

# Transformations for non-destructive evaluation of brix in mango by reflectance spectroscopy and machine learning

Ernesto Paiva-Peredo, Diego Gonzales-Rodriguez, William Trujillo Herrera,

Juan Jesús Soria Quijaite, Diana Quispe-Arpasi, Christian Ovalle Paulino

Faculty of Systems and Electronics Engineering, Universidad Tecnológica del Perú, Lima, Peru

## Article Info

### Article history:

Received May 10, 2023

Revised Jun 28, 2023

Accepted Jul 3, 2023

### Keywords:

Brix

Machine learning

Partial least squares

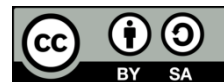
Principal component analysis

Spectroscopy

## ABSTRACT

Mango is a very popular climacteric fruit in America and Europe. Among the internal properties of the mango, total soluble solids (TSS) are an adequate indicator to estimate the quality of mango, however, the measurement of this indicator requires destructive tests. Several research have addressed similar issues; they have made use of pre-processing transformations without making it clear which of them is statistically better. Here, we created a new spectral database to build machine learning (ML) models. We analyzed a total of 18 principal component regression (PCR) models and 18 partial least squared regression (PLSR) models, where 4 types of transformations, 3 different feature extractors, and 3 different pre-processing techniques are combined. The research proposes a double cross validation (CV) both to determine the optimal number of components and to obtain the final metrics. The best model had a root mean square error (RMSE) of 1.1382 °Brix and a RMSE on the transformed scale of 0.5140. The best model used 4 components, used  $y^2$  transformation, reflectance R as the independent variable and MSC as a pre-processing technique.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Ernesto Paiva Peredo

Faculty of Systems and Electronics Engineering, Universidad Tecnológica del Perú

125 Natalio Sánchez Street, Santa Beatriz, Cercado de Lima, Lima, Peru

Email: epaiva@utp.edu.pe

## 1. INTRODUCTION

Mango is a very nutritious fruit and one of the most consumed in the world, because it contains nutrients such as proteins, carbohydrates, fats, minerals and vitamins [1], [2]. World exports of mango, mangosteen and guava have exceeded 2.3 million tons; Mexico, Thailand, Brazil, Peru and India are the largest exporters, with Europe as the main export destination [3]. Peruvian mangoes stand out due to their high quality, which is why they are increasingly in demand. The mango varieties cultivated in Peru are Haden, Edward, Keitt and Kent; being the north of Peru, exactly the city of Piura the largest producer of this fruit, followed by Lambayeque and Ancash, achieving the country to export more than 240,000 tons of mango during 2021-2022 campaign [4].

In many countries, traditional methods for estimating nutrients in mango plants are performed at stages when the mango is still growing, these techniques are invasive and time consuming. In addition, these analytical methods used for determination of total soluble solids (TSS) are destructive and are not suitable for remotely monitoring fruit quality in modern grading lines [5]. An alternative method presents more attractive approaches for the measurement of characteristics that correlate with fruit maturity where near infrared (NIR) spectroscopy technology is used, since it represents a high-speed alternative for data acquisition and with the help of calibration techniques the correct prediction can be achieved [5], [6].

Because of this, the present study focuses on using NIR spectroscopy to predict the TSS or °Brix as an indicator of the degree of maturity of Edwards mangoes using Machine Learning techniques. This aims to reduce the time in determining the internal characteristics of the products, especially without having to destroy or alter the macro and micronutrients of the fruit. Our research is focused on determining the transformations, pre-processing technique, input characteristics and dimensional reduction algorithm best suited for the creation of these models.

Initially, studies are presented that highlight the importance of using NIR spectroscopy to measure fruit quality indicators. A research work that evaluates the ability of this technology to determine the chemical composition, such as TSS and titratable acidity (TA) in 3 different fruits, shows encouraging results, affirming that this technology has a favorable use in the prediction of internal content in intact fruits. However, it is highlighted that in species such as passion fruit and tomato, the results may be affected by their physical structure [7].

In addition, a Fourier transform NIR spectroscopy study is presented as a viable option for the rapid localization of diseases in fruit [8]. In the same way, a portable spectrometer is used to monitor the evolution of the firmness of the Kent mango in a non-destructive way, evaluating its physical and chemical properties during the ripening stage, obtaining results that demonstrate the relationship between changes in the firmness of the fruit and its spectral signature [9]. Some other examples are observed in the work on the measurement of SSC, TA, vitamin C and surface color of mandarins [10]; dry matter (DM) content, potential of hydrogen (pH), TSS and acid-Brix ratio (ABR) for bananas [11]; firmness, starch index and total soluble solids content in apples [12]. It is evident that near infrared spectroscopy is a fast and non-destructive method to evidence fruit quality; therefore, it saves labor hours, manpower and improves accuracy in grading these fruits. Therefore, being able to study changes in these parameters will allow producing quality fruits and harvesting them at the right time, either for immediate consumption or to withstand export time [13].

Regarding the spectrum range, in [7] a multipurpose analyzer spectrometer was used in reflectance mode with a spectrum range from 800 to 2,700 nm with a spectral resolution of 2 nm. The general shapes of the spectra of tomato, apricot and passion fruit were quite similar, it is worth noting that the spectrum of passion fruit had a slight shift, this similarity is mainly due to the fact that these fruits contain between 80% and 90% water. Of equal importance, in [14] spectral acquisition was performed with a range of 300 to 1,000 nm, the analysis of the NIR spectrum showed that the curves were different between green and ripe mango fruits in the range of 550 to 700 nm, above 750 nm a relatively flat area was observed with different reflectance values between the two types. In the NIR range 700 to 990 the best model for predicting DM, TSS and maturity was obtained, without using any processing technique to the spectral data. Another study [13] pointed out that the use of the entire spectral range is detrimental to the calculation, since these are usually of large volume and the purchasing cost of spectral reading devices increases proportionally to the total range. For these reasons, the most relevant lengths are analyzed with artificial neural network-simulated annealing algorithm (ANN-SA). The manuscript worked on the estimation of the TSS and BrimA of the Gala apple, which determined that for TSS the wavelengths 953, 961, 977 and 983 nm are the most effective and for BrimA are 958, 966, 972 and 984 nm [13]. Likewise, in 2020, a study used an algorithm based on an artificial neural network-differential evolutionary (ANN-DE) that determined the effective wavelengths for 3 properties (firmness, acidity and starch content) of Fuji apple. This showed that in the range of 841 to 882 nm each property is predicted with a higher regression coefficient compared to working with the full wavelength range of 400 to 1,000 nm, this work demonstrates that it is possible to develop a lower cost portable device working in the optimal range and being equally efficient [15]. Similarly important, in [9] a portable visible near infrared spectrometer was used in the 310 to 1,130 nm range at a spectral resolution of 8 to 13 nm, the measurement was performed at 2 different points for a sample in order to predict the firmness of Kent mangoes at different stages of their ripening process. In this work, the interval partial least squares regression (iPLSR) algorithm was used for the selection of spectral regions of greatest relevance. This algorithm selected the range of 743 to 770 nm and 870 to 905 nm, which contribute to the prediction of firmness, in comparison to the use of the whole spectrum there is a 14% improvement in the prediction error.

On the other hand, methods based on the spectral signature analysis of different fruits are presented. A partial least squares regression (PLSR) modeling study was implemented in Brazil, using NIR spectroscopy to estimate firmness and TSS concentration in peach fruits. Spectrums were obtained as log 1/R and model performance was evaluated as a function of root mean square error (RMSE) and coefficient of determination ( $R^2$ ) values. Principal component analysis (PCA) failed to group fruit according to blush and skin background color, maturity stages and harvest season. However, NIR spectroscopy with partial least squares (PLS) proved to be a potential analytical method for determining TSS and firmness of the cultivar 'Aurora 1' [5]. In contrast, in another paper, TA, TSS and pulp content (PC) of passion fruit were predicted by NIR spectroscopy using the PLSR prediction model. Encouraging results were obtained for TA and PC with correlation coefficients of 0.91 and 0.99 respectively, however, only a value of 0.84 was obtained for TSS [16]. Then, in Cebu University of Technology - Philippines, a study was conducted to predict DM, TSS and mango maturity using PLSR

analysis, the findings found that the calibration model using NIR spectra, based on the coefficient of determination predicts DM with  $R^2=0.774$ , TSS with  $R^2=0.774$  and maturity with  $R^2=0.946$ , being optimal values that will serve as a basis for quality control and an automatic product classification system according to their fruit characteristics [14]. In another study presented, visible near-infrared spectroscopy (VNIRS) was used for nondestructive prediction of mango firmness during ripening, comparing the use of standard PLSR using the entire spectral range against an improved PLSR, which uses the variables selected by iPLSR. In response, an improved predictive model was obtained with an  $R^2=0.75$ , obtaining a better result compared to the standard PLSR model with  $R^2=0.67$ , achieving an increase of 12% in the  $R^2$  [9].

The above paragraphs present several investigations that describe encouraging results in the use of spectroscopic signals with PLSR and PCR algorithms for the estimation of internal parameters of mango fruit. Some of these studies use reflectance (R) as the model input variable [9], [17]–[19]. However, others of these must resort to preprocessing techniques to obtain good results. A common transformation is to apply the logarithm of the reciprocal reflectance  $\log(1/R)$ . For example, applied to mango to estimate softening of the flesh, total soluble solids content and acidity [19], but it has also been used to estimate internal properties of apple [20], bayberry [21], [22], pear [23]–[25], orange [26], bell peppers [18], strawberry [27], kiwifruit [28], low chilling peach [5], passion fruit, tomato and apricot [7]. Another transformation used is the first derivative of the reflectance  $dR$  for mango [19], but it has also been used in bell pepper [17], [18]. Continuing with the analysis of the derivative, other works have reported results with the first derivative of the logarithm of the reciprocal reflectance  $d\log(1/R)$  applied to mango [19], apple [20], pear [23], [25], bell pepper [17], [18], avocado [29], passion fruit, tomato and apricot [7] and mandarin fruit [10].

Spectral pretreatments known as standard normal variate (SNV) and multiplicative scatter correction (MSC) often give very similar results and are considered interchangeable [30]. Such techniques have already been applied in mango spectroscopy, on the one hand, SNV and MSC were used to estimate the TSS of the 'Nam Dok Mai Sithong' mango [31], while only MSC was used to estimate firmness in the 'Kent' mango [9]. Additionally, other investigations have used both techniques with PLSR and/or PCR showing encouraging results in fruits. For example, MSC and SNV have been used in orange fruit to estimate TSS, TA and BrimA [32], [33], also in loquat to estimate TSS and acidity [34], in *symplocos paniculata* to estimate oil content and acids [35] and in apples to detect bruise damage. Finally, some investigations have only considered using MSC to work on mandarin fruit [10] and banana [11].

As described above, different pre-processing techniques have been employed in NIR applications with PLSR and PCR. However, it has not been clear which of them would be the most suitable. Therefore, the objective of this study is to determine the most appropriate preprocessing techniques to estimate the TSS of Edwards mango by applying NIR spectroscopy with Machine Learning techniques such as PLSR and PCR. Therefore, in the following sections we analyzed a total of 18 PCR models and 18 PLSR models, where 4 types of transformations, 3 different feature extractors, and 3 different pre-processing techniques are combined.

## 2. METHOD

Figure 1 shows the methodology implemented. The methodology has been divided into 4 stages. Spectral data collection and DM measurement, a pre-processing stage of the raw data, a double cross validation (CV) stage and finally a testing phase to obtain the model metrics.

### 2.1. Getting raw data

#### 2.1.1. Spectral signature capture

Spectral signature recording was performed with an AvaSpec-NIR256-1.7-EVO NIR spectrometer in reflectance mode. The spectrometer has an operating spectral range of 900 to 1,750 nm with 221 bands. Eighteen Edward variety mangoes have been sampled; however, 12 measurement points have been identified on each of them, making a total of 216 samples. The location of the measurement points is determined by a measurement protocol detailed in Figure 2. First, the mango is placed with the dorsal shoulder to the right, then 3 circles are marked at different levels on the cheek of the mango as shown in Figure 2(a). Subsequently, the mango is turned 90 degrees and 3 new areas are marked at different levels as shown in Figure 2(b). The procedure is repeated as shown in Figures 2(c) and 2(d). In Figure 3 you can see raw spectral data taken in reflectance mode.

#### 2.1.2. Brix measurement

TSS was measured with a 0-32 °Brix hand-held refractometer with automatic temperature compensation (ATC). The objects under study correspond to 18 Edwards variety mangoes whose spectral signatures were previously recorded. The procedure consisted of quantifying the °Brix of a mango juice sample for each of the 12 points. First, the flesh of the fruit is exposed with the aid of a fruit peeler. Secondly, the

drops are extracted by inserting a cylindrical tip for 1 cm. Drops fall directly on the refractometer glass and the measurement is then made by exposing the instrument to a natural light source. Subsequently, the glass is cleaned by carefully pouring distilled water on it and finally dried with a cloth to enable the device to be used for a new measurement. Figure 4 shows the Brix values of the 216 samples taken from 18 mangoes.

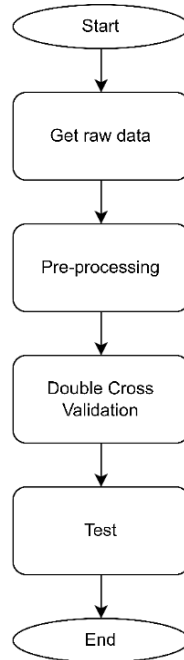


Figure 1. Methodology flowchart

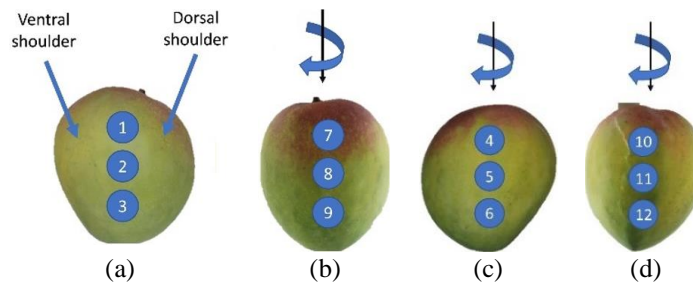


Figure 2. Location of measurement points; (a) morphology of the mango and location of points 1, 2 and 3, (b) location of points 7, 8 and 9, (c) location of points 4, 5 and 6, and (d) location of points 10, 11 and 12

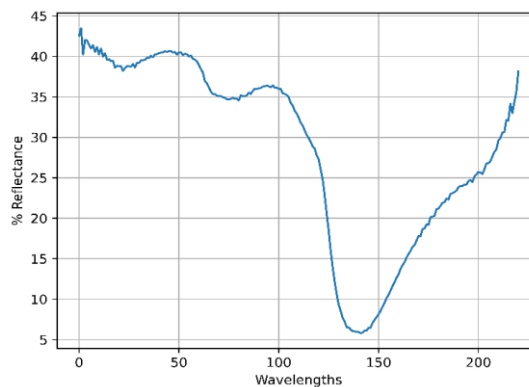


Figure 3. Raw spectral data of one sample

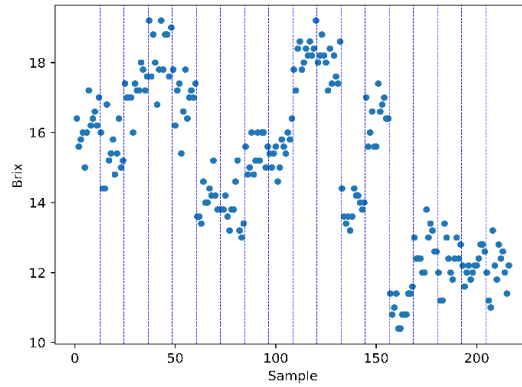


Figure 4. Brix values of the 216 samples. Samples are divided by vertical lines to represent samples of the same mango

**2.2. Pre-processing**

A flow chart of the pre-processing stage is detailed in Figure 5. First, the dependent variable went through a transformation substage (logarithmic, square root, power of two or no transformation). On the other hand, the independent variables Figure 3 went through a filtering sub-stage applying a Savitzky-Golay filter with a 7-point window and a second-degree polynomial. The most used smoothing and differencing technique in chemometrics is the Savitzky-Golay method [36]–[38], which consists of a local polynomial regression requiring equidistant spectrum values, as shown in (1).

$$x_j^* = \frac{1}{N} \sum_{h=-k}^k c_h x_{j+h} \tag{1}$$

where:  $x_j^*$  is the value of the curve to be smoothed or derived.  $N$  is the number of points in the window.  $k$  is the number of neighboring points per side.  $c_h$ : are the coefficients that depend on the degree of the regression polynomial and the objective (smoothed or derived). The filter outcomes are shown in Figure 6.

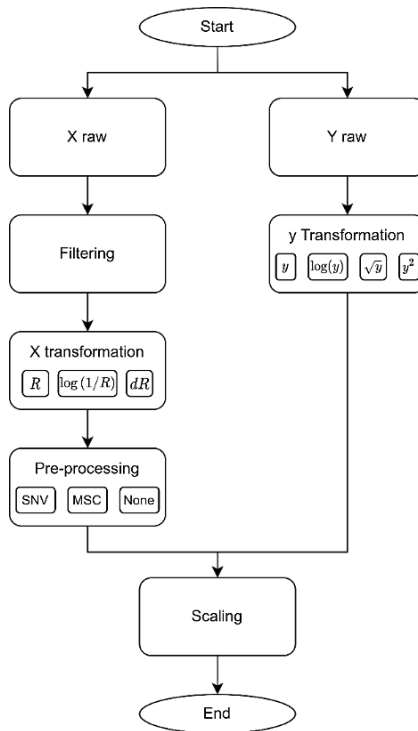


Figure 5. Flow diagram of the pre-processing stage

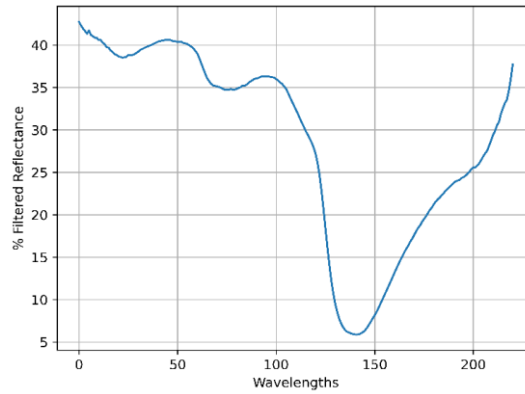


Figure 6. Reflectance signal filtered with Savitzky-Golay filter with a 7-point window and a second-degree polynomial

The filtered independent variable went through a transformation substage as shown in Figure 7, after which one of the following results is obtained: no transformation (R), Figure 7(a); logarithmic transformation  $\log(1/R)$ , Figure 7(b); and first derivative (dR) as shown in Figure 7(c). The first derivative was calculated by applying a Savitzky-Golay filter with a 7-point window and a second-degree polynomial.

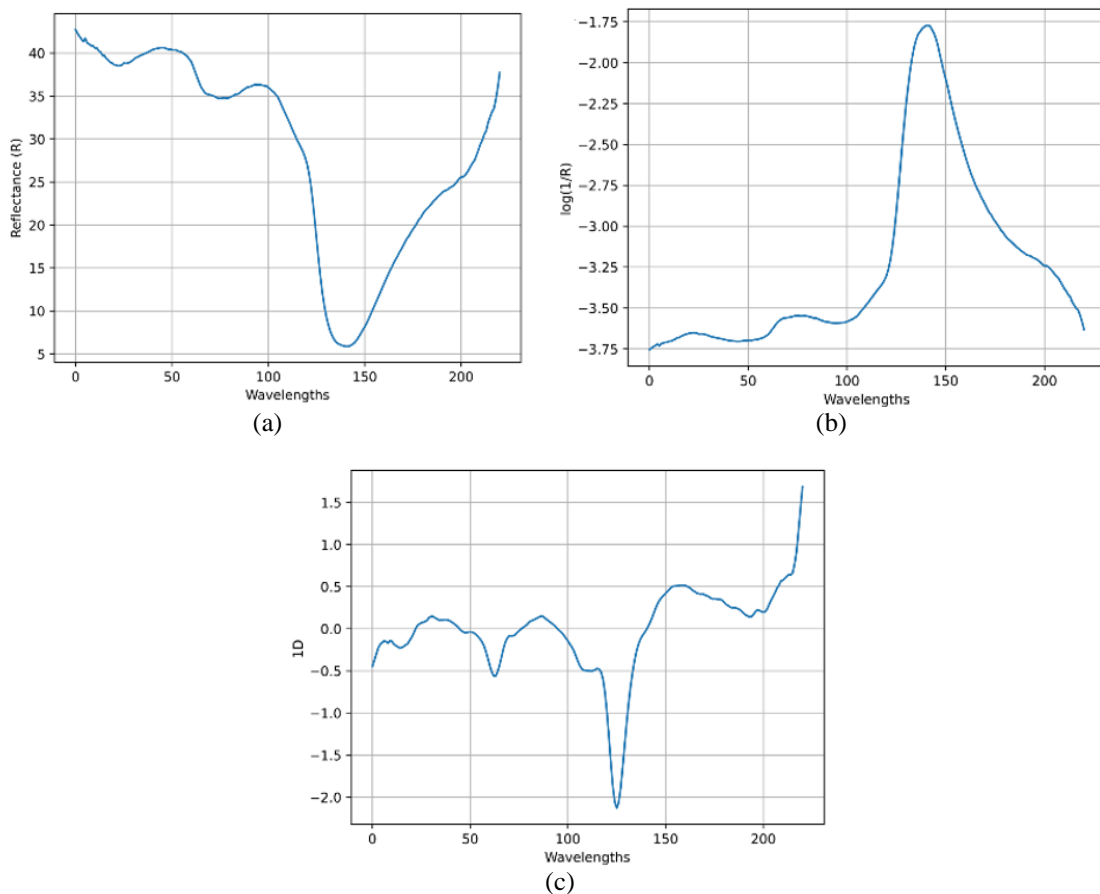


Figure 7. X transformation (a) R or no transformation, (b)  $\log(1/R)$  transformation and (c) 1D transformation

Additionally, transformed independent variables were passed to a pre-processing substage. In this stage, one of the following techniques was applied: SNV, MSC or none of the above. MSC is a preprocessing method suitable for reflectance or transmission signals in diffuse samples, since its objective is to reduce the

disturbing effect of light scattering [39], [40]. Unlike the spectral derivatives, the MSC preserves the spectral shape. With  $x_i$  being the vector spectrum  $i$  of a set  $m$  samples, the MSC model for a spectrum is shown in (2).

$$x_i = a_i \mathbf{1} + b_i \bar{x}_i + e_i \quad (2)$$

where  $\mathbf{1}$  is a column vector of ones of the same length as  $x_i$ . The parameters  $a_i$  and  $b_i$  are estimated for each spectrum  $x_i$ . Thus, each spectrum is corrected as (3).

$$x_{ij\_MSC} = \frac{x_{ij} - a_i}{b_i} \quad (3)$$

where  $x_{ij}$  is the value of absorbance  $i$  of spectrum  $j$ .

In the other hand, SNV treats each spectrum separately absorbance values as (4):

$$x_{ij\_SNV} = \frac{x_{ij} - \bar{x}_i}{s_i} \quad (4)$$

where  $\bar{x}_i$  and  $s_i$  are the mean and standard deviation of the absorbances  $x_{ij}$  in spectrum  $i$ . Thus, the transformed absorbances have zero mean and unit standard deviation in each spectrum. An example can be seen in Figure 8, where SNV is applied to the transformation R in Figure 8(a), MSC is applied to R in Figure 8(b) and no transformation is applied to R in Figure 8(c).

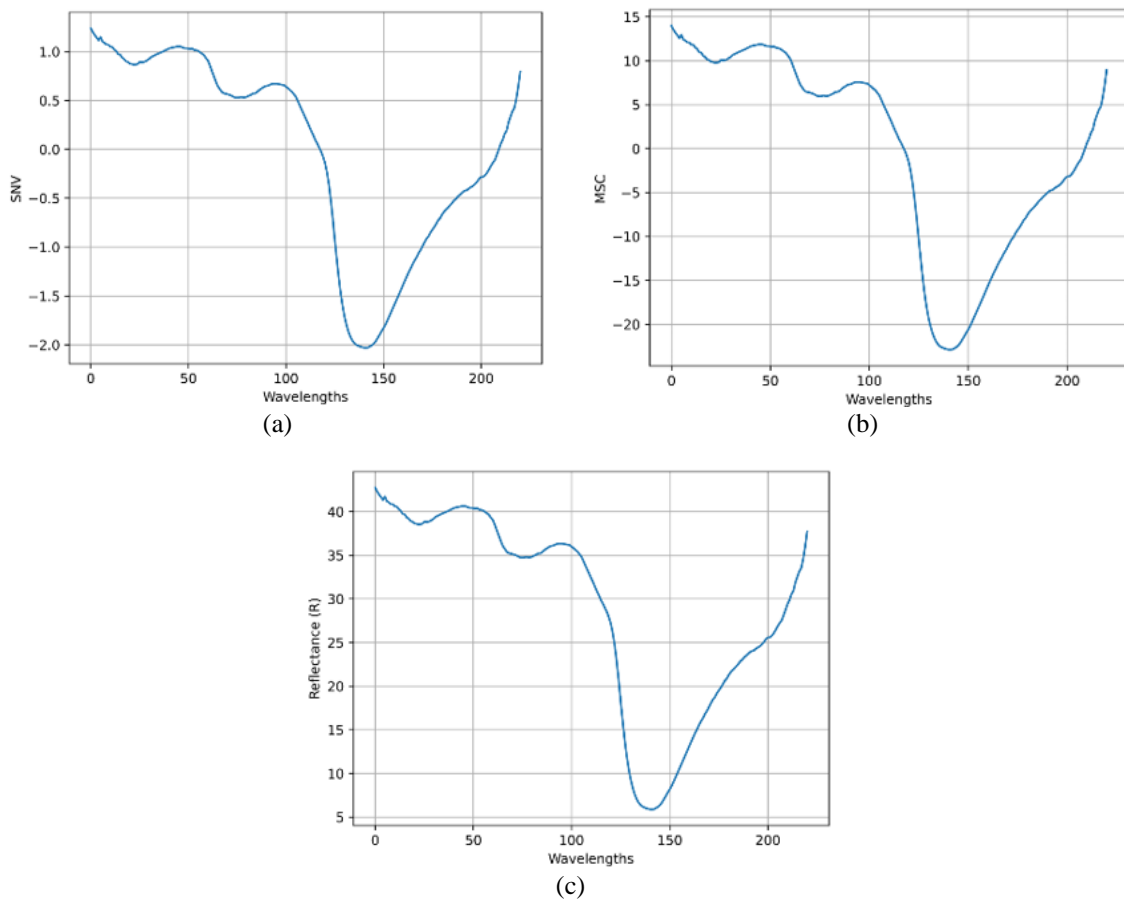


Figure 8. Pre-processing techniques applied to R transformation; (A) SNV, (b) MSC and (c) none

Finally, both the independent variables and the dependent variable went through a scaling stage. Here, the mean was removed and the standard deviation was standardized to a value of 1 for each wavelength. Therefore, as an example, look at Figure 9 which is the result of applying the scaling to Figure 8(a).

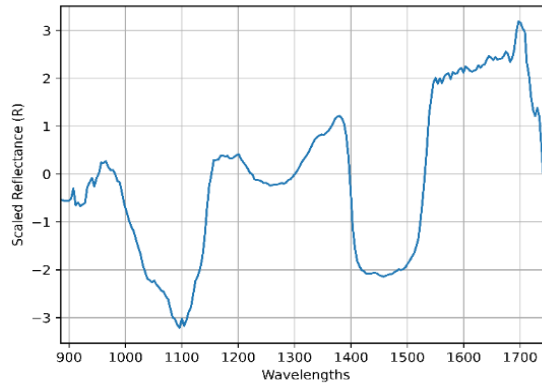


Figure 9. Output of scaling stage for R signal

**2.3. Double CV with repetitions**

A schematic of the double CV is explained in Figure 10. The internal CV defines the training and validation sets; and the optimal number of  $a_{opt}$  components. The inner CV is repeated 50 times, resulting in a list of 50  $a_{opt}$ . The external CV defines the test and calibration sets. Each external segmentation obtains the list of 50  $a_{opt}$ . At the end the  $s_{OUT}$  lists are grouped.

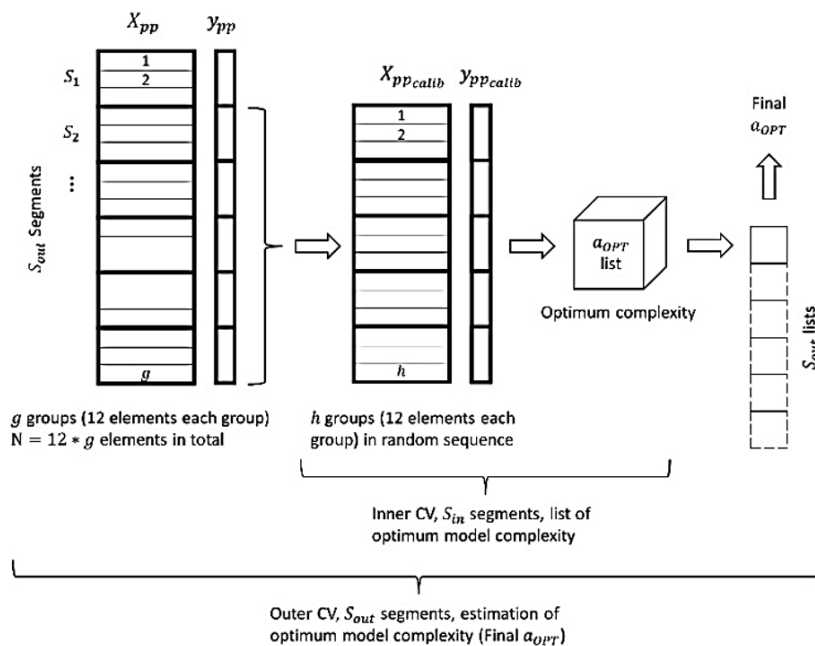


Figure 10. Double CV with  $s_{OUT}$  segments in the outer CV and  $s_{IN}$  segments in the inner CV executed to obtain the  $a_{opt}$ . Adapted from [41]

PCA is especially useful for data sets with correlated variables such as spectral data. However, the technique is sensitive to outliers and scaling. Principal least-squares (PLS) is a technique for relating the matrix  $X$  to the vector  $y$  [42]–[44]. A crucial point to build a predictive model with PCR or PLSR is to decide the number of components to be used. In principle, previously the independent variables can pass a variable selection stage, however for simplicity the project is limited to finding the optimal number of components  $a_{opt}$ . In the internal CV, the calibration set was divided into training and validation sets. The 12 samples from each handle were combined, shuffled and finally distributed in the same set to avoid false optimistic results. Then, the selection of the optimal number of principal components is based on one standard error rule [45]. Finally, the  $s_{OUT}$  lists of 50 candidates of optimal number of components were analyzed using a histogram. The final  $a_{opt}$  is determined by calculating the mode of the candidates.



**2.3. Test**

The external CV is rerun to build a model with  $a_{opt}$  components. The training uses the entire calibration set and makes predictions from the test set. The procedure is repeated for the other CV segmentations until the prediction of all data is completed as shown in Figure 11. The result of the procedure explained in Figure 12 are the test set predictions scaled by the pre-processing stage and  $\hat{y}_{pp_{test}}$  that attempt to fit  $y_{pp_{test}}$ . In addition,  $y_{pp_{test}}$  and  $\hat{y}_{pp_{test}}$  went through an inversion of the previous scaling and transformation to obtain  $y_{test}$  and  $\hat{y}_{test}$ . Finally, applying the following equation yields  $RMSE_{pp}$  and  $RMSE_{test}$ . RMSE is defined in (5):

$$RMSE = \sqrt{\frac{\sum(y-\hat{y})^2}{N}} \tag{5}$$

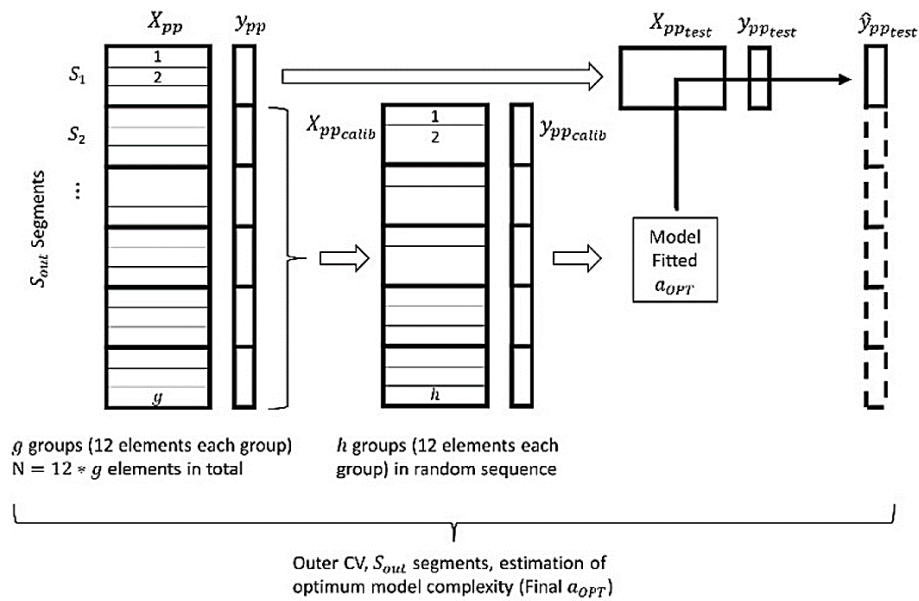


Figure 11. Double CV with  $s_{OUT}$  segments in the external CV executed to obtain test predictions of all data

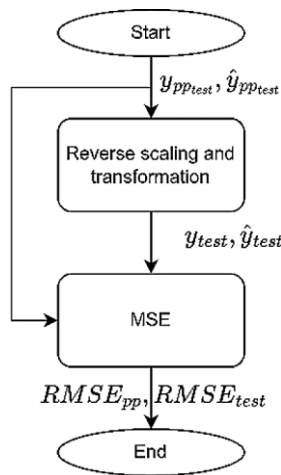


Figure 12. RMSE calculation flowchart

**3. RESULTS AND DISCUSSION**

The method presented consisted of evaluating 2 algorithms, 4 transformations, 3 types of features extractors and 3 pre-processing techniques, in other words, 36 models as shown in Table 1. The model with

the lowest RMSE applied PLSR, used the  $y^2$ -transform, using R as the independent variable and SNV as the pre-processing technique Table 1. This model obtained an RMSE of 1.1382 °Brix and an  $RMSE_{pp}$  of 0.5140 in the transformed dimensional scale.

Table 1. RMSE of the 36 models.

Feature	PP <sup>a</sup>	y		log(y)		$\sqrt{y}$		$y^2$	
		PCR	PLSR	PCR	PLSR	PCR	PLSR	PCR	PLSR
R	SNV	1.2076	1.1965	1.2449	1.2467	1.2215	1.2158	1.2053	1.1382
	MSC	1.2063	1.1969	1.2439	1.2473	1.2203	1.2162	1.2036	1.1384
	None	1.3227	1.3072	1.3494	1.3337	1.3325	1.3156	1.1662	1.2112
$Log\left(\frac{1}{R}\right)$	SNV	1.2850	1.2841	1.3369	1.3297	1.3057	1.3021	1.2670	1.2699
	MSC	1.2882	1.2877	1.3401	1.3332	1.3089	1.3056	1.2702	1.2735
	None	1.1894	1.2557	1.3343	1.2694	1.3255	1.2598	1.1786	1.1627
1D	SNV	1.3274	1.2678	1.3643	1.3063	1.3457	1.2840	1.1842	1.1788
	MSC	1.3320	1.1968	1.4009	1.3134	1.3506	1.2155	1.3197	1.1830
	None	1.2159	1.1966	1.3148	1.3105	1.2915	1.2776	1.1855	1.1674

Note: <sup>a</sup>Pre-processing

Then, the best model for Brix is analyzed. Figure 13 shows the histogram of the optimal number of components found in the double CV for the PLSR- $y^2$ -R-SNV model. Therefore, finally the model used only 4 components.

Figure 14 shows the MSE on the transformed scale of the variable for different model complexities in the Double CV. The gray line shows the MSE in each of the CVs, while the blue line represents the average of all the gray lines. The red dashed line shows the optimal number of components determined in Figure 13. Finally, the cyan dashed line shows the MSE level in the test set.

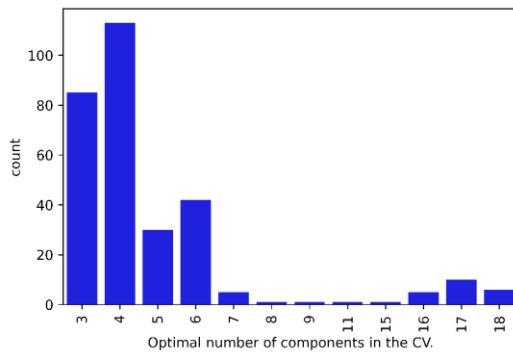


Figure 13. Histogram of the optimal number of components in the double CV for the PLSR- $y^2$ -R- SNV model

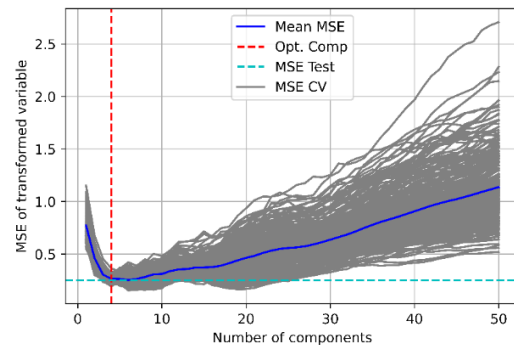


Figure 14. MSE of transformed variable to different model complexity in Double CV

Figure 15 shows predicted °Brix values of the best model versus transformed °Brix values ( $y^2$ ) after scaling in Figure 15(a), while in Figure 15(b) we can see the predicted values versus true °Brix values. In both graphs a clear trend can be observed on the blue line which represents a perfect model. In gray color are shown the predictions made in each of the CVs of the Double CV. Additionally, the predictions of the test set are represented with different colors corresponding to each mango.

Figure 16 shows prediction errors versus true values. In Figure 16(a) we can see the prediction error in the transformed ( $y^2$ ) and scaled system, while in Figure 16(b) we can see the prediction error in the true °Brix system versus the true value. In both graphs a tendency to a negative error can be observed for high °Brix values. The gray color shows the prediction errors for each of the CVs of the Double CV. Additionally, the prediction errors of the test set are represented with different colors corresponding to each mango.

Figure 17 shows the prediction errors versus sample number. In Figure 17(a) we can see the prediction error in the transformed ( $y^2$ ) and scaled system, while in Figure 17(b) we can see the prediction error in the true °Brix system versus the sample number. In both plots, a mean close to zero and a relatively constant variance distribution can be observed, so the error does not depend of the sample position. The prediction errors of the test set are represented with different colors corresponding to each mango.

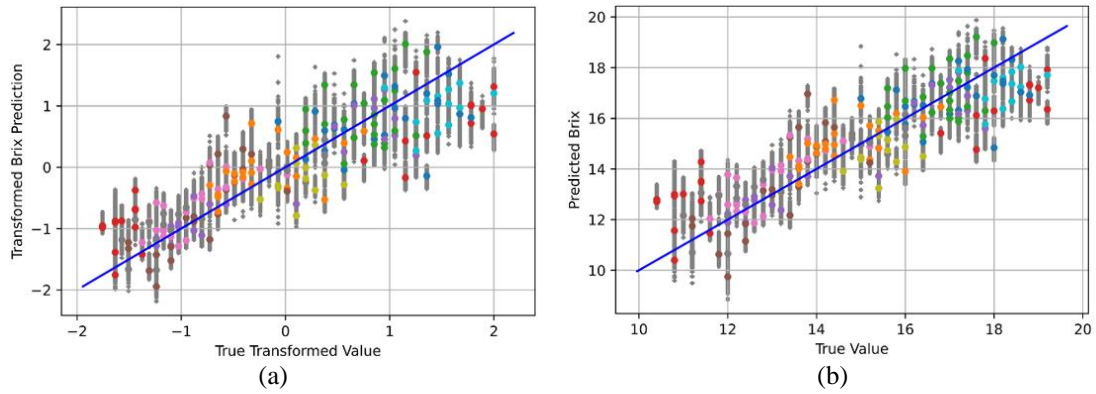


Figure 15. Brix prediction vs true value; (a) transformed and scaled values and (b) true scale

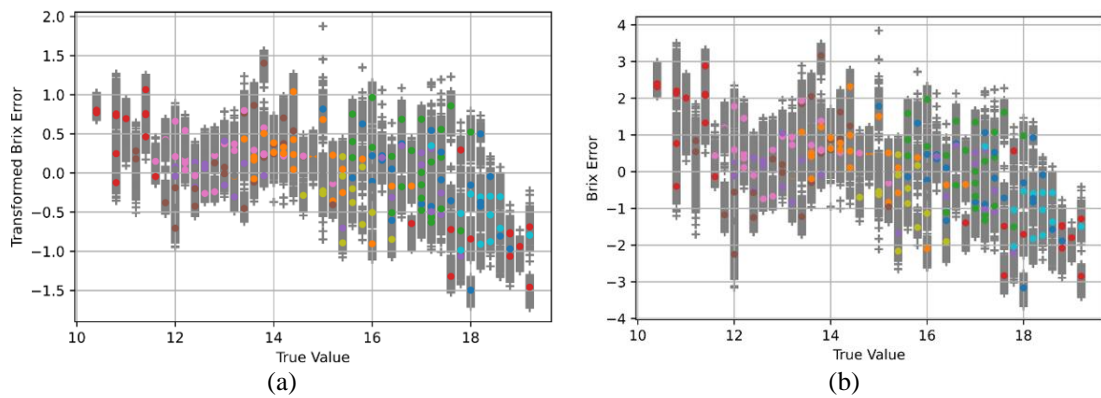


Figure 2. Brix Error Vs True value; (a) scale in transformed and scaled values and (b) true scale

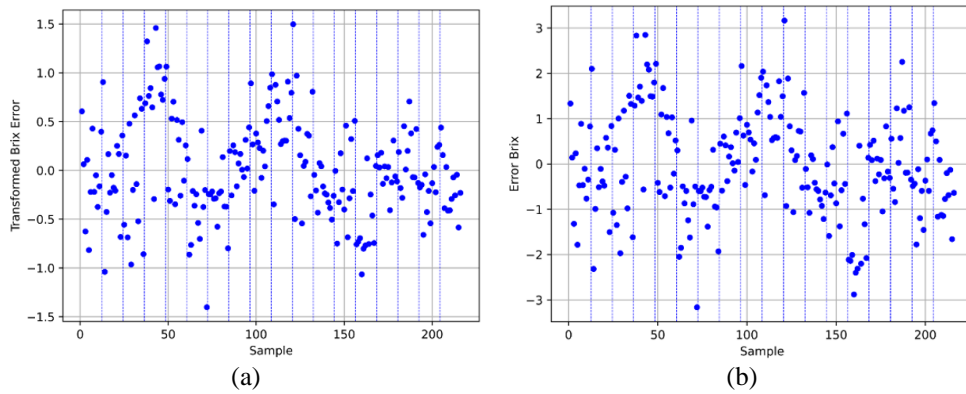


Figure 17. Brix Error Vs Sample. A) Scale in transformed and scaled values. B) Real scale

The results are analyzed based on the average RMSE and the range considering the standard deviation of the expected value according to (6):

$$Std = \hat{\sigma} / \sqrt{c} \tag{6}$$

where *Std* is the standard deviation of the expected value,  $\hat{\sigma}$  is standard deviation of the samples and *c* is the number of samples.

Table 2 shows the expected PCR and PLSR RMSEs and ranges for the Brix variable applying different transformations. The PLSR results are better than PCR when applying the  $\sqrt{y}$  transformation. However,

insufficient evidence has been found to claim that PLSR is better than PCR for the other transformations. On the other hand,  $y^2$  transformation has shown that on average it presents better results than other transformations for both techniques.

Table 2. Average and range of RMSE applying different transformations

Transformation	RMSE PCR	Interval RMSE PCR	RMSE PLSR	Interval RMSE PLSR
$y$	1.2638	[1.2454 1.2823]	1.2433	[1.2287 1.2578]
$\log(y)$	1.3255	[1.3092 1.3418]	1.2989	[1.2878 1.31]
$\sqrt{y}$	1.3002	[1.2849 1.3156]	1.2658	[1.2529 1.2786]
$y^2$	1.2200	[1.2035 1.2366]	1.1914	[1.1755 1.2074]

Note: c value is 9.

Table 3 shows the expected RMSE of PCR and PLSR, as well as the ranges for the Brix variable using different characteristics. The PLSR results are better than PCR when using the first derivative,  $1D$ . However, insufficient evidence has been found to affirm that PLSR is better than PCR for the other transformations. On the other hand, the use of R has been shown to perform on average better than other features when using PCR. Additionally, the use of R and  $1D$  has shown that on average it presents better results than  $\log(1/R)$  when using PLSR.

Table 4 shows the expected RMSE of PCR and PLSR, as well as the ranges for the Brix variable applying different pre-processing techniques. The PLSR results are better than PCR when using MSC. However, insufficient evidence has been found to claim that PLSR is better than PCR with other pre-processing techniques. On the other hand, there is not enough evidence to state that any technique presents better results.

Table 3. Average and range of RMSE applying different types of feature extraction.

Feature	RMSE PCR	Interval RMSE PCR	RMSE PLSR	Interval RMSE PLSR
R	1.2437	[1.2274 1.26]	1.2303	[1.2127 1.2479]
$\log(1/R)$	1.2858	[1.271 1.3007]	1.2778	[1.2656 1.29]
$1D$	1.3027	[1.2832 1.3222]	1.2415	[1.2258 1.2572]

Note: c value is 12.

Table 4. Average and range of RMSE applying different pre-processing techniques.

Pre-processing	RMSE PCR	Interval RMSE PCR	RMSE PLSR	Interval RMSE PLSR
SNV	1.2746	[1.2575 1.2917]	1.2517	[1.2357 1.2677]
MSC	1.2904	[1.273 1.3077]	1.2423	[1.2256 1.259]
None	1.2672	[1.2471 1.2873]	1.2556	[1.2394 1.2718]

Note: c value is 12

#### 4. CONCLUSION

A total of 18 PCR and 18 PLSR models have been tested. A methodology based on a double VC has been implemented. The internal CV was used to find the  $a_{opt}$  found with the one standard deviation rule. The model with the lowest RMSE used PLSR with 4 components, used the  $y^2$  transformation, reflectance R as the independent variable, and SNV as the pre-processing technique. This model obtained an RMSE of 1.1382 °Brix and an  $RMSE_{pp}$  of 0.514 on the transformed dimensional scale. The  $y^2$  transformation has shown better metrics than other transformations in both algorithms. Additionally, working directly with reflectance has given good results. Sufficient evidence has not been found to affirm that any pre-processing technique is better than another. Additionally, it can be affirmed that the PLSR always showed equal or better results than the PCR.

#### REFERENCES




- [1] S. Marçal and M. Pintado, "Mango peels as food ingredient/additive: nutritional value, processing, safety and applications," *Trends in Food Science & Technology*, vol. 114, pp. 472–489, Aug. 2021, doi: 10.1016/j.tifs.2021.06.012.
- [2] P. Ojha, S. Raut, U. Subedi, and N. Upadhaya, "Study of nutritional, phytochemicals and functional properties of mango kernel powder," *Journal of Food Science and Technology Nepal*, vol. 11, pp. 32–38, Dec. 2019, doi: 10.3126/jfstn.v11i0.29708.
- [3] F. and Agriculture, *Principales frutas tropicales. Análisis del Mercado 2021*. Rome, Italy, 2022.
- [4] L. L. T. Coral and H. A. Escobar-García, "Characterization of fruits of varieties of mango (*Mangifera indica*) conserved in Peru," *Revista Brasileira de Fruticultura*, vol. 43, no. 2, 2021, doi: 10.1590/0100-29452021710.
- [5] P. A. M. Nascimento, L. C. de Carvalho, L. C. C. Júnior, F. M. V. Pereira, and G. H. de A. Teixeira, "Robust PLS models for soluble solids content and firmness determination in low chilling peach using near-infrared spectroscopy (NIR)," *Postharvest Biology and Technology*, vol. 111, pp. 345–351, Jan. 2016, doi: 10.1016/j.postharvbio.2015.08.006.

- [6] B. M. Nicolai *et al.*, "Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review," *Postharvest Biology and Technology*, vol. 46, no. 2, pp. 99–118, Nov. 2007, doi: 10.1016/j.postharvbio.2007.06.024.
- [7] G. A. de Oliveira, S. Bureau, C. M.-G. C. Renard, A. B. Pereira-Netto, and F. de Castilhos, "Comparison of NIRS approach for prediction of internal quality traits in three fruit species," *Food Chemistry*, vol. 143, pp. 223–230, Jan. 2014, doi: 10.1016/j.foodchem.2013.07.122.
- [8] J. F. Isingizwe Nturambirwe, H. H. Nieuwoudt, W. J. Perold, and U. L. Opara, "Detecting bruise damage and level of severity in apples using a contactless NIR spectrometer," *Applied Engineering in Agriculture*, vol. 36, no. 3, pp. 257–270, 2020, doi: 10.13031/aea.13218.
- [9] P. Mishra, E. Woltering, and N. El Harchioui, "Improved prediction of 'Kent' mango firmness during ripening by near-infrared spectroscopy supported by interval partial least square regression," *Infrared Physics & Technology*, vol. 110, Nov. 2020, doi: 10.1016/j.infrared.2020.103459.
- [10] S. Xudong, Z. Hailiang, and L. Yande, "Nondestructive assessment of quality of nanfeng mandarin fruit by a portable near infrared spectroscopy," *International Journal of Agricultural and Biological Engineering*, vol. 2, no. 1, pp. 65–71, 2009.
- [11] P. Jaiswal, S. N. Jha, and R. Bharadwaj, "Non-destructive prediction of quality of intact banana using spectroscopy," *Scientia Horticulturae*, vol. 135, pp. 14–22, Feb. 2012, doi: 10.1016/j.scienta.2011.11.021.
- [12] T. Ignat *et al.*, "Forecast of apple internal quality indices at harvest and during storage by VIS-NIR spectroscopy," *Food and Bioprocess Technology*, vol. 7, no. 10, pp. 2951–2961, Oct. 2014, doi: 10.1007/s11947-014-1297-7.
- [13] V. R. Sharabiani, S. Sabzi, R. Pourdarbani, M. Szymanek, and S. Michałek, "Inner properties estimation of gala apple using spectral data and two statistical and artificial intelligence based methods," *Foods*, vol. 10, no. 12, Dec. 2021, doi: 10.3390/foods10122967.
- [14] Y. Q. Polinar, K. F. Yaptenco, E. K. Peralta, and J. U. Agravante, "Near-infrared spectroscopy for non-destructive prediction of maturity and eating quality of 'Carabao' mango (*Mangifera indica* L.) fruit," *Agricultural Engineering International: CIGR Journal*, vol. 21, no. 1, pp. 209–219, 2019.
- [15] R. Pourdarbani, S. Sabzi, D. Kalantari, and J. I. Arribas, "Non-destructive visible and short-wave near-infrared spectroscopic data estimation of various physicochemical properties of Fuji apple (*Malus pumila*) fruits at different maturation stages," *Chemometrics and Intelligent Laboratory Systems*, vol. 206, Nov. 2020, doi: 10.1016/j.chemolab.2020.104147.
- [16] P. Maniwaru *et al.*, "Evaluation of NIRS as non-destructive test to evaluate quality traits of purple passion fruit," *Scientia Horticulturae*, vol. 257, Nov. 2019, doi: 10.1016/j.scienta.2019.108712.
- [17] T. Ignat, Z. Schmilovitch, J. Fefoldi, B. Steiner, and S. Alkalai-Tuvia, "Non-destructive measurement of ascorbic acid content in bell peppers by VIS-NIR and SWIR spectrometry," *Postharvest Biology and Technology*, vol. 74, pp. 91–99, Dec. 2012, doi: 10.1016/j.postharvbio.2012.06.010.
- [18] Z. Schmilovitch, T. Ignat, V. Alchanatis, J. Gatker, V. Ostrovsky, and J. Felföldi, "Hyperspectral imaging of intact bell peppers," *Biosystems Engineering*, vol. 117, pp. 83–93, Jan. 2014, doi: 10.1016/j.biosystemseng.2013.07.003.
- [19] Z. Schmilovitch, A. Mizrach, A. Hoffman, H. Egozi, and Y. Fuchs, "Determination of mango physiological indices by near-infrared spectrometry," *Postharvest Biology and Technology*, vol. 19, no. 3, pp. 245–252, Jul. 2000, doi: 10.1016/S0925-5214(00)00102-2.
- [20] Y. Liu and Y. Ying, "Use of FT-NIR spectrometry in non-invasive measurements of internal quality of 'Fuji' apples," *Postharvest Biology and Technology*, vol. 37, no. 1, pp. 65–71, Jul. 2005, doi: 10.1016/j.postharvbio.2005.02.013.
- [21] Y. Shao and Y. He, "Nondestructive measurement of the internal quality of bayberry juice using Vis/NIR spectroscopy," *Journal of Food Engineering*, vol. 79, no. 3, pp. 1015–1019, Apr. 2007, doi: 10.1016/j.jfoodeng.2006.04.006.
- [22] X. Li and Y. He, "Non-destructive measurement of acidity of Chinese bayberry using Vis/NIRS techniques," *European Food Research and Technology*, vol. 223, no. 6, pp. 731–736, Oct. 2006, doi: 10.1007/s00217-006-0260-x.
- [23] Y. Liu, X. Chen, and A. Ouyang, "Nondestructive determination of pear internal quality indices by visible and near-infrared spectrometry," *LWT - Food Science and Technology*, vol. 41, no. 9, pp. 1720–1725, Nov. 2008, doi: 10.1016/j.lwt.2007.10.017.
- [24] F. Bexiga *et al.*, "A TSS classification study of 'Rocha' pear (*Pyrus communis* L.) based on non-invasive visible/near infra-red reflectance spectra," *Postharvest Biology and Technology*, vol. 132, pp. 23–30, Oct. 2017, doi: 10.1016/j.postharvbio.2017.05.014.
- [25] L. Yan-De, C. Xing-Mao, S. Xu-Dong, and Y. Yi-Bin, "Non-destructive measurement of pear internal quality indices by visible and near-infrared spectrometric techniques," *New Zealand Journal of Agricultural Research*, vol. 50, no. 5, pp. 1051–1057, Dec. 2007, doi: 10.1080/00288230709510385.
- [26] Y. Zheng *et al.*, "Predicting oleocellosis sensitivity in citrus using VNIR reflectance spectroscopy," *Scientia Horticulturae*, vol. 125, no. 3, pp. 401–405, Jun. 2010, doi: 10.1016/j.scienta.2010.04.036.
- [27] Y. Shao and Y. He, "Nondestructive measurement of acidity of strawberry using Vis/NIR spectroscopy," *International Journal of Food Properties*, vol. 11, no. 1, pp. 102–111, Feb. 2008, doi: 10.1080/10942910701257057.
- [28] H. Singh, A. Sridhar, and S. S. Saini, "Ultra-low-cost self-referencing multispectral detector for non-destructive measurement of fruit quality," *Food Analytical Methods*, vol. 13, no. 10, pp. 1879–1893, Oct. 2020, doi: 10.1007/s12161-020-01810-7.
- [29] L. Dvash *et al.*, "Determination by near-infrared spectroscopy of perseitol used as a marker for the botanical origin of avocado (*persea americana* mill.) honey," *Journal of Agricultural and Food Chemistry*, vol. 50, no. 19, pp. 5283–5287, Sep. 2002, doi: 10.1021/jf020329z.
- [30] T. Fearn, C. Riccioli, A. Garrido-Varo, and J. E. Guerrero-Ginel, "On the geometry of SNV and MSC," *Chemometrics and Intelligent Laboratory Systems*, vol. 96, no. 1, pp. 22–26, Mar. 2009, doi: 10.1016/j.chemolab.2008.11.006.
- [31] S. Sharma, P. Sirisomboon, and P. Pornchaloempong, "Application of a Vis-NIR spectroscopic technique to measure the total soluble solids content of intact mangoes in motion on a belt conveyor," *The Horticulture Journal*, vol. 89, no. 5, pp. 545–552, 2020, doi: 10.2503/hortj.UTD-168.
- [32] Y. Liu, X. Sun, and A. Ouyang, "Nondestructive measurement of soluble solid content of navel orange fruit by visible-NIR spectrometric technique with PLSR and PCA-BPNN," *LWT - Food Science and Technology*, vol. 43, no. 4, pp. 602–607, May 2010, doi: 10.1016/j.lwt.2009.10.008.
- [33] B. Jamshidi, S. Minaei, E. Mohajerani, and H. Ghassemian, "Reflectance Vis/NIR spectroscopy for nondestructive taste characterization of Valencia oranges," *Computers and Electronics in Agriculture*, vol. 85, pp. 64–69, Jul. 2012, doi: 10.1016/j.compag.2012.03.008.
- [34] X. Fu *et al.*, "Determination of soluble solid content and acidity of loquats based on FT-NIR spectroscopy," *Journal of Zhejiang University SCIENCE B*, vol. 10, no. 2, pp. 120–125, Feb. 2009, doi: 10.1631/jzus.B0820097.
- [35] Q. Liu *et al.*, "Determination of fruit oil content and fatty acid composition in *symplocos paniculata* using near infrared reflectance spectroscopy," *Journal of Biobased Materials and Bioenergy*, vol. 10, no. 4, pp. 272–278, Aug. 2016, doi: 10.1166/jbmb.2016.1607.
- [36] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964, doi: 10.1021/ac60214a047.
- [37] O. Y. Rodionova, "Chemometrics: Data analysis for the laboratory and chemical plant," *Journal of Analytical Chemistry*, vol. 60, no. 10, pp. 994–996, Oct. 2005, doi: 10.1007/s10809-005-0223-6.




- [38] J. Steinier, Y. Termonia, and J. Deltour, "Smoothing and differentiation of data by simplified least square procedure," *Analytical Chemistry*, vol. 44, no. 11, pp. 1906–1909, Sep. 1972, doi: 10.1021/ac60319a045.
- [39] P. Geladi, D. MacDougall, and H. Martens, "Linearization and scatter-correction for near-infrared reflectance spectra of meat," *Applied Spectroscopy*, vol. 39, no. 3, pp. 491–500, May 1985, doi: 10.1366/0003702854248656.
- [40] T. Naes, T. Isaksson, and B. Kowalski, "Locally weighted regression and scatter correction for near-infrared reflectance data," *Analytical Chemistry*, vol. 62, no. 7, pp. 664–673, Apr. 1990, doi: 10.1021/ac00206a003.
- [41] K. Varmuza and P. Filzmoser, *Introduction to multivariate statistical analysis in chemometrics*. CRC Press; 1st edition, 2009.
- [42] S. Wold, M. Sjostrom, and L. Eriksson, "Partial least squares projections to latent structures (PLS) in chemistry," in *Encyclopedia of Computational Chemistry*, Chichester, UK: John Wiley & Sons, Ltd, 2002. doi: 10.1002/0470845015.cpa012.
- [43] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, Oct. 2001, doi: 10.1016/S0169-7439(01)00155-1.
- [44] I. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, May 1993, doi: 10.2307/1269656.
- [45] T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning*. New York, NY: Springer New York, 2001. doi: 10.1007/978-0-387-21606-5.

## BIOGRAPHIES OF AUTHORS






**Ernesto Paiva-Peredo**    received the title of electrical mechanical engineer from the University of Piura, Peru, in 2013; He has completed a master's degree in electrical mechanical engineering with a mention in automation and optimization at the University of Piura funded by CONCYTEC 2016. He was a research assistant at the Department of Technology and Innovation (DTI) - SUPSI. Now, he is a Professor-Researcher at Universidad Tecnológica del Perú. He can be contacted at email epaiva@utp.edu.pe.






**Diego Gonzales-Rodriguez**    received the B.Sc. degree in electronic engineering from the Universidad Tecnológica del Perú, Peru, in 2022. He worked on the Development of Machine Learning algorithms as solutions to problems. He made a research article to support the work done. He collected research data through experimentation and entered information into databases and other types of computer programs. He worked on the design of experiments to collect data under the orders of the head of research. He can be contacted at email U17202491@utp.edu.pe.






**Wiliam Trujillo Herrera**    received the B.Sc. degree in physics from National San Marcos University, Perú, in 2005 and the M.Sc. and Ph.D. degrees in physics from Brazilian Center for Research in Physics (CBPF), Rio de Janeiro, Brazil, in 2007 and 2011, respectively. He completed a specialization in "Innovation and entrepreneurship" at Singularity University, Silicon Valley, California, USA. Currently, he is Research Professor at the Department of Industrial Engineering, Technological University of Peru and he is partner of WAREM company, specializing in the sale of scientific equipment and related services to universities and technology-based companies. His research interests include nanoscience and nanotechnology, green synthesis of nanomaterials, UV-Vis and NIR spectroscopy, flavonoids and polyphenols, anthocyanins and betalains of endemic fruits and plants of Andes, melanin pigment from mushrooms, UV effect on biosynthesis of secondary metabolites in plants and mushrooms, legume fermentation with filamentous fungi. He can be contacted at email wtrujillo@utp.edu.pe.






**Juan Jesús Soria Quijaite**    master in applied mathematics from the Universidad Nacional de Ingeniería (1997), PhD in Systems Engineering (2010) and specialist in Statistics for Research at the Universidad Peruana de la Unión (2013). Currently, Research Professor at the Department of Systems Engineering of the Universidad Tecnológica del Perú. He has more than 7 years of experience participating in R&D projects in artificial intelligence and machine learning. His research interests include mathematical optimization, numerical methods, differential equations, functional analysis, predictive models, machine learning, neural networks, data mining, clustering with artificial intelligence. He can be reached by email at C20723@utp.edu.pe.



**Diana Quispe-Arpasi**    food engineer trained at the Universidad Peruana Unión (2012), with a master's (2016) and PhD (2021) in the environmental biotechnology research line from the University of Sao Paulo (USP) in Brazil, where he developed his thesis on the application of anaerobic digestion and hydrothermal liquefaction for the energy use of cyanobacteria, in collaboration with the Environmental Sanitation Laboratory of the Federal University of Pernambuco (Brazil) and the Ecotoxicology and Applied Ecology Center of USP (Brazil). In 2022 she belonged to the Innovative Energy Technologies for Biofuels, Bioenergy and a Sustainable Irish Bioeconomy (IETS BIO3) research group at the University of Galway (Ireland) working on Biorefinery and Bioremediation Projects. She has more than seven years of experience participating in Biotechnology Research Projects, especially in the area of valorization of agro-industrial waste by anaerobic digestion. She is currently a Research Professor at the Universidad Peruana Unión and is interested in research related to the Circular Economy of Waste (Biorefineries) and Bioremediation of Contaminated Soils. She can be contacted at email [C22967@utp.edu.pe](mailto:C22967@utp.edu.pe).



**Christian Ovalle Paulino**    he is an Associate Professor at the Faculty of Engineering of the Technological University of Peru, Lima-Peru. He has a Ph.D. in systems engineering with a specialization in artificial intelligence. His research areas are process mining, business data analysis and pattern recognition. He is CEO of the 7D consultancy dedicated to the investigation of intelligent solutions. He has participated in different research projects, receiving awards from the Peruvian Ministry of Defense and the Armed Forces Army for the best general researcher, which is a technology-based company and his innovative products received national and international recognition. Dr. Ovalle has filed a number of patents and industrial designs on his innovative ideas. His research interests include data mining, artificial intelligence, image/signal processing, bibliometrics, and pattern recognition. He can be contacted by email [dovalle@utp.edu.pe](mailto:dovalle@utp.edu.pe).