Non-Destructive Prediction of Papaya Maturity Level Using Near-Infrared Spectroscopy and Artificial Neural Network

Herlina Abdul Rahim[1*], Norliana Shukri[1], Nur Athirah Syafiqah Noramli[2], Indrabayu[3], Aisyah Mohd Akram[4]

[1*]Department of Control and Mechatronics Engineering, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[2]School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, Shah Alam, Selangor. Malaysia
[3]Department of Informatics, Faculty of Engineering, Hasanuddin University, Makassar, Indonesia
[4]Regional Development & Strategy, Siemens Energy, Cyberjaya, Selangor

Corresponding author* email: herlina@utm.my
Available online 30 December 2025

ABSTRACT

This paper presents a non-invasive technique for evaluating the maturity of Carica Papaya L. by predicting its Soluble Solid Content (SSC) using Near-Infrared (NIR) spectroscopy integrated with an Artificial Neural Network (ANN) model. Traditional destructive methods using refractometers hinder quality preservation and consumer usability. In contrast, the proposed approach utilizes NIR spectral reflectance data, pre-processed with Savitzky-Golay (SG) smoothing and derivatives, and is analyzed via a nonlinear ANN regression model. Experimental results based on 49 papaya samples show high predictive accuracy (R^2 = 0.9063 for training, 0.8768 for testing; RMSE = 0.4406 and 0.7047 respectively) using second derivative data. This study demonstrates the feasibility of portable NIR systems for real-time fruit maturity classification, supporting broader applications in agricultural and supply chain contexts.

Keywords: Carica papaya, Soluble Solid Content, Near-Infrared Spectroscopy, Artificial Neural Network, Non-destructive testing, Savitzky-Golay, SSC prediction

1. Introduction

Soluble Solid Content (SSC), a key index for determining the sweetness and ripeness of fruits, is directly associated with consumer preference and commercial value. It is commonly measured using refractometers, which require juice extraction and thus result in fruit wastage. Such destructive testing is unsuitable for in-line or post-harvest quality control processes where fruit integrity must be preserved.

In order to meet human needs for greater fruit quality, the fruit industry must deal with new technologies and methods on a global scale. The maturity level of fruits is heavily dependent on when they are received, how they are transported and how they are stored before being purchased by customers. The conventional method to check fruit quality by means of Soluble Solid Content (SSC) and other parameters is time consuming and tedious [1]. As a result, a brand-new non-destructive method using (NIR) spectroscopy was invented to check the quality attribution of various products. NIR spectroscopy measures the electromagnetic spectrum between 780 nm and 2526 nm [2]. The idea underlying NIR spectroscopy is based on the optical properties associated with light. The reflectance mode calculated the amount of reflected light that was emitted from the product surface [3].

Papaya or genetically known as Carica Papaya L. is from the family Caricaceae. Papaya is a tropical, semi-woody and herbaceous plant that grows quickly. According to reports, the largest papaya producer was India, with profit of 13.9 million metric tonnes daily [4]. The value of the SSC of papaya ranges from 5.19±0.33°Brix for the unripe to 11.06±0.13 °Brix for the overripe [5]. A good predictive model method should be selected to process the spectral data. There are two calibration models for regression which are linear and nonlinear. In some studies, the nonlinear model specifically the Artificial Neural Network (ANN) improved the accuracy of prediction far beyond the linear model, the Partial Least Squares (PLS) regression model. However, the ANN model is very complex to be established [6].

Near-Infrared (NIR) spectroscopy presents an efficient and non-invasive alternative to traditional SSC measurement methods. This technique capitalizes on the vibrational absorption of specific molecular bonds (e.g., O-H, C-H, C=O) by NIR light, which provides unique spectral patterns that can be associated with sugar content and internal fruit quality. Recent advances in machine learning have enabled improved prediction performance by modeling complex nonlinear relationships in spectral data. In this context, Artificial Neural Networks (ANNs) are well-suited to approximate the intricate mapping between NIR spectra and SSC values.

Evaluating the maturity level of fruit before marketing is very crucial. However, if the destructive method is used to measure the fruit maturity level, it is unavoidable to destruct the fruit, hindering the later purchase process. Moreover, the process of using refractometer is very time consuming and require large amount of labour [6]. This research investigates the feasibility of a low-cost, portable NIR spectroscopy setup, combined with an ANN regression model, to estimate SSC levels in papaya fruits at different maturity stages. The study further explores the impact of spectral pre-processing, particularly the Savitzky-Golay (SG) smoothing and derivative transformations, on model performance.

2. Literature Review

Prior studies have explored the use of spectroscopy for non-destructive assessment of fruit quality. Huang et al. [1] demonstrated successful SSC prediction in tomatoes using Vis/NIR transmittance spectra and multivariate calibration models. Similarly, Shah et al. [3] achieved accurate classification of mango maturity using NIR reflectance. Regression techniques such as Partial Least Squares Regression (PLSR) have been widely used but tend to perform poorly with complex and nonlinear relationships inherent in biological data. In contrast, ANN models have shown superior performance due to their ability to model such nonlinearities. For instance, Basile et al. [7] and Zhang et al. [8] reported improved SSC prediction accuracy using ANN models trained on pre-processed spectral data.

Spectral pre-processing methods significantly affect model reliability. Techniques like SNV and MSC reduce the influence of scattering effects, while SG derivatives help enhance peak visibility and minimize background noise. The choice of pre-processing method should be aligned with the type of spectral noise and the target chemical features. A common practice is to divide data into training and testing sets in a 70:30 ratio. This method balances learning and validation, allowing for robust performance assessment and minimizing overfitting.

2.1 Soluble Solid Content

Due to the NIR radiation, light scattering and absorption processes will occur inside the sample. When energy transitions from the ground state to the excited state, the atoms and molecules will vibrate such as by being stretched, oscillated or bent. During the energy transition, the photon energy can have similarity with a certain NIR wavelength and produce absorption.

There are two types of absorption, combination and overtones. Combination emphasized the distinctive absorption caused by several chemical bond vibrations at the total of the fundamental frequency wave numbers. Overtones refer to the distinctive absorption formed at multiples of fundamental frequency of the chemical bonds. The chemical bonds that react with the absorption are C-O, O-H, and N-H bonds inside the fruit.

However, the organic molecules that represent fruit ripening are C-C, C-O, O-H and C-H bonds [2]. These bonds represent the fruit sweetness that is correlated with fruit ripening. Fructose, glucose and sucrose are the three main components of SSC. Glucose becomes predominant sugar during fruit development. As fruit ripens, glucose and fructose undergo several enzyme metabolisms to be transformed into sucrose. Hence, the main sugar that contributes to fruit sweetness is sucrose [9].

2.2 Comparison Between Destructive and Non-Destructive Method

Nowadays, a non-destructive method is widely used for product quality in agriculture, pests and diseases by means of machine vision technology, NIR spectroscopy technology, hyperspectral imaging technology [10], acoustic technology and Computed Tomography (CT) scans [2]. NIR spectroscopy technology uses spectral information and chemometric techniques to detect qualitative and quantitative data. However, the efficiency of NIR also depend on environmental conditions.

The non-destructive method is preferable to the destructive method due to minimal damage to the sample [10], has simple procedure, open source and easy software to process the spectral data, a rapid process, pollution-free and cheaper [2]. In contrast, destructive methods prohibit repetitive testing and analysis for the same sample [11], have tedious procedure, need expertise to conduct the experiment and also need laboratory facilities [12].

2.3 Destructive Method Using Refractometer

A refractometer is as instrument used to measure the concentration of sugar in a substance and liquid in unit Brix, where one Brix degree is equivalence to 1% SSC by mass [13]. It simply calculated the percentage of glucose, fructose and sucrose in the sample [9]. It is commonly used in agriculture, food, chemical and manufacturing applications. In foods department, sugar and salt contents of liquid substances can be measured such as for winemaking, beverages and bakery products. There are two types of refractometer, analog and digital as depicted in Figure 1.

The principle of refractometer lies on the concept of refraction. The refraction occurs once the light passed through the substance at a certain angle and changes course. Thus, this refraction angle will determine the SSC value of the sample. The liquid with high sugar intensity will cause more refraction.

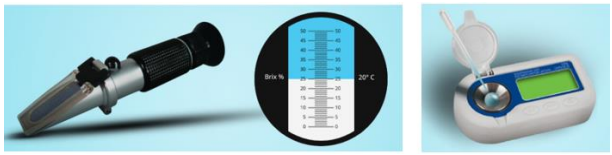

Figure 1. Analog and digital refractometer

2.5 Non-Destructive Method Using NIR Spectroscopy

NIR is an electromagnetic wave with a range of 780-2526 nm that is between the visible (Vis-NIR) range of 400-780 nm and the mid-infrared (MIR). NIR can be divided into two parts, which are the near-infrared shortwave range of 780-1100 nm and the near-infrared longwave range of 1100 – 2526nm [2]. Figure 2 shows the electromagnetic spectrum range, specifically the infrared wavelength.

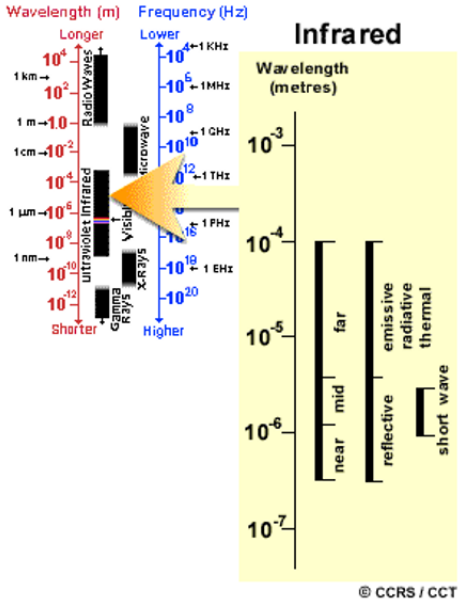

Figure 2. The electromagnetic spectrum of infrared [15]

The theory of NIR relies on the interaction between light and objects. When the infrared is emitted towards the fruit, the obtained spectral is in the mode of reflectance, transmittance or absorption [13]. The reflectance mode is commonly where the light scattered from the fruit surface is easily received, without touching the sample and the collected samples are distinguishable [2]. NIR depends on the correctness and precision of the collected data with reference methods before applying the prediction technique. It is advisable to select the samples wisely before the analysis for the sake of simplicity and high model performance [12].

2.6 Pre-Processing

The raw data should undergo pre-treatment to remove noise before establishing the predictive model such as scatter correction (Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), smoothing for noise reduction (Savitzky-Golay smoothing (SG-smoothing) with different window widths, polynomials and derivative orders) and scaling (mean center).

A derivative is used to remove the influence of baseline drift, smoothing and distinguish overlapping peaks from the spectrum [14]. However, the noise may increase in the process, so SG-smoothing is better to use after the derivation process. The selection of derivation order is highly dependent on the result of the output data. A higher order derivative may not provide any improvement to the data [11]. Data normalization refers to the process of rescaling the input data into a range of [-1,1]. Both normalization and denormalization can be done using the '*mapminmax*' function.

In data splitting, the calculation and optimization of the regression model can be done by cross-validation in the training set. For further assessment of the model's capability for prediction, the test set was used. The training set to test set ratios proposed are 8:2 [11], 3:1 [1], 7:3, 6:4 and 5:5. Note that the division for training and test samples can be randomly chosen as there is no clear instruction to find the optimal ratio [15].

The accuracy of training and test sets will then be evaluated by the Root Mean Squared Error (RMSE), coefficient of determination (*R*), bias and residual predictive deviation (RPD, ratio of standard error of performance to standard deviation) index [11]. However, the studies involving prediction on fruit quality usually used R and RMSE. RMSE can further divided into RMSEC that represent the calibration set, RMSECV that represent cross-validation set and RMSEP for prediction set. A higher value of *R* and lower value of RMSEC, RMSECV and RMSEP shows a good predictive model.

2.7 Artificial Neural Network (ANN)

ANN is a potent non-linear predictive model with pattern recognition capabilities. This technique can perfectly extract quantitative information from large databases that are intrinsically influenced by complex biological, environmental and instrumental variations. The only downside of ANN is the complexity of implementing the network, the setup for the method, training and parameter estimation compared to a linear regression model [11].

Analogously, the ANN principle is to build a mathematical model by simulating the structure and activity of human nerves. There are three layers of ANN: the input layer, the hidden layer and the output layer, where each layer consists of several neurons as shown in Figure 3. A directed curvature with an adjustable weight coefficient can be established by neurons in adjacent layers. The adjustable weight coefficient is determined by rapidly learning the information and finding the best processing technique [2].

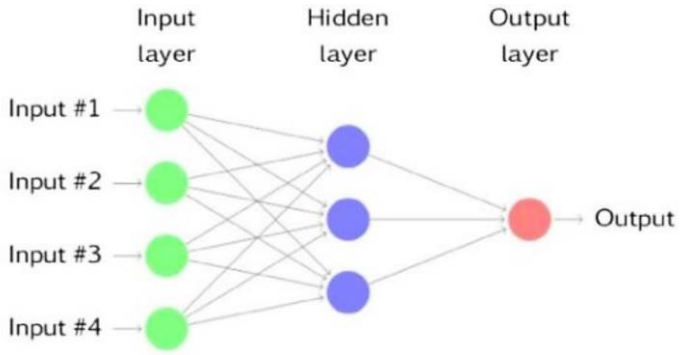

Figure 3. ANN architecture

3. Methodology

3.1 Data Acquisition

Papaya fruits (*Carica Papaya L.*) at three ripeness stages (underripe, ripe, overripe) were obtained from a local market. Each fruit was dissected into 20 zones to account for intra-fruit variability as shown in Figure 4. From this, 49 representative samples were obtained. Measurements were conducted in a controlled environment to reduce ambient light interference.

Figure 4. Each papaya was divided into 20 sections and labelled

Figure 5 shows the NIR spectra were collected using an Ocean Optics USB4000 NIR spectrometer (700–1000 nm range), coupled with a tungsten halogen lamp as the light source and a bifurcated fiber-optic probe for sample interaction. The probe was fixed 10 mm above the sample surface with a 90° orientation as shown in Figure 6. Integration time and gain settings were calibrated to avoid saturation and ensure high signal-to-noise ratio.

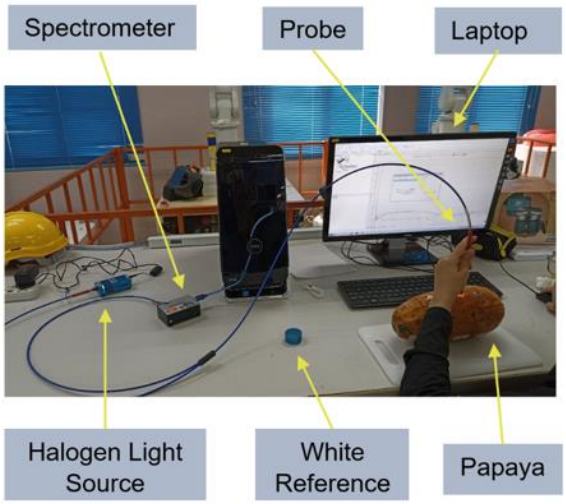

Figure 5. NIR spectroscopy method setup

Figure 6. The spectrometer probe was handled at 90°

3.2 Experimental Setup

Two techniques were performed for SSC analysis.

(a) Non-destructive NIR Measurement: Each sample's spectral reflectance data was recorded using SpectraSuite software. Measurements were repeated thrice and averaged as shown in Figure 7.

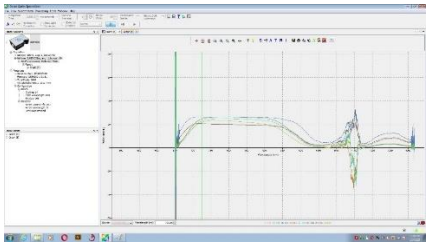

Figure 7. SpectraSuite software interface

(b) Destructive Refractometer Testing: Juice was extracted manually from each papaya segment using a garlic press, filtered, and analyzed using a digital refractometer to obtain SSC in °Brix as shown in Figure 8 and 9, respectively.

Figure 8. Papayas were cut into blocks

Figure 9. Procedure for refractometer method : a) The papaya block was crushed into juice, b) Zero calibration was obtained using distilled water, c) The residue on the optical prism was wiped out and d) The SSC value of papaya sample was obtained

3.3. Spectral Pre-Processing

Raw reflectance data was truncated to 700–1000 nm, as signal-to-noise ratio beyond this range deteriorated. Spectra were converted to absorbance. SG smoothing was applied with parameters.

(a) 0th derivative as shown in Figure 10, window size = 15, polynomial order = 2

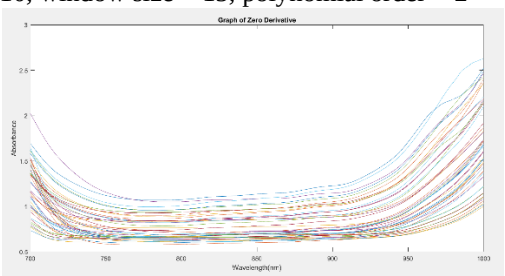

Figure 10. Zero derivative predicted SSC

(b) 2nd derivative as shown in Figure 11, window size = 35, polynomial order = 3

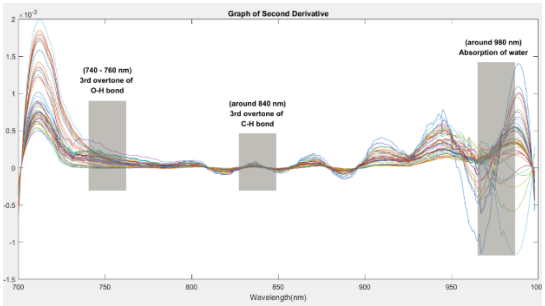

Figure 11. Second derivative predicted SSC

These settings were chosen to optimize the visibility of absorption bands without introducing distortion. Key absorbance bands appeared near 740–760 nm (O-H overtone), 840 nm (C-H stretch), and 980 nm (water absorption band).

3.4. ANN Model Development

A feedforward backpropagation ANN was developed using MATLAB R2022a Neural Network Toolbox. The architecture comprised.

(a) Input Layer: 300 spectral features (after preprocessing)
(b) Hidden Layer: 10 neurons, logsig activation
(c) Output Layer: 1 neuron, purelin activation
(d) Training Algorithm: Levenberg-Marquardt (trainlm)

Input and output data were normalized to the range [-1, 1] using mapminmax. The dataset was split randomly into training (70%) and testing (30%) using a fixed random seed (43) for reproducibility. Network performance was evaluated using coefficient of determination (R^2), root mean square error (RMSE), and regression slope.

Table 1. Training and testing distribution

Sample Set	Number of Sample	SSC (°Brix)			
		Min	Max	Mean	Std
Training Set	34	3.3682	8.8095	6.1260	1.2619
Testing Set	15	3.5202	7.5160	5.8083	1.3161
Total	49	3.3682	8.8095	6.0288	1.3143

4. Results and Discussion

4.1 Prediction Analysis

The performance of the ANN model was evaluated by the R^2 and RMSE values as shown in Table 2. A high R^2 value depicted a strong correlation between the actual and predicted SSC values of papaya samples. The prediction error, RMSE represented the deviation of value from the data point in regression line [16]. The pre-processing of the second derivative revealed a better performance than the zero derivative in terms of accuracy. The R^2 for training and testing sets of zero derivative (R^2_C=0.8886 and R^2_P=0.7365) show increment by using second derivatives (R^2_C =0.9063 and R^2_P =0.8768). Additionally, the RMSE for both training and testing of the zero derivative (RMSEC=0.5264 and RMSEP=0.9108) display a decrement when using second derivative (RMSEC=0.4406 and RMSEP=0.7047).

Table 2. Pre-processing result

Pre-processing	Training		Testing	
	R^2_C	RMSEC	R^2_P	RMSEP
Zero derivative	0.8886	0.5264	0.7365	0.9108
Second derivative	0.9063	0.4406	0.8768	0.7047

* R^2_C - R^2 for the training, RMSEC – Root Mean Squared Error for training,
R^2_P - R^2 for the testing, RMSEP – Root Mean Squared Error for testing

This indicates the model explained over 87% of the variance in the test dataset, affirming its predictive capability.

4.2 Regression Analysis

Regression plots (Figures 12 and 13) illustrate strong agreement between predicted and actual SSC values. The regression slope of the training set (0.84) and testing set (0.69) reflect slight underestimation, especially in lower SSC ranges. Nonetheless, the tight clustering around the ideal line demonstrates low variance and acceptable model bias.

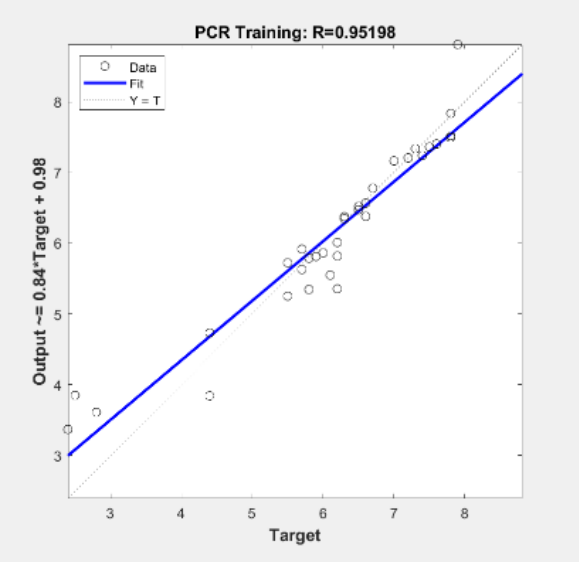

Figure 12. Regression model for training set

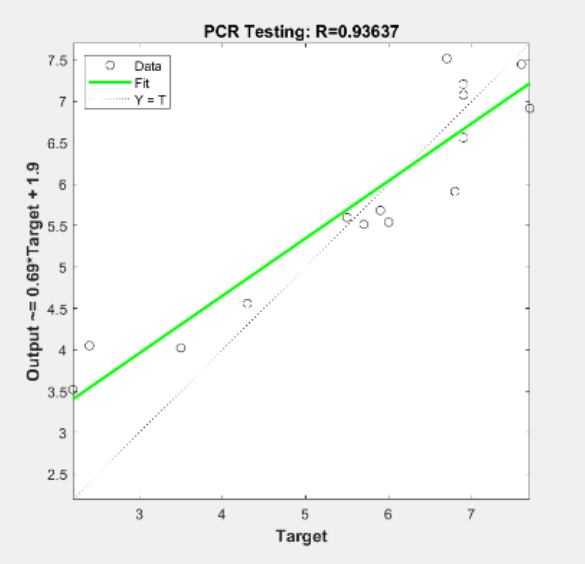

Figure 13. Regression model for testing set

5. Conclusion and Future Work

This study confirms the feasibility of using NIR spectroscopy in combination with ANN modeling for non-destructive SSC prediction in papaya. With appropriate pre-processing, the proposed method achieved high predictive accuracy and is suitable for potential integration into portable sorting systems. Further research should focus on real-time hardware implementation, optimization of network architecture, and field deployment across different environments and papaya cultivars.

Acknowledgment

The author would like to thank Universiti Teknologi Malaysia (UTM) for providing facilities for this research.

References

[1] Y. Huang et al., "Online detection of soluble solids content and maturity of tomatoes using Vis/NIR full transmittance spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 210, Mar. 2021, doi: 10.1016/j.chemolab.2021.104243.

[2] M. Wang, Y. Xu, Y. Yang, B. Mu, M. A. Nikitina, and X. Xiao, "Vis/NIR optical biosensors applications for fruit monitoring," Biosensors and Bioelectronics: X, vol. 11. *Elsevier Ltd*, Sep. 01, 2022. doi: 10.1016/j.biosx.2022.100197.

[3] S. Sohaib Ali Shah et al., "Towards fruit maturity estimation using NIR spectroscopy," *Infrared Phys Technol*, vol. 111, Dec. 2020, doi: 10.1016/j.infrared.2020.103479.

[4] B. Koul et al., "Carica papaya L.: A Tropical Fruit with Benefits beyond the Tropics," *Diversity*, vol. 14, no. 8. MDPI, Aug. 01, 2022. doi: 10.3390/d14080683.

[5] N. Sanchez, G. F. Gutiérrez-López, and G. Cáez-Ramírez, "Correlation among PME activity, viscoelastic, and structural parameters for Carica papaya edible tissue along ripening," *J Food Sci*, vol. 85, no. 6, pp. 1805–1814, Jun. 2020, doi: 10.1111/1750-3841.15130.

[6] A. Zeb et al., "Is this melon sweet? A quantitative classification for near-infrared spectroscopy," *Infrared Phys Technol*, vol. 114, May 2021, doi: 10.1016/j.infrared.2021.103645.

[7] T. Basile, A. D. Marsico, and R. Perniola, "Use of Artificial Neural Networks and NIR Spectroscopy for Non-Destructive Grape Texture Prediction," *Foods*, vol. 11, no. 3, Feb. 2022, doi: 10.3390/foods11030281.

[8] G. Zhang, X. Tuo, S. Zhai, X. Zhu, L. Luo, and X. Zeng, "Near-Infrared Spectral Characteristic Extraction and Qualitative Analysis Method for Complex Multi-Component Mixtures Based on TRPCA-SVM," *Sensors*, vol. 22, no. 4, Feb. 2022, doi: 10.3390/s22041654.

[9] Z. Zhou, I. Bar, R. Ford, H. Smyth, and C. Kanchana-Udomkan, "Biochemical, Sensory, and Molecular Evaluation of Flavour and Consumer Acceptability in Australian Papaya (Carica papaya L.) Varieties," *Int J Mol Sci*, vol. 23, no. 11, Jun. 2022, doi: 10.3390/ijms23116313.

[10] "Structure and Function of Carbohydrates | Biology for Majors I." https://courses.lumenlearning.com/wm-biology1/chapter/reading-types-of-carbohydrates/ (accessed Jul. 02, 2023).

[11] T. Basile, A. D. Marsico, and R. Perniola, "Use of Artificial Neural Networks and NIR Spectroscopy for Non-Destructive Grape Texture Prediction," *Foods*, vol. 11, no. 3, Feb. 2022, doi: 10.3390/foods11030281.

[12] M. Noguera, B. Millan, and J. M. Andújar, "New, Low-Cost, Hand-Held Multispectral Device for In-Field Fruit-Ripening Assessment," *Agriculture*, vol. 13, no. 1, p. 4, Dec. 2022, doi: 10.3390/agriculture13010004.

[13] "What Is A Brix Refractometer And How Does It Work? - Mega Depot." https://megadepot.com/resource/what-is-a-brix-refractometer-and-how-does-it-work (accessed Feb. 02, 2023).

[14] G. Zhang, X. Tuo, S. Zhai, X. Zhu, L. Luo, and X. Zeng, "Near-Infrared Spectral Characteristic Extraction and Qualitative Analysis Method for Complex Multi-Component Mixtures Based on TRPCA-SVM," *Sensors*, vol. 22, no. 4, Feb. 2022, doi: 10.3390/s22041654.

[15] V. R. Joseph, "Optimal ratio for data splitting," *Stat Anal Data Min*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.

[16] A. A. M. Ismail, N. Ali, S. Amirul, R. Endut, and S. A. Aljunid, "Prediction Model for Spectroscopy Using Python Programming," 2021.