Estimation of Harumanis (Mangifera indica L.) Sweetness using Near-Infrared (NIR) Spectroscopy

M S Amirul¹,², R Endut¹,², C B M Rashidi¹,², S A Aljunid¹,², N Ali³, M H Laili⁴, A R Laili¹,⁴ and M N M Ismail²

¹School of Computer and Communication Engineering (SCCE), Universiti Malaysia Perlis, 02600, Arau, Perlis, Malaysia.
²Centre of Excellence Advance Communication Engineering (ACE) Optics, SCCE, Universiti Malaysia Perlis, 01000, Kangar, Perlis, Malaysia.
³School of Microelectronic Engineering (SoME), Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia.
⁴Photons R&D, Department of Industrial Electronics Technology (IET), MIMOS Berhad, Technology Park Malaysia, 43300 Kuala Lumpur, Malaysia.

roshdisham@unimap.edu.my

Abstract. Harumanis mango quality demanded by consumers is depending on the sweetness level of the fruit. The sweetness is evaluated by brix percentage using refractometer as a representative factor correlated with near-infrared (NIR) spectroscopy spectral absorbance. NIR spectroscopy method of sampling have been tested to overcome the time consuming, complex chemical analysis more importantly invasive sampling methods in order to determine the sugar content in mangoes. Spectral absorbance data from range 941 nm to 1685 nm of mango skin is correlated with Brix reading then tested through five pre-processing techniques. Data calibration and prediction of both data is evaluated using Partial Least Square Regression (PLSR) model. In the final analysis, Unit vector normalization (UVN) technique has achieved as a best pre-processing technique for predicting results, with the coefficient of determination ($R^2$) values of 0.9836 and root mean square error (RMSE) values of 0.3131. Overall, the correlation of NIRS absorbance data and Brix data can be obtained using PLSR model with UVN pre-processing technique. Henceforth, we can conclude that the NIRS method of sampling can be used to identify sugar content in Harumanis mango by using time saving, non-invasive and less laborious method of sampling.

1. Introduction

Harumanis mango (Mangifera indica L.) is one of the most popular tropical fruit in Malaysian markets, which is mainly cultivated and produced in the state of Perlis. The Harumanis prefer the lime structure of the soil and climax changes in Perlis and Kedah, which make it the only state that suited the mango plant growth [1]. Therefore, the limited size of cultivation, quality produced and seasonal has made this fruit really expensive and consumer prefer the best quality they ever produced. Consumer preference is mainly driven by sweetness, which is dependent on physiological parameters described by various researchers [2–4]. Increase of total soluble solids (TSS), carotenoid pigments and decrease in acidity are some indicators of the sweetness of mango [5]. Currently, most consumers determine these by experiencing surface firmness, gloss, aroma, flavour, etc., which is often misleading [6]. Even for other...
quality demands fruits and vegetables, many industries have been conducting studies for maturity and quality indices, but most of them are of a chemical or destructive method, and their determination involve very laborious techniques [7–9].

Currently, near-infrared (NIR) technology is a non-destructive method for analysing the ingredients of organic products, and it is being successfully implemented throughout the food industry [7]. NIR systems are applied both in laboratories and in on-line production processes and has been used for quality evaluation [10]. These techniques are rapid and easy to conduct. Many physical characteristics of fruits and vegetables have been determined non-destructively [11, 12]. The objective of this study was to examine the potential of NIR spectrometer to predict the sweetness of Harumanis using non-invasive, rapid and non-laborious method of sampling. Besides that, this study will analyse the performances of Partial Least Square Regression (PLSR) model with different data pre-processing techniques for the prediction of mango sweetness using NIR spectroscopy. Last but not least, to find the range of wavelength in NIR spectrum, which is highly correlated with sweetness properties.

2. Materials and methods

2.1. Mango skin and juice sampling

For sampling, 30 Harumanis mangoes is taken at the peak time of the fruit season, which is between April to June [13]. The mango is selected randomly covering different size, firmness, maturity and color, but plucked from the tree at the same time and same farm. The mango then will be kept at room temperature for about 1 week to achieve the required maturity which will consequently increase the sucrose content in the fruit [14]. The mango sample was split into two types of sample which is the juice and skin. Firstly, the mango is labelled as 24 sections at both sides, including the bottom and upper part of the fruit as shown in figure 1, giving in 1440 readings for overall juice and skin data.

![Figure 1](image.png)

**Figure 1.** For every sides of mango, 12 section is labelled on the skin

The juice sample will also be taken at the same spot according to the skin section labelled. Juice sample will be used by refractometer for sweetness calibration data, while skin sample will be used by NIR spectrometer for prediction in further correlation between both data.

2.2. Calibration and spectral data measurement

Hanna HI96801 (Hanna Instrument Inc., Woonsocket, USA) portable refractometer has been used to identify the sugar content of the Harumanis mango juice samples. The device converts the refractive index of food sample to % Brix and ideal for the analysis of fruits and various foods. The measurement is required to be prepared in liquid form. The conversion is based on the tables found in the International Commission for Uniform Methods of Sugar Analysis (ICUMSA) Methods Book that documents the changes in refractive index with temperature for a percent by weight sucrose solution [15]. The refractometer has a Brix range of 0% to 85% with an accuracy of ± 0.2% and this will be used as calibration data.

For spectral data, the absorbance will be measured using the DLPNIRscan Nano Evaluation Module (EVM) (Texas Instrument, Texas, USA). NIRscan data are presented in absorbance absorbance contrary
to reflectance given by the light flashed to the surface of the mango skins and juice. Absorbance is
detailed in spectrum and measured from 900 nm to 1700 nm for every data reading [16]. Before taking
the measurement, NIRscan need to calibrate with Polytetrafluoroethylene (PTFE) compound, which
giving the highest reflectance and lowest absorbance in order for the device to give best result with less
calibration error [17]. For every measuring point, 10 readings are taken consequently just in about 3
second by the NIRscan. Then, the reading is presented as one reading only after averaging automatically
by the NIRscan systems itself.

![Figure 2](Image)

**Figure 2.** Technique of sampling the mango skin and juice using NIR spectrometer.

Firstly, NIRscan is used to measure the spectral data of mango skin which have been labeled. This
data will be used to correlate with Brix data in order to achieve this study objective. After that, the
labelled section of the mango is sliced one by one, blended and sieved to prepare the juice samples for
refractometer measurement. Besides, the spectral data of the juice samples are also taken using NIRscan
with a specialized setup which needed the juice samples to be prepared in a cuvettes. Figure 2 shows the
technique of sampling the mango skin and juice. This data will be used to check the correlation between
Brix data and mango juice spectral data. Finally, after saving both calibration and spectral data, the
spectral data are plotted to analyze the pattern and identify measurement error from the absorbance
response of the soil samples specifically to avoid further error in the correlation.

### 2.3. Pre-processing technique

Several first and last data points were removed from original absorbance spectra to avoid data
interruption and noise [18], leaving remaining 941 nm to 1685 nm from overall NIR spectrum
wavelength. Besides that, the data are also going through the pre-treatment process to remove negative,
null and harsh data. Moreover, this process also can improve performance of the PLSR model [4]. Five
out of various pre-processing techniques have been selected after being tested and analyzed
consequently. Unit vector normalization (UVN), Multiplicative scatter correction using common
amplification (MSCCA), multiplicative scatter correction using common offset (MSCCO),
multiplicative scatter correction (MSC) and mean normalization (MN) technique is selected to treat the
spectral data in this study because of their optimum performance in the PLSR model testing. The
comparison of these pre-processing has been carried out to determine the best pre-processing technique
and has been concluded in the results by evaluating the coefficient of determination ($R^2$) and root mean
square error (RMSE).

### 2.4. Prediction model

Partial least square regression (PLSR) is a method for constructing predictive models when the factors
are many and highly collinear [19]. In other words, the model emphasis on predicting the responses and
not necessarily trying to understand the underlying relationship between variables. The X and Y-scores
are selected with the goal that the connection between progressive sets of scores is as solid as could be
expected under the circumstances. Fundamentally, this resembles a robust form of redundancy analysis,
looking for direction in the factor space that is related to high variability in the reactions yet biasing
them toward a direction that is accurately predicted [19].
PLSR generalizes and combines features from principle component analysis (PCA) and multiple regression (MR) to predict or analyze a group of dependent variables from a group of variables or predictors. Eventually, a prediction is retrieved by extracting it from the predictors set of orthogonal factors called latent variable (LV) which consisting the best predictive power. The orthogonal basis of LV is constructed by PLSR, aligned along the direction of maximal covariance between the spectra matrix X and the response Y. As for this analysis, the maximum number of LV is set at 10 because any number higher than this is resulting in non-actual data [20]. For this reason, PLSR will be preferred as the prediction model to validate the correlation between calibration data of Brix and NIR spectral data. Before developing the calibration model, two groups of data consisting of 75% for calibration and 25% for prediction are randomly distributed from the overall sample data. This ratio is the ideal ratio preferred for the prediction model analysis that is performed in previous studies [21]. Full cross-validation is selected to evaluate the quality and avoid over-fitting of the graph to achieve the calibration models while the prediction set validates the models. PLSR analysis and pre-processing techniques are tested and evaluated using statistical software named ‘The Unscrambler X’ (version 10.4, Camo Process AS, Oslo, Norway).

3. Result and discussion

Figure 3 shows a spectral plot for the overall 720 reading of Harumanis mango skins in the NIR spectral range of 941 nm to 1685 nm. The curve show respond of absorbance varies with a wavelength between all of the skin samples. From this graph, skin sample absorption is intensively high at range 1424 nm to 1560 nm. Henceforth, these absorbance were then analyzed using a prediction model to evaluate the correlation between absorbance and Brix data reading.

![Figure 3. Spectral plot of mango skin samples absorbance.](image)

As a result, both Brix and spectral absorbance data is then evaluate using the PLS regression prediction model and the top performer’s pre-processing technique is then selected. Out of 20 pre-processing techniques, five that shown interest are UVN, MSCCA, MSCCO, MSC, and MN. In table 1, the evaluation of Brix data and spectral absorbance is presented with a different pre-processing technique that has been tested.

The selected absorbance spectral is spilled into 7 range section that is distributed evenly, and section 7th (1578 nm to 1684 nm) is selected as the observation range to be analyzed because this range shows a highly anticipated correlation for both data. For this analysis, $R^2$ and RMSE will be the benchmark of performance for the correlation between Brix data and spectral absorbance data. In spite of $R^2$, the best performance for the correlation to happen is between the ranges of 0.70 to 0.90, or even better if possible,
nearer to 1.00. On the contrary, best RMSE are evaluated in ranges of 0.10 to 0.90 or even better if possible to reach nearer 0.00 [22]

Table 1. PLSR evaluation result with optimum pre-processing technique

| Pre-processing technique                                      | Calibration | Prediction |
|--------------------------------------------------------------|-------------|------------|
|                                                             | $R^2$       | RMSE       | $R^2$       | RMSE       |
| Unit Vector Normalization (UVN)                              | 0.9856      | 0.2926     | 0.9836      | 0.3131     |
| Multiplicative Scatter Correction using Common Amplification (MSCCA) | 0.9859      | 0.2831     | 0.9832      | 0.3175     |
| Multiplicative Scatter Correction using Common Offset (MSCCO) | 0.9214      | 0.6843     | 0.9096      | 0.7366     |
| Mean Normalization (MN)                                     | 0.9072      | 0.7439     | 0.8865      | 0.8265     |
| Multiplicative Scatter Correction (MSC)                     | 0.8830      | 0.8353     | 0.7962      | 1.1075     |

Figure 4. PLSR prediction using UVN technique

Figure 5. PLSR prediction using MSCCA technique

Figure 6. PLSR prediction using MSCCO technique

Figure 7. PLSR prediction using MN technique
Figure 8. PLSR prediction using MSC technique

Graph plot in figure 4 to figure 8 shows the correlation of both calibration and prediction of Brix and spectral data. The red plot represent predicted data and the blue plot represents calibrated data. The line represent regression line for the correlation. From table 1, the MSCCA technique shows the most anticipated and intended calibration with R² values of 0.9859 and RMSE of 0.2831. However, the prediction result is at the second-best with R² values of 0.9832 and RMSE values of 0.3175 as shown in figure 5. More importantly, the prediction result of UVN is in the best-expected range as shown in figure 4, in which R² is at the values of 0.9836 and RMSE of 0.3131 which little lower compared to calibration. For calibration, the UVN technique able to achieve R² values of 0.9856 and RMSE values of 0.2926. This evaluation shows that prediction for Brix data is highly correlated to specified spectral absorbance data using this technique.

As we can observe from figure 7, the MN technique has achieved less linear plot if compared with MSCCA and UVN technique. With the values of 0.9072 for R² and 0.7439 for RMSE, it is considered 2nd lowest performance for calibration after the MSC technique. Looking at the prediction, it can only achieve R² values of 0.8865 with RMSE values of 0.8265. This is not even close is compared to the UVN technique, but nearly following the prediction values for the MSCCO technique. Comparatively, the MSCCO technique as shown in figure 6 have accomplished as 3rd highest calibration performance with R² values of 0.9214 and RMSE of values 0.6843. Although, compared to the predicted result, the MSCCO technique also achieved 3rd highest performer with R² values of 0.9096 and RMSE values of 0.7366. Meanwhile, observing the MSC technique in figure 8, the predicted result is still acceptable even though the prediction RMSE is out of the targeted range. MSC technique achieved calibration R² values of 0.8830 and RMSE values of 0.8353, while for prediction R² values of 0.7962 and RMSE values of 1.1075. Given this point, we can assure that the UVN technique has scored the best pre-processing technique so far compared to the other four techniques tested in this study for correlation of Brix and spectral absorbance data. As a suggestion for future research, a lot of other variables should be considered for observation which can also affect the performance of the prediction model such as firmness of the mango and also the color intensity of mango. Besides that, to increase correlation and the accuracy of the prediction model, a larger number of samples are expected.

4. Conclusion

In brief, after going through overall testing and observation, UVN pre-processing technique has been identified as the best performer for the analysis of Brix data correlation with spectral absorbance data of the skin samples of the Harumanis mango. Moreover, we can summarize that using NIRS, sweetness level of Harumanis mango can be identified more efficiently using time-saving, non-invasive and less laborious methods of sampling, enhanced with the UVN pre-processing technique and PLSR model. For future research, numerous fruits can be proposed and tested for sweetness estimation using the NIRS instrument as an efficient tool.
Acknowledgements
The author would like to acknowledge the support from the Fundamental Research Grant Scheme (FRGS) under a grant number of FRGS/1/2015/TK04/UNIMAP/03/2 from the Ministry of Education Malaysia.

References
[1] Dermawan A. Harumanis takes root in Kedah. New Straits Times, [available online]: https://www.nst.com.my/news/nation/2019/06/493403/harumanis-takes-root-kedah (2019).
[2] Jha S N, Jaiswal P, Narasiah K, et al. Non-destructive prediction of sweetness of intact mango using near infrared spectroscopy. *Sci Hortic (Amsterdam)* 2012 **138** p 171–175.
[3] Taira E, Nakamura S, Hiyane R, et al. Development of a nondestructive measurement system for mango fruit using near infrared spectroscopy. *Eng Appl Sci Res* 2017 **44** p 189–192.
[4] dos Santos Neto J P, de Assis M W D, Casagrande I P, et al. Determination of ‘Palmer’ mango maturity indices using portable near infrared (VIS-NIR) spectrometer. *Postharvest Biol Technol* 2017; **130** p 75–80.
[5] Wanitchang P, Terdwongworakul A, Wanitchang J, et al. Non-destructive maturity classification of mango based on physical, mechanical and optical properties. *J Food Eng* 2011; 105: 477–484.
[6] Jha S N and Ruchi G. Non-destructive prediction of quality of intact apple using near infrared spectroscopy. *J Food Sci Technol* 2010; **47** p 207–213.
[7] Khodabakhshian R, Emadi B, Khojastehpour M, et al. Non-destructive evaluation of maturity and quality parameters of pomegranate fruit by visible/near infrared spectroscopy. *Int J Food Prop* 2017 **20** p 41–52.
[8] Magwaza L, Tesfay S. A Review of Destructive and Non-destructive Methods for Determining Avocado Fruit Maturity. *Food Bioprocess Technol*; 2015; **8**. DOI: 10.1007/s11947-015-1568-y.
[9] Kumari P, Feza M and Mir H. Non-destructive quality evaluation by sensing maturity and ripening of fruits and vegetables. 2018 p 84–89.
[10] Porep J U, Kammerer D R and Carle R. On-line application of near infrared (NIR) spectroscopy in food production. *Trends Food Sci Technol* 2015; **46** p 211–230.
[11] Beghi R, Giovenzana V, Tugnolo A, et al. Application of visible/near infrared spectroscopy to quality control of fresh fruits and vegetables in large-scale mass distribution channels: a preliminary test on carrots and tomatoes. *J Sci Food Agric* 2018; **98** p 2729–2734.
[12] Wang H, Peng J, Xie C, et al. Fruit Quality Evaluation Using Spectroscopy Technology: A Review. *Sensors* 2015; **15** p 11889–11927.
[13] Zakaria A, Shakaff A Y M, Masnain M J, et al. Improved Maturity and Ripeness Classifications of Magnifera Indica cv. Harumanis Mangos through Sensor Fusion of an Electronic Nose and Acoustic Sensor. *Sensors* 2012, **12**. DOI: 10.3390/s120506023.
[14] Saad F S A, Shakaff A Y M, Zakaria A, et al. Bio-inspired Vision Fusion for Quality Assessment of Harumanis Mangos. In *2012 Third International Conference on Intelligent Systems Modelling and Simulation*, 2012, p 317–324.
[15] Hanna Instruments Inc. *HI 96801 Refractometer for Sucrose Measurements*. Woonsocket, USA: Hanna Instruments Inc., 2014.
[16] Thangappan A and Pathangay V. Non-Invasive Blood Glucose Estimation from Near Infrared Spectrum using Gradient Boosted Tree Models. *IEEE-EMBC* 2016; **3**.
[17] Inácio M R C, de Lima K M G, Lopes V G, et al. Total anthocyanin content determination in intact açaí (Euterpe oleracea Mart.) and palmitero-juçara (Euterpe edulis Mart.) fruit using near infrared spectroscopy (NIR) and multivariate calibration. *Food Chem* 2013; **136** p 1160–1164.
[18] Hervé A. Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdiscip Rev Comput Stat* 2010; **2** p 97–106.
[19] Tobias R D. An introduction to partial least squares regression. *SAS Conf Proc SAS Users Gr Int 20 (SUGI 20)* 1995; p 2–5.
[20] Anusia H, Jayaselan J, Nawi N M, et al. Application of Spectroscopy for Nutrient Prediction of Oil Palm. 2017; 15 p 1–9.

[21] Mat Nawi N, Chen G, Jensen T, et al. Prediction and classification of sugar content of sugarcane based on skin scanning using visible and shortwave near infrared. Biosyst Eng 2013; 115 p 154–161.

[22] Kyprianidis K G and Skvaril J. Developments in Near-Infrared Spectroscopy. Croatia: Janeza Trdine 9, 51000 Rijeka, Croatia, 2017. DOI: http://dx.doi.org/10.5772/62932.