Scientia Horticulturae 304 (2022) 111263

Contents lists available at ScienceDirect

Scientia Horticulturae

journal homepage: www.elsevier.com/locate/scihorti

Online measurement of carambola (*Averrhoa carambola* L.) physicochemical properties and estimation of maturity stages using a portable NIR spectrometer

Ingrid A. de Moraes, Luis J.P. Cruz-Tirado, Douglas Fernandes Barbin *

Department of Food Engineering, School of Food Engineering, University of Campinas, Campinas, SP, Brazil

ARTICLE INFO

Keywords:
Machine learning
Star fruit
Ripeness
Spectroscopy

ABSTRACT

Carambola is a tropical fruit with rising value in developed countries due to its nutritional value and exotic aspect. It is important to assess carambola quality in different maturity stages to estimate a "fair price" and to assign fruit for specific applications and markets. This work reported the use of a portable NIR spectrometer in the range of 900 to 1700 nm as a non-destructive, chemical-free technique for determination of carambola physicochemical properties, according to maturity stage. Colour, total soluble solids, ascorbic acid, moisture, pH and titratable acidity analysis were performed for 177 fruit from two clones and four maturity stages (MS1, MS2, MS3 and MS4). PLS-DA and PLSR models were built to classify carambola according to maturity stage and to estimate its physicochemical properties, respectively. Several pre-processing were tested and among them the new algorithm introduced in the (SNV) pre-processing, the variable sorting for normalization (VSN), allows the improvement of the signal shape and the model interpretation. Genetic algorithm (GA) and interval partial least square (iPLS) were tested for improving model performance. The PLS-DA model based on important variables selected by iPLS achieved the best performance with 84.2% accuracy to classify carambola according to maturity stage. Variable selection (iPLS and GA-PLS) allowed an improvement in the performance of the PLSR models, with pH and moisture content achieving (R^2_P of 0.78 and 0.74), (RMSEP of 0.2 and 0.87), (RPD of 2.01 and 2.23) and (RER of 8.02 and 10.38), respectively, which is acceptable for screening. Portable NIR spectrometer, which can be considered low-cost when compared to benchtop spectrometers, in tandem with chemometrics can be a promising tool to assess the composition and to classify carambola according to maturity stage.

List of abbreviations and symbols

GA	genetic algorithm
iPLS	interval partial least squares
LV	latent variables
MC	mean centering
PCA	principal component analysis
PLS-R	partial least squares regression
PLS-DA	partial least square discriminant analysis
RMSEC	root mean square error of calibration
RMSECV	root mean square error of cross-validation
RMSEP	root mean square error of prediction
RER	ratio error
RPD	ratio of standard deviation
R^2_C	coefficient of determination of calibration
R^2_{CV}	coefficient of determination of cross-validation
R^2_P	coefficient of determination of prediction
S-G	Savitzky–Golay
SNV	standard normal variate
VSN	variable sorting for normalization

1. Introduction

Carambola (*Averrhoa carambola L.*), also known as starfruit, is a rich source of minerals such as magnesium and potassium, and natural antioxidants such as ascorbic acid and carotenoids. Carambola is mainly sold as fresh fruit, but it has been consumed in different ways and directed by regional medicine for various treatments (Muthu et al., 2016). For this reason, it is important to assess carambola quality and physicochemical properties in different maturity stages to estimate a

* Corresponding author.
E-mail address: dfbarbin@unicamp.br (D.F. Barbin).

https://doi.org/10.1016/j.scienta.2022.111263
Received 9 February 2022; Received in revised form 27 April 2022; Accepted 9 June 2022
Available online 13 June 2022

"*fair price*" and to derivate fruit for specific applications (Yahaya and Omar, 2017).

Harvesting the fruit at the appropriate maturity stage is essential for the best quality and desired flavour. During the ripening phase, the fruit becomes soft, produces volatile compounds, decreases acidity and increases total soluble solids (TSS) (Yahaya and Omar, 2017). Fruits harvested in early stages contain less sugar and are prone to wilting and suffering mechanical damage. On the other hand, overripe fruits produce compounds with undesirable flavours, and tend to become soft, floury, and easily damaged right after harvest (Yahaya and Omar, 2017). Carambolas harvested in early maturity stages are destined for exportation due to their longer shelf life but have not yet developed full flavour and have limited use for garnishes, salads and drinks (FAMA, 2014). On the other hand, in advanced maturity stages, carambola can be marketed for jellies, puddings, pies and especially fresh, although they are directed only to deliveries to nearby destinations (FAMA, 2014; Pauziah et al., 2010). Therefore, it is important to know physicochemical properties of carambola at different maturity stages.

Currently, carambola quality inspection is based on colour distribution of the fruit, size and shape (FAMA, 2014; Stan, 1981). Fruit classification is carried out by visual inspection of trained evaluators, generating a great variability in the results, since this task is subjective and does not allow adequate standardization (Yahaya and Omar, 2017). On the other hand, even with a correlation between colour and physicochemical parameters related to flavour in the literature, skin colour does not predict the appropriate day of harvest, fruit quality or sweetness and acidity levels. Therefore, the quick and reliable prediction of physicochemical properties such as total soluble solids, titratable acidity, pH and ascorbic acid can improve the correct classification of carambola (Kyriacou and Rouphael, 2018).

Traditional analytical techniques used to assess physicochemical properties related to the quality of the fruit are destructive, slow and involve the use of chemicals (Alander et al., 2013).Fruit industries require fast, non-destructive and chemical free analytical techniques for implementing *on*- and *in*-line measurements. Near-Infrared Spectroscopy (NIRS) requires a minimum sample preparation, is chemical-free and allows quick characterization and measurement of several attributes in agricultural products, which has gained interest in recent years. In addition, according to Magwaza (2012), NIRS is more advanced compared to other non-destructive techniques in terms of instrumentation, accessories, availability of suitable chemometric software packages and even applications in the food sector. Recently, the reduced components size, such as LEDs and chips, allowed the development of handheld/portable NIR spectrometers. Portable NIR spectrometers have an ergonomic design, are transportable for different environmental work, and when well-calibrated, can be an alternative for benchtop devices (Pasquini, 2018). Portable NIR spectrometers have demonstrated good performance to estimate the quality of fresh tomatoes in the field (Borba et al., 2021), instantaneous and simultaneous prediction of anthocyanins and sugar in whole fresh raspberries (Gales et al., 2021), the internal quality of oranges, lemons and tangerines (Santos et al., 2021), prediction of some physicochemical parameters in pear (Mishra et al., 2021), and apple (Fan et al., 2020; Pourdarbani et al., 2020). However, to the best of our knowledge, there is no research available in the literature on the evaluation of the quality of carambola using a portable NIR spectrometer.

Therefore, the objective of this work is to develop a methodology based on a portable NIR spectrometer for online determination of physicochemical properties and the maturity stage of carambola.

2. Material and methods

2.1. Samples

Several batches of fruits were obtained from local commerce (CEASA supply station, Campinas-SP) on different dates, totalling 177 samples from B17 and B10 varieties. These are the most popular clones that are commercially grown: 'B17' are sweeter, less acidic and has better-eating quality as compared to 'B10', but they have a shorter shelf-life (Abd Rahman and Ahmad Hafiz, 2013). First, fruits were disinfected in a chlorine-based solution (0.2 g L^{-1}) and stored at 23°C until the time of analysis to standardize temperature. After colour analysis, samples were classified into 7 levels, according to the Federal Agricultural Marketing Authority (FAMA, 2014). However, since the colour change between the stages was very subtle, a new classification standard with 4 maturity stages (MS) was proposed. The fruits were classified based on the hue angle (h_{ab}) using the colorimeter (the details on the device are explained in paragraph 2.2) into MS1 (maturity stage 1, green, hue angle$>$ 100 °), MS2 (maturity stage 2, green/yellow, 92 ° $<$hue angle $\leq$ 100 °), MS3 (maturity stage 3, yellow, 83 ° $<$hue angle$\leq$ 92 °) and MS4 (maturity stage 4, yellow/orange hue angle, (hue angle $\leq$ 83 °) (Fig. 1).

2.2. Physical and physicochemical analyses

The physical and physicochemical parameters were measured after the acquisition of the spectra. Colour parameters were measured using a colorimeter (CM-2600D, illuminant D65 Konica Minolta Sensing Inc., Osaka, Japan), calibrated with a white ceramic standard, and the result was expressed in terms of lightness (L*), green-red (a*) and blue-yellow (b*) of the CIELAB system (CIE, 1976; McLAREN, 2008). In addition, the amount of Chroma (C^*_{ab}) (Eq. (1)) and hue angle (h_{ab}) (Eq. (2)) were calculated to compare the maturity stages (Theanjumpol et al., 2019).

$$C^*_{ab} = \sqrt[2]{a^{*2} + b^{*2}} \quad (1)$$

$$h_{ab} = \tan^{-1}\left(\frac{b}{a}\right) \quad (2)$$

Moisture content, total soluble solids content (TSS), ascorbic acid content (AA), pH and titratable acidity (TTA) were analysed. The moisture content of the samples was obtained by the gravimetric method after drying 4 g of sample in a vacuum oven at 60°C until constant weight (AOAC, 2006). The content of total soluble solids was determined using a manual refractometer model (KASVI, K52-032), on a scale from 0 to 32%, and calibrated with distilled water AOAC 932.12. The pH was determined using a pH meter (model MB-10; Marte, São Paulo, Brazil), titratable acidity and ascorbic acid were measured from the juice of the fruit (Nielsen, 2017).

2.3. Acquisition of NIR spectra

Acquisition of NIR spectra in the 902 - 1700 nm range, with 4 nm intervals, was performed on the whole fruit in absorbance mode, directly on the peel, using a portable spectrometer (DLPR NIRscanTM Nano, Texas Instruments, USA), with a 10 W halogen lamp, and NIRscan^{TM} Nano software. For each sample, three measurements (and ten replicates for the same measurement) were recorded along the equatorial region but at different positions in the fruit, to obtain a better representation of the samples.

2.4. Spectra pre-processing

The original data set was randomly divided into two subsets; 70% of the data (123) were used for calibration and the remaining 30% (54) for prediction (external test set). The spectra were mean-centered (MC), and then pre-processing methods were tested: smoothing Savitzky-Golay (S-G) utilizing a window size of 7 to 13 by performing a local polynomial regression (order 0) to remove noise. The first and second derivatives were used to correct the baseline; the derivatives were calculated using the Savitzky-Golay algorithm, with a window size of 7 to 15 data points. Multiplicative scatter correction (MSC), and the standard normal variate (SNV) were applied to correct the effects of light scattering (Barnes et al.,

Analyze	Range	Mean ± SD	Range	Mean ± SD	Range	Mean ± SD	Range	Mean ± SD
h_{ab}	100.14 - 107.12	103.10A ± 1.69	92.02 - 99.74	95.84B ± 2.49	83.03 - 91.87	88.24C ± 78.60	71.55 - 82.95	78.06D ± 2.87
C^*_{ab}	26.34 - 42.33	35.98B ± 3.84	24.88 - 42.89	32.99C ± 4.10	29.99 - 42.12	36.36B ± 2.52	33.30 - 44.12	38.58A ± 2.28

Fig. 1. Maturity stages for carambola according to hue angle (h_{ab}) value.

1989; Martens and Geladi, 1983). The variable sorting for normalization (VSN) was also tested. This novel algorithm, when introduced in SNV pre-processing, significantly improves signal shape and model interpretation. The method assumes that not all bands are equally altered by the unwanted effects and, consequently, it assigns a weight to each variable in the interval [0.1] corresponding to its probability of being affected only by the dispersion. This estimates the extent to which a wavelength is affected by size effects (additive and multiplicative displacements) rather than shape effects (resources attributable to chemically relevant contributions) (Rabatel et al., 2020).

2.4.1. Exploratory analysis

As an exploratory technique, principal component analysis (PCA) was performed to identify whether the spectra were affected by carambola maturity stages. PCA outputs were developed with the full wavelength range, using the singular value decomposition algorithm (SVD) (95% confidence level). Outlier's (anomalous spectra) were detected and eliminated using the amount of Q_residual and Hotelling T^2.

2.4.2. Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) is a commonly used technique for dimensionality reduction and classification problems. The mean and covariance for a Gaussian distribution are estimated for each class during classifier training. The model was developed by using the chemical physical parameters (pH, SST, TTA, moisture and AA) of 177 samples as predictors, divided into four maturity stages, and the performance of the classification models was expressed by Sensitivity (Eq. (3)), Specificity (Eq. (4)) and Accuracy (Eq. (5)) using Minitab software (MINITAB, LLC).

$$\text{Sensitivity } (\%) = \frac{TP}{(TP + FN)} \times 100 \quad (3)$$

$$\text{Specificity } (\%) = \frac{TN}{(TN + FP)} \times 100 \quad (4)$$

$$\text{Accuracy } (\%) = \frac{(TP + TN)}{TOTAL} \times 100 \quad (5)$$

$$\text{Error rate } (\%) = \frac{(FP + FN)}{TOTAL} \times 100 \quad (6)$$

Where, TP = true positive; FN = false negative; TN = true negative; and FP = false positive.

2.4.3. Classification of fruits according to maturity stage

Partial least squares discriminant analysis (PLS-DA) was applied to classify carambola according to its maturity stage considering four classes (MS1, MS2, MS3 and MS4). The number of latent variables (LV) for the PLS-DA models was selected using the lowest root mean squared error for cross-validation (RMSECV), which was performed using *ten subsets randomly constituted* (Barbin et al., 2013). This procedure was similar to partial least squares regression (PLSR). The performance of the classification models was expressed by Sensitivity (Eq. (3)), Specificity (Eq. (4)), Accuracy (Eq. (5)) and Error rate (Eq. (6)) and in the calibration, cross-validation and external validation data sets.

2.4.4. Prediction of physicochemical properties

Partial Least squares Regression (PLSR) was applied using NIR spectra as predictors for each characteristic analysed. The spectral region from 902 – 913 nm and 1649 – 1698 nm was removed to avoid noise. The performance of the regression models was assessed by the coefficient of determination (R^2), the number of latent variables (LV) and root mean squared error (RMSE) for calibration (RMSEC), cross-validation (RMSECV) and prediction (RMSEP), residual predictive deviation (RPD) and the range error ratio (RER) (Barbin et al., 2015). RPD is useful to assess the overall performance of prediction models, where RPD < 1.5 indicates unusable model, 1.5 < RPD < 2.0 states that the model can distinguish between high and low values, 2.0 < RPD < 2.5 suggests a model with approximate quantitative predictions possible, 2.5 < RPD <3.0 indicates a good prediction model and, RPD > 3, indicates an excellent predictive capacity of the model (Saeys et al., 2005). According to the American Cereal Chemicals Association (AACC), the PLSR models' performance also can be accessed by RER values: RER ≥ 4 is qualified for screening calibration when the RER ≥ 10 the model is acceptable for quality control, and if the RER ≥ 15 the model is very good for quantification (Rambo et al., 2013).

2.4.5. Variable selection

2.4.5.1. Interative partial least square (iPLS). The iPLS variable selection method is an extension developed for PLS, where a partial square linear regression is performed in each equidistant interval along the entire spectrum. The results are presented in the form of a graph, variance versus the number of latent variables, to facilitate comparison across the spectral range. In the application of the iPLS algorithm, the spectra were divided into equidistant intervals by automatic configuration with tested sub-windows from 1 to 60 (Xiaobo et al., 2010).

2.4.5.2. Genetic algorithm. The genetic algorithm (GA) is a random search and global optimization algorithm and is very effective in selecting the characteristic wavelength. The RMSECV was used as a model fitness scale, and windows were tested with sub-windows from 1 to 50, double crossover, the maximum number of generations of 100, the maximum number of latent variables of 20, and one iteration were used to obtain the wavelengths associated with the best correlation between spectra, and reference values (Rady and Guyer, 2015).

2.5. Statistical analysis

The physicochemical results for each class were compared using the analysis of variance (ANOVA). Subsequently, the Tukey multiple comparison test was performed ($p < 0.05$), in order to analyse the difference between the sample sets of each maturity stage. Multivariate statistical analysis was performed using PLS Toolbox (Eigenvector Research, Inc. Manson, WA, USA) was used in Matlab R2019a environment (Mathworks, Natick, USA).

3. Results and discussion

3.1. Physicochemical properties

The results showed relevant changes between the extreme maturity stages for all physicochemical properties ($p < 0.05$), which is shown in the values of pH, TSS, AA, TTA (Table 1) and colour (Fig. 1). As it was previously seen during carambola's ripening, the fruit becomes soft, produces volatile compounds, increases sugar content, and decreases the water content and acidity values (Yahaya and Omar, 2017). However, after the classification of maturity stages, it was observed samples with a similar value for hue angle but discrepant values of TSS, pH, and TTA. It was also seen that the physicochemical parameters, mainly the SST, did not show a high correlation with the colour parameters (Table 2). In addition, linear discriminant analysis (LDA) of physicochemical data at different maturation stages showed a low overall sensibility of 0.763 (Table 1, in supplementary material) compared to other classification works based on hue angle in other fruits (Ghazal et al., 2021). Thus, it was seen that not all samples presented a direct relationship between the hue angle and the physicochemical data related to the respective maturation stage. This evidences the spectroscopy's technique relevance because besides allowing the classification of the fruits based on their hue angle values, the technique also allows the prediction of physicochemical parameters. Thus, it is possible to identify ripped fruits with low sweetness and high acidity, and green fruits yet sweet.

3.2. Spectra profile and exploratory analysis

Fig. 2C shows the spectra profile and PCA analysis for carambola in different maturity stages (MS1-4). Fruit contains about 80–90% water, and the remaining fraction is composed of carbohydrates, organic acids, vitamins and minerals (Magwaza et al., 2012). The NIR spectra of samples, with a high percentage of water such as carambola (Fig. 2A), have a combination of overtones and fundamental vibrations, mainly attributed to the organic bonds (O - H, C - H and N - H), that make absorption bands relatively broad and complex (Magwaza et al., 2012). The spectral regions of 970 nm and 1450 nm associated to the second and first harmonic of O-H, respectively, showed differences in absorption for the classes (Magwaza et al., 2012), which are related to moisture content. The bands near 1445 nm and 1000 nm are related to the organic acids present in the carambola, such as oxalic and ascorbic acids. These regions correspond to the first and second overtone regions of O – H bonds, respectively and are difficult to see because they overlap with the bands assigned to water. The difference in the absorbance value of these regions may be associated with a drop in the levels of organic acids and an increase in the pH seen during the maturation process (Weyer and Lo, 2006). The absorption bands associated with second overtones (O–H) in 920 nm, third overtones (C-H) in 910 nm and the second connection (C - H) in 1190 nm are associated to sugars such as glucose and fructose in carambola (Magwaza et al., 2012). These regions showed differences in absorption for the stages and may be related to the increase in soluble solids contents from the stages MS1 to MS4. These bands are very close to the regions where the water has strong absorption, which makes it difficult to visualize and generates overlapping (Magwaza et al., 2012).

The principal component analysis (PCA) was performed using spectral region from 902 to 1637 nm. The spectral region from 1637 – 1698 nm was removed because it presented some noise. The raw spectral data was mean-centered (MC) and then pre-processed using smoothing S-G + 1st S-G derivative (Fig. 1, in supplementary material). Fig. 2C shows the score plot, where the first two principal components explained 81.46%

Table 1
Statistical results of the physicochemical properties of the carambola samples in four different maturity stages.

Class/ Analyse	MS1 Range	Mean ± SD	MS2 Range	Mean ± SD	MS3 Range	Mean ± SD	MS4 Range	Mean ± SD
AA (10^{-2} g Gallic acid L^{-1})	1.94 - 41.31	17.28D ± 9.71	2.80 - 53.85	21.63BC± 12.56	5.32 - 63.85	28.03B ± 15.48	10.38 - 75.77	37.38A ± 14.00
pH	2.51 - 3.08	2.80D ± 0.14	2.72 - 3.56	3.09C ± 0.27	2.85 - 4.15	3.40B ± 0.27	3.49 - 4.13	3.76A ± 0.16
TTA (% oxalic acid)	0.36 - 0.73	0.56C ± 0.10	0.26 - 0.65	0.40BC ± 0.08	0.22 - 0.58	0.37B ± 0.07	0.22 - 0.57	0.34A ± 0.08
TSS (%)	6.60 - 10.00	8.15D ± 0.84	7.59 - 14.60	9.51C ± 1.69	7.99 - 13.60	10.87B ± 1.57	8.60 - 15.89	11.65A ± 1.32
Moisture (%)	86.95 - 92.37	89.44A ± 1.23	83.89 - 91.90	88.81AB ± 2.27	83.57 - 91.99	88.17B ± 2.20	85.19 - 92.58	88.36B ± 1.32
L*	34.59 - 48.00	42.22A ± 3.66	33.94 - 58.38	43.60A ± 4.92	35.01 - 49.94	42.84A ± 4.14	29.31 - 40.60	36.90B ± 2.27
a*	(-11.27) - (-4.92)	-8.19D ± 1.57	(-6.84) - (-1.15)	-3.31C ± 1.41	(-1.12) - 4.52	1.16B ± 1.62	4.29 - 12.18	8.00A ± 1.99
b*	25.88 - 40.83	35.02B ± 3.66	24.64 - 42.67	32.79C ± 4.12	29.99 - 42.39	36.30AB ± 2.50	32.59 - 43.49	37.69A ± 2.21

* The data correspond to the mean ± SE of two repetitions. Different letters for the same parameter analysed indicate significant differences, between the maturity stages, by ANOVA ($P < 0.05$). (TSS) content of total soluble solids, (AA) content of ascorbic acid, (TTA) titratable acidity.

Table 2
Correlation of values of pH, TSS, AA, TTA and colour parameters.

	AA	pH	TTA	TSS	moisture	L*	a*	b*	C^*_{ab}	hue_{ab}
AA	1.000									
pH	0.320	1.000								
TTA	-0.222	-0.718	1.000							
TSS	0.663	0.571	-0.412	1.000						
moisture	-0.623	-0.050	0.022	-0.673	1.000					
L*	-0.002	-0.550	0.330	-0.057	-0.329	1.000				
a*	0.536	0.861	-0.682	0.701	-0.253	-0.445	1.000			
b*	0.399	0.286	0.043	0.457	-0.406	-0.019	0.324	1.000		
C^*_{ab}	0.391	0.270	0.080	0.423	-0.383	-0.066	0.309	0.992	1.000	
hue_{ab}	-0.543	-0.865	0.672	-0.716	0.276	0.428	-0.996	-0.373	-0.355	1.000

* Green values in modulus are >00 and ≤0.5; yellow values in modulus are >0.5 and <0.9 and blue values in modulus are ≥0.9 and <1.0.

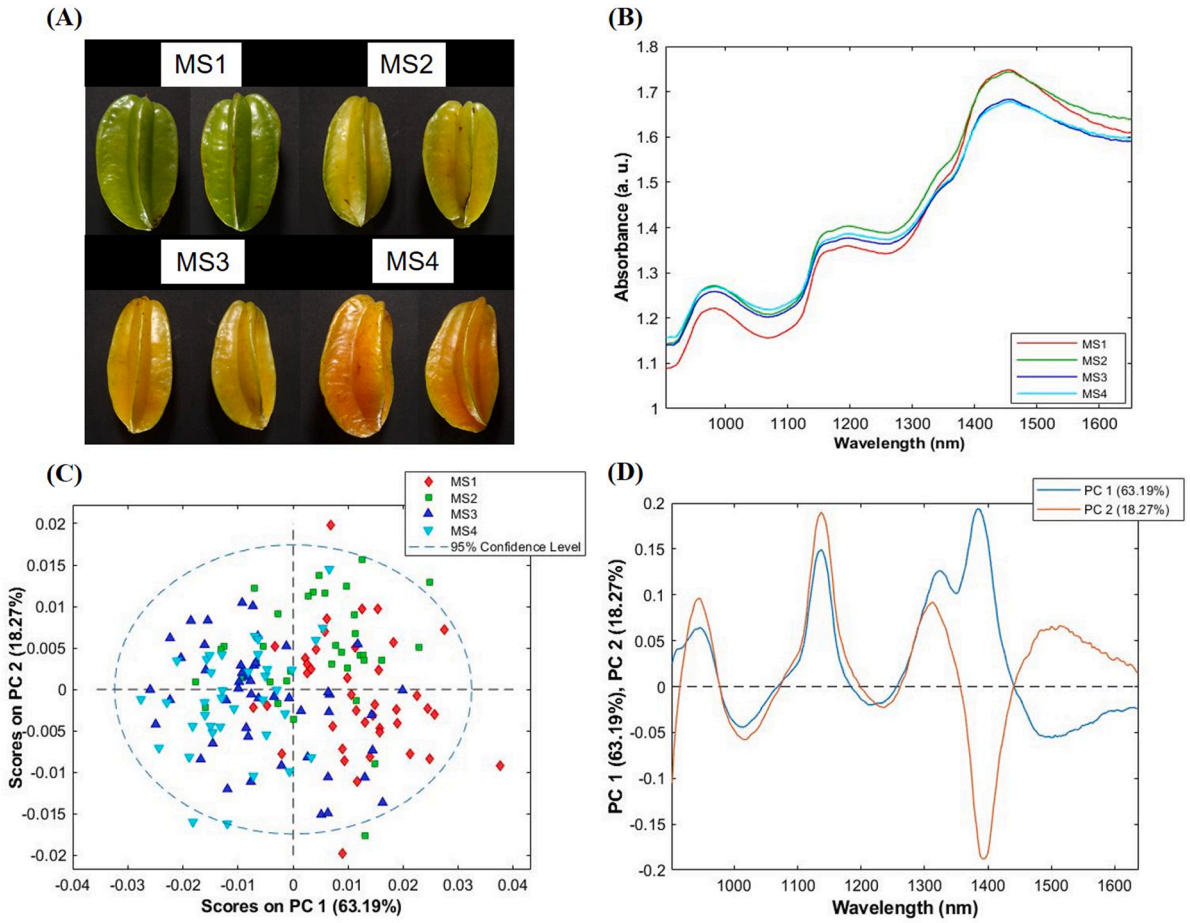

Fig. 2. A) Carambola in different maturity stages (MS1, MS2, MS3 and MS4), B) Mean spectra of carambola in different maturity stages, C) Scores plot based on 1st S-G derivative + smoothing S-G spectra (902–1637 nm), and D) Loadings plot of PCA analysis.

of variability for carambola in different maturity stages. The scores for the carambola moved to positive PC1 scores to negative PC1 scores, which is from MS1 stage (unripe) to MS4 stage (ripe). MS1 fruit is clearly located in the positive region of PC1 scores, while MS4 is located in the negative region of PC1 scores. Both MS2 and MS3 are overlapped, and they are in the center of the score plot, indicating the transition phase from MS1 stage to MS4 stage. The loadings plot in Fig. 2D showed that the peaks associated with water (970 and 1450 nm), organic acids (1000 and 1445 nm) and sugars (910, 920 and 1190 nm) are important for class discrimination.

3.3. Discrimination model

The confusion matrix of PLS-DA models for carambola classification according to maturity stages is shown in Fig. 3. Smoothing S-G with 11-point (order 0; window 11) + 1st S-G derivative (order 2; window 7) showed the best performance. The best PLS-DA model in the prediction set reached an accuracy of 97.22% for samples from MS1 (green), 75.15% for samples from MS2 (green/yellow), 74.03% for samples with from (MS3 (yellow, and 90.27% for samples from MS4 (yellow/orange) (Table 2, in supplementary material). These results were similar to those found in the LDA of the physical-chemical data (Table 1, in supplementary material). The accuracy obtained in the prediction set was similar to the accuracy for the calibration set, indicating that models are free from over and underfitting (Fig. 3). The loadings of the PLS-DA model (Fig. 2, in supplementary material) showed that regions related to the contents of water, acids and sugars are important for the classification of carambola. PLS-DA model was less sensitive and specific for

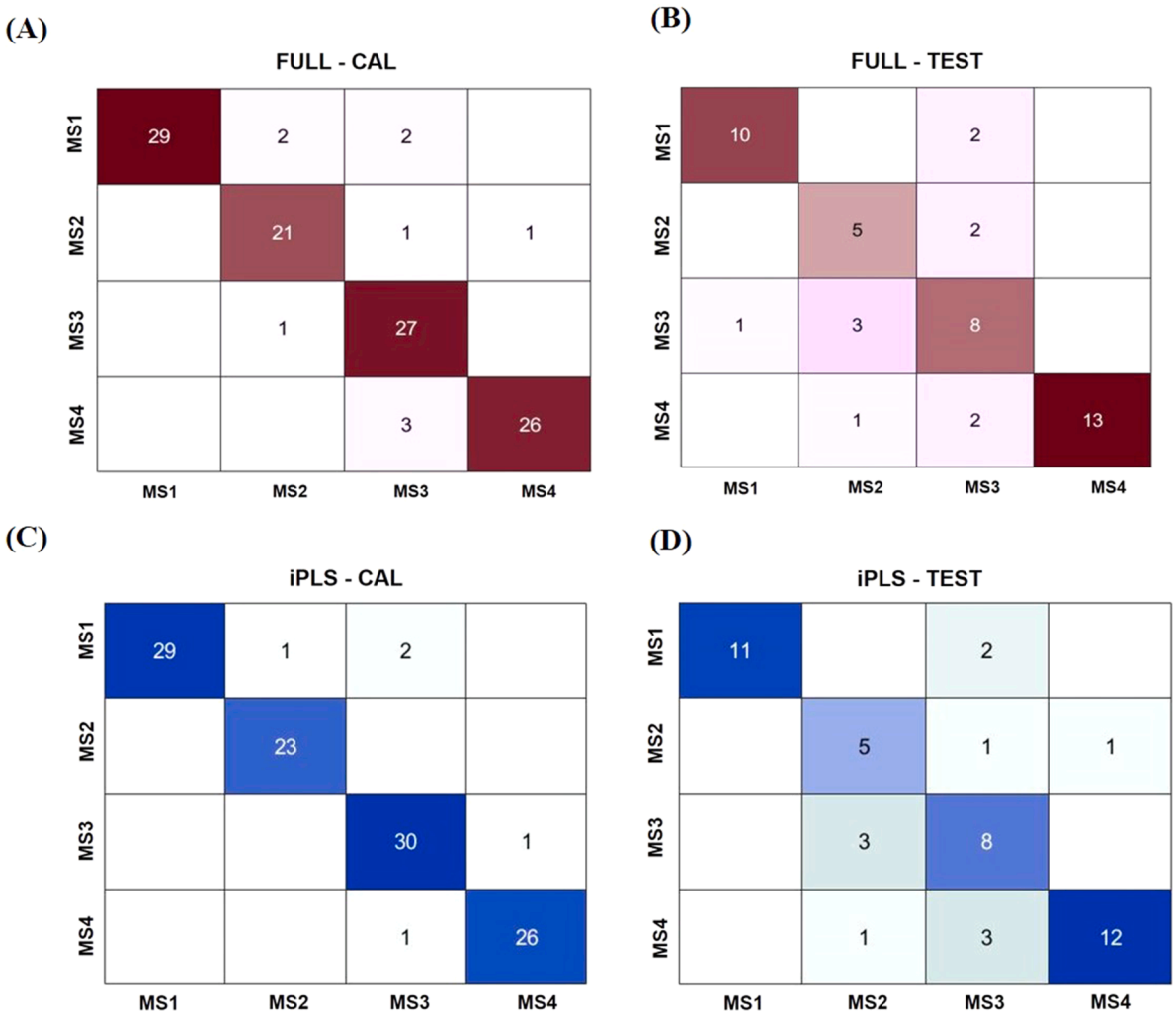

Fig. 3. Confusion matrix of PLS-DA models for calibration (CAL) set and validation (TEST) set using full spectra (A and B) and reduced spectra by iPLS (C and D).

MS2 and MS3 samples than MS1 and MS4 samples in both calibration and the prediction set (Table 2, in supplementary material). In addition, these intermediate stages had more samples incorrectly classified, which may be related to the smooth and gradual change of some components during the maturation process, such as moisture, TSS, TTA and AA (Table 1) (Yahaya and Omar, 2017). Based on the results, it is possible to assume that the discriminant models obtained are robust to discriminate between MS1 and MS4 classes, MS1 and MS2, and MS3 and MS4, but they have lower classification performance for MS2 and MS3 classes. Costa et al. (2019) classified grapes in three stages of ripeness and also obtained a lower performance of distinction in the stage of intermediate ripening, which was due to the smooth change of colour of grapes during its ripening process.

Variable selection was performed using both genetic algorithm (GA) and interval partial least square (iPLS). The GA did not improve the classification model. In contrast, iPLS algorithm selected the spectral regions from 905 to 1640 nm (eliminating 21 variables), which showed better performance than the full PLS-DA model, with an increase in the accuracy of 4.9% for MS1 and 2.0893 % in MS3 in the test set (Fig. 3).

3.4. Prediction of physicochemical properties by PLSR

PLSR regression was used to model the relationships between the NIR spectra and the physicochemical properties of carambola (TSS, TTA, moisture, pH and AA) (Fig. 4). The spectral region from 902 – 913 nm and 1649 – 1698 nm was removed to avoid noise. In addition, Fig. 4 shows the best PLSR model for each physicochemical measured parameter. The variable selection improved PLSR model, both by iPLS (in TSS and pH) or GA (in% moisture, TTA and AA) (Table 3). Magwaza et al. (2012) have argued that models involving a few carefully selected wavelengths were effective for the prediction of chemical composition in citrus fruit. The iPLS algorithm seeks a particularly informative spectral range in relation to the analysed parameter to generate simple models and increase its predictive capacity (Nørgaard et al., 2000). The comparison between the selected range and the full spectrum models helps to interpret the relationship between the model and the sample compositions. For its part, GA algorithms also have a great ability to improve results, as it is a random search algorithm and guarantees the model's reliability; the calculation procedure is repeated many times until the best variables are found. This decreases the values RMSEP and RMSECV without reducing prediction capacity (Whitley, 1994).

The R^2 (Cal and Test set) for pH and moisture prediction models were superior to TTA, TSS and AA parameters, as shown in Table 1. Previous works reported similar R^2 values of 0.656 for TTA and 0.808 for pH in carambola using Vis/NIR spectroscopy (Yahaya and Omar, 2017), and R^2= 0.58 for TTA in strawberry (Amodio et al., 2017). This could be related to the short-range of values for some physicochemical properties such as TTA, and low concentrations, such as i.e. ascorbic acid.

The best PLS model, after the variables selection, for TSS (104 selected variables), TTA (120 selected variables) and AA (145 selected variables) showed 1.5 < RPD < 2 values, which suggests that the model can distinguish between high and low values (Saeys et al., 2005), and 4

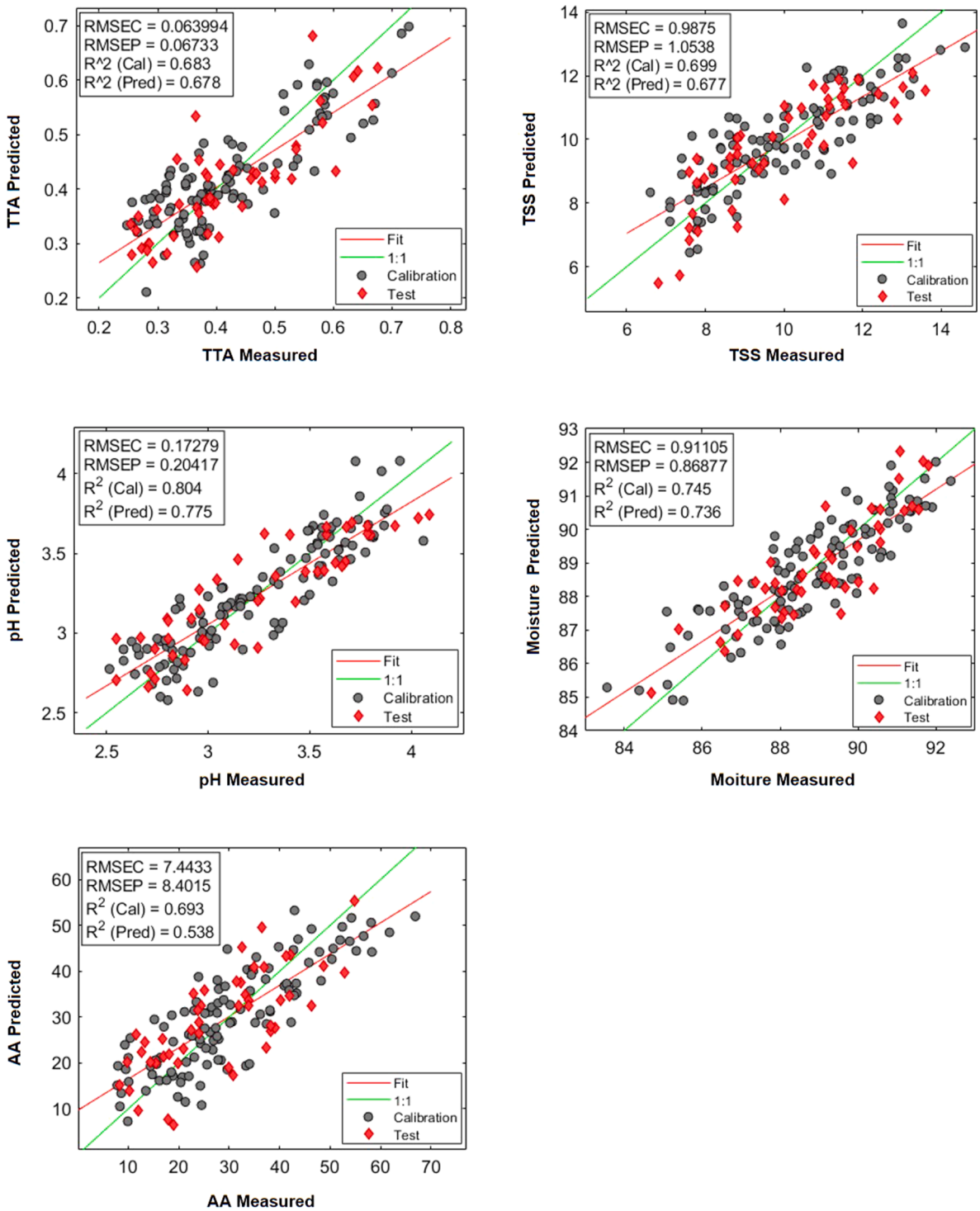

Fig. 4. PLSR models with the best performance for total titratable acidity (TTA), total of soluble solids (TSS), pH, moisture and ascorbic acid (AA).

< RER < 10, which it represents a model for screening calibration (Rambo et al., 2013). According to Nicolaï et al. (2007) the prediction of TTA using NIRS is more difficult since organic acids tend to be found in relatively lower concentrations, near the limit of detection of the method. However, RPD values in this work for TTA prediction were superior to those reported for strawberries (RPD = 1.33) (Amodio et al., 2017). On the other hand, Yahaya and Omar (2017) reported similar values $R^2_P = 0.656$ for the prediction of TSS in carambola with Vis/NIR spectroscopy. In addition, it was reported better prediction models for AA in apples (RPD = 2) (Pissard et al., 2013) TTA in purple passion fruit (RPD = 1.8) (Maniwara et al., 2019).

Prediction models with selected variables for pH (81 selected variables) and moisture (133 selected variables) showed RPD > 2, which suggests a model with approximate quantitative predictions possible (Saeys et al., 2005). The pH model reached RER of 8.02, showing that the model can be used for screening. Models for moisture prediction presented RER = 10.38, suggesting that the model is acceptable for quality control (Rambo et al., 2013). According to Walsh et al. (2020), the water absorption bands dominate the spectrum of fruit, which probably influences the results for moisture prediction. Similar results

Table 3
Statistics for the calibration and prediction sets for physical and physicochemical analysis in carambola using portable NIR and PLSR, iPLS and GA-PLS.

Physicochemical parameter	Pre-processing	Model	Number of variables	LV	R^2C	RMSEC	R^2CV	RMSECV	R^2P	RMSEP	RPD	RER
AA	Smoothing S-G (order 0; window 15) + 1st S-G derivative (order 2; window 13)	Full	208	9	0.67	7.72	0.52	9.43	0.53	8.47	1.85	8.72
		iPLS	175	9	0.52	7.54	0.52	9.42	0.54	8.42	1.86	8.77
		GA	145	12	0.69	7.44	0.50	9.68	0.54	8.40	1.87	8.79
pH	Smoothing S-G (order 0; window 11) + 1st S-G derivative (order 2; window 11)	none	208	9	0.82	0.17	0.71	0.21	0.74	0.21	1.91	7.63
		iPLS	81	9	0.80	0.17	0.73	0.20	0.78	0.20	2.01	8.02
		GA	173	11	0.81	0.17	0.66	0.23	0.76	0.21	1.95	7.76
TTA	Smoothing S-G (order 0; window 9)	Full	222	9	0.69	0.06	0.56	0.08	0.66	0.07	1.79	7.55
		iPLS	74	9	0.70	0.06	0.57	0.08	0.68	0.07	1.81	7.66
		GA	120	9	0.70	0.06	0.54	0.08	0.69	0.06	1.84	7.80
TSS	Smoothing S-G (order 0; window 13) + 1st S-G derivative (order 2; window 13)	none	208	7	0.69	1.00	0.61	1.12	0.64	1.12	1.75	8.33
		iPLS	104	7	0.70	0.99	0.63	1.10	0.68	1.05	1.85	8.81
		GA	18	7	0.71	0.97	0.66	1.05	0.66	1.08	1.80	8.61
Moisture	Smoothing S-G (order 0; window 13) + 1st S-G derivative (order 2; window 9)	none	208	9	0.78	0.85	0.68	1.04	0.72	0.90	2.14	9.98
		iPLS	56	9	0.75	0.90	0.68	1.02	0.74	0.89	2.18	10.16
		GA	133	9	0.74	0.91	0.63	1.11	0.74	0.87	2.23	10.38

were reported for pH prediction Yahaya and Omar (2017) with RMSEP = 0.188 for carambola using VIS / NIR spectroscopy.

Carambola contains a high content of water and other chemicals, such as vitamins and soluble solids, and these chemical and physical parameters can be affected by some external factors, such as temperature and relative humidity at the time of harvest, which affects the accuracy and robustness of prediction (Walsh et al., 2020). Previous works reported that the growth conditions, the size of the fruit cells and the tissue structure, in some cases translucent as a star fruit (Fig. 3, supplementary material), also could interfere with the performance of the model in NIR. This is because they are less predictable interferences related to the transmission of light at the time of acquisition of the spectra (Xie et al., 2021). On the other hand, the use of more than one variety, as used here, can assign robustness to the prediction model in NIR (Walsh et al., 2020; Xie et al., 2021).

The need for alternative methods, which do not generate large amounts of chemical residues and /or costly analyses, makes the NIR method an adequate alternative. The determination of properties through predictive models capable of distinguishing between high and low values, in this study, would economize long periods for traditional analyses. Research on some fruit has generated great advances; The Australian industry has adopted the use of portable NIRS to evaluate fruit in the field, to aid in the decision making during harvest and post-harvest classification (Walsh et al., 2020). Therefore, research on the prediction of physicochemical properties and classification according to maturity stage of carambola using portable NIR and chemometrics should be addressed and standardized to help fruit producers.

4. Conclusion

It is essential to know the maturity stage and physicochemical properties of carambola to supply high-quality fruits to customers. During ripening, the content of total soluble solids, ascorbic acid and pH increased and the content of titratable acidity and moisture content decreased. The PLS-DA method showed that the fruit spectra contains information that allows correct classification of carambola according to their maturity stage, with an overall accuracy of 97.1% in the calibration set and 84.2% in the test set. For physicochemical properties, the variable selection by iPLS and GA-PLS showed improvements in the performance of the prediction models. Portable NIR-based models allowed the prediction of the content of TSS, AA, TAA, pH and moisture content, which can be used for online analysis in the industry for carambola quality control.

CRediT authorship contribution statement

Ingrid A. de Moraes: Software, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Visualization. **Luis J.P. Cruz-Tirado:** Software, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Visualization. **Douglas Fernandes Barbin:** Methodology, Investigation, Writing – review & editing, Resources, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This study was financed in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001; and São Paulo Research Foundation (FAPESP) (project number 2015/24351-2). J. P. Cruz-Tirado acknowledges scholarship funding from FAPESP, grant number 2020/09198-1. Ingrid Alves de Moraes acknowledges scholarship funding from FAPESP, grant number 2019/12625-1. Prof. Douglas Fernandes Barbin is CNPq research fellow (308260/2021-0).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.scienta.2022.111263.

References

Abd Rahman, M., Ahmad Hafiz, B., 2013. Genetic improvement of fruit quality traits in starfruit (*Averrhoa carambola*) hybrids. Acta Hortic. 1012, 259–264. https://doi.org/10.17660/ActaHortic.2013.1012.30.

Alander, J.T., Bochko, V., Martinkauppi, B., Saranwong, S., Mantere, T., 2013. A review of optical nondestructive visual and near-infrared methods for food quality and safety. Int. J. Spectrosc. 1–36. https://doi.org/10.1155/2013/341402, 2013.

Amodio, M.L., Ceglie, F., Chaudhry, M.M.A., Piazzolla, F., Colelli, G., 2017. Potential of NIR spectroscopy for predicting internal quality and discriminating among strawberry fruits from different production systems. Postharvest Biol. Technol. 125, 112–121. https://doi.org/10.1016/j.postharvbio.2016.11.013.

AOAC, 2006. Official Methods of Analysis of AOAC International, 18th ed. Maryland.

Barbin, D.F., ElMasry, G., Sun, D.-W., Allen, P., 2013. Non-destructive determination of chemical composition in intact and minced pork using near-infrared hyperspectral imaging. Food Chem. 138, 1162–1171. https://doi.org/10.1016/j.foodchem.2012.11.120.

Barbin, D.F., Kaminishikawahara, C.M., Soares, A.L., Mizubuti, I.Y., Grespan, M., Shimokomaki, M., Hirooka, E.Y., 2015. Prediction of chicken quality attributes by near infrared spectroscopy. Food Chem. 168, 554–560. https://doi.org/10.1016/j.foodchem.2014.07.101.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43, 772–777. https://doi.org/10.1366/0003702894202201.

Borba, K.R., Aykas, D.P., Milani, M.I., Colnago, L.A., Ferreira, M.D., Rodriguez-Saona, L. E., 2021. Portable near infrared spectroscopy as a tool for fresh tomato quality control analysis in the field. Appl. Sci. 11 https://doi.org/10.3390/app11073209.

CIE, (Commission International De l' Eclairage), 1976. Colorimetry. Vienna, Switzerland.

Costa, D., dos, S., Mesa, N.F.O., Freire, M.S., Ramos, R.P., Mederos, B.J.T., 2019. Development of predictive models for quality and maturation stage attributes of wine grapes using vis-nir reflectance spectroscopy. Postharvest Biol. Technol. 150, 166–178. https://doi.org/10.1016/j.postharvbio.2018.12.010.

FAMA, 2014. Ministry of Agriculture and Food Industry [WWW Document]. SIRI Pandu. Kual. CARAMBOLA.

Fan, S., Wang, Q., Tian, X., Yang, G., Xia, Y., Li, J., Huang, W., 2020. Non-destructive evaluation of soluble solids content of apples using a developed portable Vis/NIR device. Biosyst. Eng. 193, 138–148. https://doi.org/10.1016/j.biosystemseng.2020.02.017.

Gales, O., Rodemann, T., Jones, J., Swarts, N., 2021. Application of near <scp>infra-red</scp>spectroscopy as an instantaneous and simultaneous prediction tool for anthocyanins and sugar in whole fresh raspberry. J. Sci. Food Agric. 101, 2449–2454. https://doi.org/10.1002/jsfa.10869.

Ghazal, S., Qureshi, W.S., Khan, U.S., Iqbal, J., Rashid, N., Tiwana, M.I., 2021. Analysis of visual features and classifiers for Fruit classification problem. Comput. Electron. Agric. 187, 106267 https://doi.org/10.1016/j.compag.2021.106267.

Kyriacou, M.C., Rouphael, Y., 2018. Towards a new definition of quality for fresh fruits and vegetables. Sci. Hortic. (Amsterdam). 234, 463–469. https://doi.org/10.1016/j.scienta.2017.09.046.

Magwaza, L.S., Opara, U.L., Nieuwoudt, H., Cronje, P.J.R., Saeys, W., 2012. NIR spectroscopy applications for internal and external quality analysis of citrus fruit- a review. Food Bioprocess Technol. 425–444 https://doi.org/10.1007/s11947-011-0697-1.

Maniwara, P., Nakano, K., Ohashi, S., Boonyakiat, D., Seehanam, P., Theanjumpol, P., Poonlarp, P., 2019. Evaluation of NIRS as non-destructive test to evaluate quality traits of purple passion fruit. Sci. Hortic. (Amsterdam). 257, 108712 https://doi.org/10.1016/j.scienta.2019.108712.

Martens, J.H.S., Geladi, P., 1983. Multivariate linearity transformation for near-infrared reflectance spectrometry. Proc. Nord. Syrup. Appl. Stat. 205–234.

McLAREN, K., 2008. XIII-the development of the CIE 1976 (L* a* b*) uniform colour space and colour-difference formula. J. Soc. Dye Colour 92, 338–341. https://doi.org/10.1111/j.1478-4408.1976.tb03301.x.

Mishra, P., Marini, F., Brouwer, B., Roger, J.M., Biancolillo, A., Woltering, E., Echtelt, E. H., 2021. Sequential fusion of information from two portable spectrometers for improved prediction of moisture and soluble solids content in pear fruit. Talanta 223, 121733. https://doi.org/10.1016/j.talanta.2020.121733.

Muthu, N., Lee, S.Y., Phua, K.K., Bhore, S.J., 2016. Nutritional, medicinal and toxicological attributes of star-fruits (*Averrhoa carambola* L.): a review. Bioinformation 12, 420–424. https://doi.org/10.6026/97320630012420.

Nicolaï, B.M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K.I., Lammertyn, J., 2007. Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: a review. Postharvest Biol. Technol. 46, 99–118. https://doi.org/10.1016/j.postharvbio.2007.06.024.

Nielsen, S.S., 2017. Food Analysis Laboratory Manual, Food Science Text Series. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-44127-6.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval partial least-squares regression (i PLS): a comparative chemometric study with an example from near-infrared spectroscopy. Appl. Spectrosc. 54, 413–419. https://doi.org/10.1366/0003702001949500.

Pasquini, C., 2018. Near infrared spectroscopy: a mature analytical technique with new perspectives – a review. Anal. Chim. Acta 1026, 8–36. https://doi.org/10.1016/j.aca.2018.04.004.

Pauziah, M., Tarmizi, S.A., Mohd Salleh, P., Norhayati, M., 2010. Quality of starfruit harvested at advanced maturity stage. Acta Hortic. 880, 231–235. https://doi.org/10.17660/ActaHortic.2010.880.26.

Pissard, A., Fernández Pierna, J.A., Baeten, V., Sinnaeve, G., Lognay, G., Mouteau, A., Dupont, P., Rondia, A., Lateur, M., 2013. Non-destructive measurement of vitamin C, total polyphenol and sugar content in apples using near-infrared spectroscopy. J. Sci. Food Agric. 93, 238–244. https://doi.org/10.1002/jsfa.5779.

Pourdarbani, R., Sabzi, S., Kalantari, D., Karimzadeh, R., Ilbeygi, E., Arribas, J.I., 2020. Automatic non-destructive video estimation of maturation levels in Fuji apple (Malus Malus pumila) fruit in orchard based on colour (Vis) and spectral (NIR) data. Biosyst. Eng. 195, 136–151. https://doi.org/10.1016/j.biosystemseng.2020.04.015.

Rabatel, G., Marini, F., Walczak, B., Roger, J., 2020. VSN: variable sorting for normalization. J. Chemom. 34, e3164.

Rady, A.M., Guyer, D.E., 2015. Evaluation of sugar content in potatoes using NIR reflectance and wavelength selection techniques. Postharvest Biol. Technol. 103, 17–26. https://doi.org/10.1016/j.postharvbio.2015.02.012.

Rambo, M.K.D., Amorim, E.P., Ferreira, M.M.C., 2013. Potential of visible-near infrared spectroscopy combined with chemometrics for analysis of some constituents of coffee and banana residues. Anal. Chim. Acta 775, 41–49. https://doi.org/10.1016/j.aca.2013.03.015.

Saeys, W., Mouazen, A.M., Ramon, H., 2005. Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy. Biosyst. Eng. 91, 393–402. https://doi.org/10.1016/j.biosystemseng.2005.05.001.

Santos, C.S.P., Cruz, R., Gonçalves, D.B., Queirós, R., Bloore, M., Kovács, Z., Hoffmann, I., Casal, S., 2021. Non-destructive measurement of the internal quality of citrus fruits using a portable NIR device. J. AOAC Int. 104, 61–67. https://doi.org/10.1093/jaoacint/qsaa115.

Stan, C., 1981. CODEX STAN 82 Page 1 of 4 1–4.

Theanjumpol, P., Wongzeewasakun, K., Muenmanee, N., Wongsaipun, S., Krongchai, C., Changrue, V., Boonyakiat, D., Kittiwachana, S., 2019. Non-destructive identification and estimation of granulation in 'Sai Num Pung' tangerine fruit using near infrared spectroscopy and chemometrics. Postharvest Biol. Technol. 153, 13–20. https://doi.org/10.1016/j.postharvbio.2019.03.009.

Walsh, K.B., McGlone, V.A., Han, D.H., 2020. The uses of near infra-red spectroscopy in postharvest decision support: a review. Postharvest Biol. Technol. 163, 111139 https://doi.org/10.1016/j.postharvbio.2020.111139.

Weyer, L.G., Lo, S.-C., 2006. Spectra- structure correlations in the near-infrared. In: Griffiths, P.R. (Ed.), Handbook of Vibrational Spectroscopy. John Wiley & Sons, Ltd, Chichester, UK, pp. 140–141. https://doi.org/10.1002/0470027320.s4102.

Whitley, D., 1994. A genetic algorithm tutorial. Stat. Comput. 4 https://doi.org/10.1007/BF00175354.

Xiaobo, Z., Jiewen, Z., Povey, M.J.W., Holmes, M., Hanpin, M., 2010. Variables selection methods in near-infrared spectroscopy. Anal. Chim. Acta 667, 14–32. https://doi.org/10.1016/j.aca.2010.03.048.

Xie, D., Liu, D., Guo, W., 2021. Relationship of the optical properties with soluble solids content and moisture content of strawberry during ripening. Postharvest Biol. Technol. 179, 111569 https://doi.org/10.1016/j.postharvbio.2021.111569.

Yahaya, O.K.M., Omar, A.F., 2017. Non-spectroscopic Techniques for the Assessment of Quality Attributes. In: Nurolaini, N., Isa, N.M. (Eds.), SPECTROSCOPY OF TROPICAL FRUITS Sala Mango and B10 Carambola. PENERBIT UNIVERSITI SAINS MALAYSIA (Universiti Sains Malaysia, Penerbit Universiti Sai, pp. 35–38.