance mode. Some samples were scanned two times in different environment, and some were scanned only one time. A total of 80 spectra were collected from 60 samples. Spectral data was transfer in a form of the datasheet to both Orange software and MATLAB® simulation software for analysis.

The data measured by the spectrometer contains irrelevant data and affected by noise and scattering [6]. Per-processing spectral data is essential to get an efficient and accurate model [6], [7]. For this paper, Extended multiplicative signal correction (EMSC) is used to pre-process the spectra of polynomial order of 2. It is effective for elimination of uncontrollable path length or scattering effects. Under ideal conditions, the absorbance spectra data \( Z_{i,chem} \) can be seen as a sum of the contributions from the different chemical constituents with spectra \( K=\{k_j, j=1,2,...,J \} \) and concentrations \( c=\{c_{i,j}|j=1,2,...,J\} \):

\[
Z_{i,chem} = c_{i,1}k_1 + \cdots + c_{i,j}k_j + \cdots + c_{i,J}k_J. \tag{1}
\]

EMSC can be expressed as follows [1]:

\[
Z_i \approx a_iI_r + b_iZ_{i,chem} + d_i\lambda + e_i\lambda^2. \tag{2}
\]

Where \( Z_i \) is a measured spectrum, \( Z_{i,chem} \) is the ideal spectrum, \( I_r \) is the identity matrix, \( \lambda \) is the wavelength, and \( a_i, b_i, d_i, \) and \( e_i \) are scalar parameters obtained by calibration. Spectrum \( Z_{i,chem} \) may be rewritten as a deviation from a reference spectrum \( m \), which could be, for example, the average of a set of empirical spectra as follows:

\[
Z_{i,chem} = m + \Delta c_{i,j}k_j \tag{3}
\]
where $\Delta c_{ij}$ represents the deviations in the analyte and interference concentration compared with that of the reference sample.

![Figure 1. The handheld micro NIR spectrometer](image)

2.2. Brix level collection

Real data need to be collected to calibrate and validate the predictive model accurately to achieve a predictive model by using spectroscopy data. Refractometer – A tool used to measure plant juices refractive index to determine the plant cell protoplasm mineral/sugar ratio.

The MA871 is an optical instrument that employs the measurement of the refractive index to determine the % Brix of sugar in aqueous solutions as shown in Figure 2 [8]. The method is both simple and quick. Samples are measured after a simple user calibration with deionized or distilled water. Within seconds the instrument measures the refractive index of the sample and converts it to % Brix concentration units. The MA871 digital refractometer eliminates the uncertainty associated with mechanical refractometers and is easily portable for measurements in the field [8].

The measurement technique and temperature compensation employ methodology recommended in the ICUMSA Methods Book (Internationally recognized body for Sugar Analysis). Temperature (in °C or °F) is displayed simultaneously with the measurement on the large dual-level display along with icons for low power and other helpful message codes.

![Figure 2. MA871 Digital Refractometer](image)
2.3. **Support vector machine (SVM)**

Support vector machines (SVM) could provide a learning method that is used for both regression and classification, with a fast algorithm that yields good results for many learning tasks [10]. Support vectors are the training examples that comprise the support vector machine [11]. Support vector machines cannot handle nominal data, necessitating pre-processing that transforms the nominal data to numerical data. SVM models, which are based on the statistical learning theory, are a new class of models that can be used for predicting values.

2.4. **Partial Least Squares regression (PLS)**

Partial Least Squares Regression (PLS Regression) is a statistical method which has some relation to principal component regression; it finds a linear regression model by projecting the predicted variables and the variables observable into another space instead of finding hyperplanes of highest variance between responses and independent variables. Due to the projection of both X and Y data in new spaces, the PLS processes family are known as bilinear factor models. Partial least squares (PLS-DA) Discriminant Analysis is a variation used when the Y is classified as categorical.

PLS is used to identify basic relationships between two matrices (X and Y) i.e. a latent approach to modeling covariance structures in both spaces. A model PLS attempts to locate the multidimensional direction in the X space that explains the maximum multidimensional direction of variance in the Y space. PLS regression is particularly suitable when there are more variables than observations in the predictor matrix, and when the X values are multi-linear. In contrast, in such cases (unless it is regularized), standard regression will fail. [9], [12]

3. Results and Discussions

3.1. **Pre-processing**

The raw absorption spectra from 80 samples of 3 different types of mango acquired from the spectrometer of wavelength ranging from 900nm to 1700nm are all recorded in Figure 3. This spectral data is then preprocessed by Extended multiplicative signal correction (EMSC) to eliminate the uncontrollable path length or scattering effects from the spectra data of polynomial order of 2 as shown in Figure 4.

![Figure 3. Raw spectral data](image-url)
3.2. **Support vector machine (SVM) Regression**

The transformed data is then divided into two subsets where 80% of the data is used to train the SVM prediction model and the remaining 20% of the data are used for testing. Regression cost (C) set for 100 and Complexity bound (v) set to 0.50. The coefficient of determination ($R^2$) obtained from SVM for training and testing dataset are 0.96 and 0.95 respectively. Table 1 shows the results for the predictive values compared to the actual values.

3.3. **Partial Least Squares regression (PLS) Regression**

PLS regression methods were used to find a linear regression model which gave a linear relationship between the multivariate NIR data and the Brix levels.

| Sample | Actual values | Predicted values |
|--------|---------------|------------------|
| 1      | 10.0          | 10.5             |
| 2      | 11.4          | 10.9             |
| 3      | 13.6          | 12.4             |
| 4      | 13.7          | 13.6             |
| 5      | 14.1          | 15.5             |
| 6      | 15.0          | 15.8             |
| 7      | 16.4          | 16.6             |
| 8      | 16.7          | 18.2             |
| 9      | 13.9          | 13.3             |
| 10     | 14.3          | 12.4             |
| 11     | 14.5          | 14.2             |
| 12     | 15.6          | 14.8             |
| 13     | 14.6          | 14.7             |
| 14     | 15.8          | 15.0             |
| 15     | 16.2          | 14.6             |
| 16     | 14.7          | 14.8             |
| 17     | 15.9          | 15.5             |
| 18     | 16.6          | 17.5             |
| 19     | 14.8          | 16.5             |
| 20     | 16.1          | 15.2             |
MATLAB® simulation software was used for generating PLS calibration models and testing these model’s accuracy. The 80 samples pre-processed data is then divided into two subsets where 80% of the data is used to calibrate the PLS prediction model and the remaining 20% of the data are used for testing. The coefficient of determination (R²) obtained from PLS for calibrating and predicting dataset are respectively 0.89 and 0.86. Table 2 shows the results for the predictive values compared to the actual values. This regression result for both predicted values and actual values is also shown in Figure 5. Furthermore, Table 3 shows the comparison for regression using both PLS and SVM that indicate that the SVM regression produces better result in this project in terms of coefficient of determination (R²) and RMSE (root mean squared error).

Table 2. PLS prediction values compared to actual values (%Brix)

| Sample | Actual values | Predicted values |
|--------|---------------|------------------|
| 1      | 9.6           | 11.2             |
| 2      | 15.8          | 15.8             |
| 3      | 16.1          | 14.9             |
| 4      | 17.1          | 14.4             |
| 5      | 17.8          | 16.8             |
| 6      | 7.8           | 8.6              |
| 7      | 7.0           | 9.1              |
| 8      | 12.2          | 12.8             |
| 9      | 14.1          | 14.8             |
| 10     | 18.1          | 20.2             |
| 11     | 9.9           | 10.2             |
| 12     | 11.4          | 11.6             |
| 13     | 14.3          | 13.3             |
| 14     | 15.8          | 14.5             |
| 15     | 15.9          | 15.2             |
| 16     | 16.2          | 15.4             |
| 17     | 10            | 8.6              |
| 18     | 13.7          | 12.8             |
| 19     | 16.4          | 16.9             |
| 20     | 16.6          | 14.8             |

Figure 5. The plot of regression result using PLS for calibration ($R^2 = 0.89$) and testing ($R^2 = 0.86$)
4. Conclusions
Methods for mango fruit quality are normally destructive in nature. Several methods were attempted to measure the inner quality of fruits, but most of these techniques are still invasive and take a lot of time and work. The ease of operation, speed, and reliability of processes make various mechanical, electromagnetic, and non-destructive methods increasingly important today. Recent studies have investigated an accelerated and simpler technique of fruit quality assessment using optical spectroscopy technology.

The results in this study showed that the NIR spectroscopy technique can be a promising tool to assess fruits quality and other products non-destructively. The NIR spectral data collected from the mango fruits and Brix levels are interrelated. The coefficient of determination ($R^2$) obtained from SVM for calibration and testing dataset are 0.96 and 0.95 respectively. Meanwhile, the coefficient of determination ($R^2$) obtained from PLS for calibration and testing dataset are 0.89 and 0.86 respectively. This project will be further improved, for example adding additional mango samples and software integration to test the calibrated tool on-site.

Acknowledgment
The authors would like to express the gratitude to the Centre of Excellence for Research, Value Innovation, and Entrepreneurship (CERVIE) – UCSI University for supporting this project.

References
[1] Martens H, Stark E, “extended multiplicative signal correction, Journal of Raman Spectroscopy, pp. 119–136. 2016.
[2] Skogholt, Joakim & Liland, “Preprocessing of spectral data in the extended multiplicative signal correction, Journal of Raman Spectroscopy, 10.1002/jrs.5520. 2018.
[3] Prusky, D., Kobiler, I., Zaubermann, G. and Fuchs, Y. Pre harvest conditions and post harvest treatments affecting the incidence of decay in mango fruits during storage. International Society of Horticulture Science (International Mango Symposium) 5: 307-320. 1993.
[4] Williams,P., Norris, K., Near Infrared Technology in the Agricultural and FoodIndustries. American Association of Cereal Chemists, Inc, St. Paul, MN.G.Y. 1987.
[5] Manley, Marena., “Near-Infrared Spectroscopy and Hyperspectral Imaging: Non-Destructive Analysis of Biological Materials. Chemical Society reviews. 43. 10.1039/c4cs00062e., 2014.
[6] 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.
[7] X. Wang, Near-infrared spectroscopy for food quality evaluation. Elsevier Inc., 2019.
[8] “Milwaukee - MA871 Digital Brix Refractometer.” [Online]. Available: http://www.milwaukeeinst.com/site/products/products/digital-refractometers/165-products-g-digital-refractometers-g-ma871. [Accessed: 22-Aug-2019].
[9] H’sokuldson, A. Pls regression methods. Journal of Chemometrics, 2, 211-228. 1988.
[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, vol. 345 AICT, no. PART 2, pp. 623–632, 2011.
[12] Geladi, P., & Kowalski B. Partial least square regression: A tutorial. Analytica Chemica Acta, 35, 1-17.Biomol. Spectrosc., vol. 120, pp. 625–629, 1986.