


Nondestructive detection of mango soluble solid content in hyperspectral imaging based on multi-combinatorial feature wavelength selection

J.J. Lin¹, Q.H. Meng^{1,2*} , Z.F. Wu^{1,2}, S.Y. Pei^{1,2}, P. Tian¹, X. Huang¹, Z.Q. Qiu¹, H.J. Chang¹, C.Y. Ni¹, Y.Q. Huang³ and Y. Li⁴

¹ School of Physics and Electronics, Nanning Normal University, Nanning 530001, China

² Key Laboratory of New Electric Functional Materials of Guangxi Colleges and Universities, Nanning Normal University, Nanning 530001, China

³ Key Laboratory of Environmental Evolution and Resource Utilization of the Beibu Gulf, Ministry of Education & Guangxi Key Laboratory of Earth Surface Processes and Intelligent Simulation, Nanning Normal University, Nanning 530001, China

⁴ Guangxi Technical Instruction Office for Fruit, Nanning 530022, China

ORIGINAL RESEARCH PAPER

Received: January 20, 2023 • Accepted: June 26, 2023

Published online: August 9, 2023

© 2023 Akadémiai Kiadó, Budapest



ABSTRACT

This paper explores the prediction of the soluble solid content (SSC) in the visible and near-infrared (400–1,000 nm) regions of Baise mango. Hyperspectral images of Baise mangoes with wavelengths of 400–1,000 nm were obtained using a hyperspectral imaging system. Multiple scatter correction (MSC) was chosen to remove the effect of noise on the accuracy of the partial least squares (PLS) regression model. On this basis, the characteristic wavelengths of mango SSC were selected using the competitive adaptive reweighted sampling (CARS), genetic algorithm (GA), uninformative variable elimination (UVE), and combined CARS + GA-SPA, CARS + UVE-SPA, and GA + UVE-SPA characteristic wavelength methods.

* Corresponding author. E-mail: mqhgx@163.com

The results show that the combined MSC-CARS + GA-SPA-PLS algorithm can reduce redundant information and improve the computational efficiency, so it is an effective method to predict the SSC of mangoes.

KEYWORDS

hyperspectral imaging, mango, nondestructive, variable selection, soluble solids content (SSC), partial least squares (PLS)

1. INTRODUCTION

Baise, Guangxi, China, has a subtropical monsoon climate. The special climatic conditions and geographical environment give mangoes their unique flavour. Baise mango has the characteristics of a small core, rich in nutrients, and less fibre, so it is loved by people. The soluble solid content (SSC) is an important indicator of the internal quality of mangoes (Gao et al., 2021), which allows one to determine the time of harvest, assess and grade the post-harvest quality of mangoes. Traditional methods to test the internal quality of fruits include chemical analysis (Weingerl and Unuk, 2015). This method of chemical analysis and testing, which destroys the appearance of fruit samples, is a time-consuming and inefficient process that cannot achieve real-time online detection. Therefore, there is an urgent need for a fast, non-destructive, and real-time method to nondestructively inspect the internal quality of mangoes.

Hyperspectral imaging (HSI) techniques can acquire both image information and spectral information about a sample to determine the internal quality and external quality, respectively. Hyperspectral imaging technology has advantages in terms of simple operation, fast and accurate detection, and nondestructive environmental protection. So it is widely used to nondestructively inspect the internal quality of fresh fruits (Sun et al., 2020; He et al., 2021). Li, L.S. et al. (2020) achieved the calibration of spectral data across samples by transferring calibration models using the existing NIR spectral data of mango and apple standard samples. A wavelength selection method based on the calibration transfer of standard samples was proposed. By comparing the predictive performance of models before and after calibration transfer, the wavelength with better predictive performance for quality properties of mango and apple was selected. Cortés et al. (2016) proposed a new comprehensive index to evaluate the internal quality of mango more comprehensively. Besides, they developed a prediction model to predict the internal mass index of mangoes by measuring the reflected light from the surface of mangoes. Rungpichayapichet et al. (2017) studied the physical and chemical properties of several mangoes to achieve a predictive mapping of the physical and chemical properties of mangoes. Moreover, the visualisation of the spatial distribution on the fruit surface also can be demonstrated. However, the current study suffers from a lack of sample size and diversity; a lack of detailed description of the sample selection, model building and validation process, and a discussion of the wavelength range and feature selection limit the comprehensive evaluation of the method. These steps are essential for building accurate and stable prediction models.

This paper aims to study mangoes after harvesting by using hyperspectral imaging system combined with mango SSC content. The spectral information of mangoes was collected, the effects



of different characteristic wavelength screening methods on model accuracy and stability were compared, the optimal wavelength combination that can characterise the SSC content of mangoes was identified, and the optimal prediction model of spectral information and SSC content of mangoes was established to provide a theoretical basis for high-quality harvesting of mangoes.

2. MATERIALS AND METHODS

2.1. Mango samples

The samples in the experiment were harvested from Baise mango orchards in Guangxi Zhuang Autonomous Region. To nondestructively detect the mango soluble solid content by hyperspectroscopy, 134 mango samples were purchased, which were free from any mutilation, disease, or damage and of uniform size. There were no significant differences in fruit surface colour, but data measured by a digital soluble solids meter indicated that individual fruits were not uniformly ripe. After the mangoes had been brought to the laboratory for labelling (m1–m134), hyperspectral images and SSC were acquired after 24 h at room temperature.

2.2. Hyperspectral image acquisition

The hyperspectral image data were acquired using a hyperspectral imaging system that consisted of a headwall Micro-Hyperspec VNIR A hyperspectral imager from the United States, a 300-W halogen lamp, and a movable head, as shown in Fig. 1.

The hyperspectral camera swings horizontally through the head to photograph the sample. The sample position is roughly 30 cm directly in front of the lens. In total, 327 bands of hyperspectral images in the spectral range between 400 and 1,000 nm were acquired in this

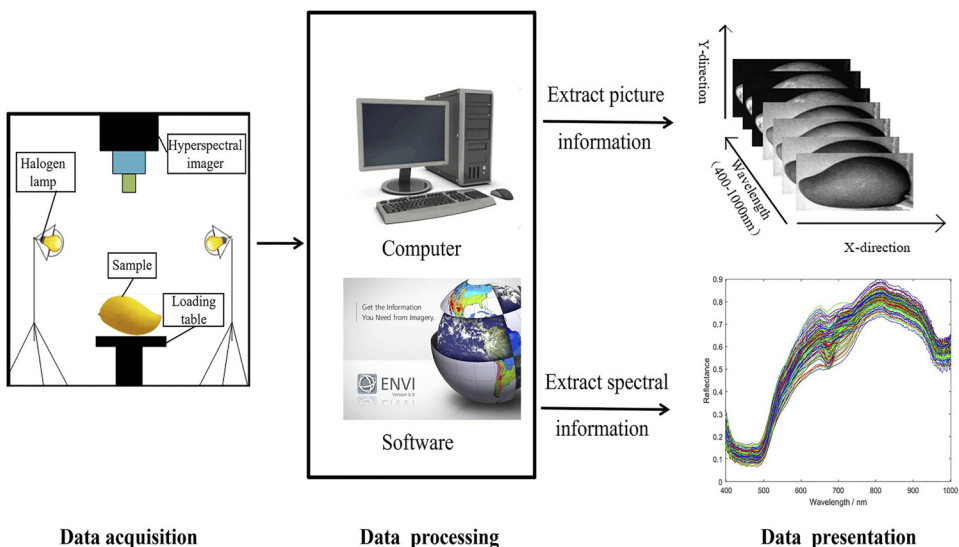


Fig. 1. Hyperspectral image acquisition



experiment. To reduce the effect of dark current noise and light from the hyperspectral imaging system to produce noise in the image, black and white correction of the hyperspectral image was required before sample acquisition. The correction was performed using the white reference image I_{white} and black reference image I_{dark} , which were obtained from a standard whiteboard scan and a light-free coverage lens, respectively. Theoretically, the corrected image R is transformed from the original hyperspectral image I according to the following equation:

$$R = \frac{I - I_{\text{dark}}}{I_{\text{white}} - I_{\text{dark}}} \times 100\% \quad (1)$$

2.3. Extracting spectral and image data

The mangoes were numbered and marked in order, and the areas at the top, middle, and bottom of the mangoes were marked as mango soluble solid test sites. After opening the original spectral image with ENVI software and extracting the original spectral data in a pixel square with a region size of 10×10 , the average spectral data of the region were manually selected and extracted. After the corresponding spectral and image information had been extracted and saved, the MATLAB R2018b software was used to perform spectral data modelling and the original segmentation of the image data.

2.4. Reference measurements

To evaluate the effect of the internal quality of the fruit on the spectrum, the SSC was measured by the portable digital refractometer PAL-1. Measurements were taken three times separately, and the average value was calculated as the reference value of soluble solids of mango samples. We could classify the fruits into ripeness sets based on SSC and study the ripeness classification.

2.5. Data pretreatment

The original dataset contained 134 samples, which were divided into two independent datasets by the KS algorithm. Table 1 shows the minimum, maximum, mean, standard deviation (SD), and coefficient of variation (CV) of the quality attributes for the calibration and prediction sets. Since spectral data are easily disturbed by light, noise, baseline drift, and other factors, the raw data must be preprocessed (Mishra et al., 2019). We have used five methods of multivariate scattering correction (MSC), standard normal variables transformation (SNV), normalisation (Nor), moving-average (MA) method, and Savitzky-Golay convolution smoothing (SG) for preprocessing. A comparative analysis of various preprocessing methods revealed that the MSC (Helland et al., 1995) processing method had the best effect, as shown in Table 2. So MSC is finally chosen for data preprocessing in this experiment.

Table 1. Measurement results statistics of the mango soluble solid form (Unit/Brix)

Sample set	Number	Min	Max	Average	SD	CV
Calibration set	101	12.9	18.7	14.648	1.538	0.105
Prediction set	33	11.0	19.5	15.282	1.968	0.129
Total	134	19.5	11.0	14.831	1.674	0.113



Table 2. Statistics of the results of different pretreatment methods

Quality attribute	Preprocessing	Index cmp	R ² c	RMSEC	R ² p	RMSEP	RPD
SSC(%)	SNV	7	0.8252	0.6430	0.7546	0.9747	2.0186
	MSC	10	0.9026	0.4800	0.7491	0.9855	1.9964
	Nor	8	0.8340	0.6267	0.7254	1.0311	1.9083
	MA	10	0.7950	0.6963	0.6511	1.1621	1.6929
	SG	9	0.8007	0.6867	0.6869	1.1010	1.7871

2.6. Data analysis methods

Hyperspectral images contain irrelevant spectral noise information and information to the detection metrics, which may hinder the predictive performance and application of a particular model; therefore, feature wavelength extraction (Weng et al., 2020) is required to simplify the model and improve the model accuracy. Various feature wavelength extraction methods have been proposed in the past few years to reduce the data dimensionality and increase the computational speed (Wang et al., 2018). The CARS, UVE (Liu et al., 2018), GA and continuous projection algorithm (SPA) (Fan et al., 2015) methods were used to select the optimal solution by combining and re-extracting four feature wavelength extraction methods. CARS is suitable for multivariate analysis and pattern recognition tasks, which can reduce the dimensionality of wavelength subsets and improve modelling speed and accuracy; GA can handle high-dimensional and nonlinear problems with high global search capability; UVE is simple to use, fast to compute, and applicable to a variety of data types and models.

The partial least squares regression (PLS) model (Zhang et al., 2018) was used to select fewer new variables to replace the original larger number of variables without losing the main spectral information, which solved the difficulty of not being able to analyse the spectral bands due to their overlap. PLS regression reveals a linear relationship between the spectral variables (X) and the sample properties (Y) (Feng et al., 2013), and the resulting model can be expressed as:

$$Y = Xb + e \quad (2)$$

where b and e are the regression coefficient and prediction error, respectively. The predictive ability of all models was assessed by coefficient of determination (R^2p), root mean square error of prediction (RMSEP) and residual predictive deviation (RPD) for prediction set, respectively. A RPD value between 2.5 and 3 or above corresponds to good and excellent prediction accuracy (Zhang et al., 2020). The overall process is shown in Fig. 2.

3. RESULTS AND DISCUSSION

3.1. Spectral analysis

The average spectral reflectance curves of the original spectral data of Baise mango in the range of 400–1,000 nm are shown in Fig. 3A, and the average spectral reflectance curves in the range of 400–1,000 nm after the MSC pretreatment are shown in Fig. 3B. The results show that all samples exhibited similar trends in spectral profiles, and the differences in the visible regions were attributed to differences in colour of the samples. The broad absorption



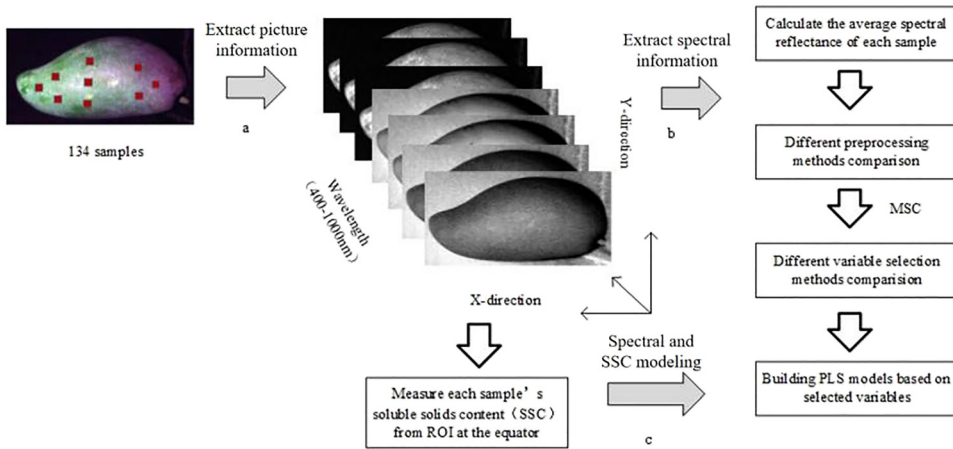


Fig. 2. Hyperspectral image analysis process

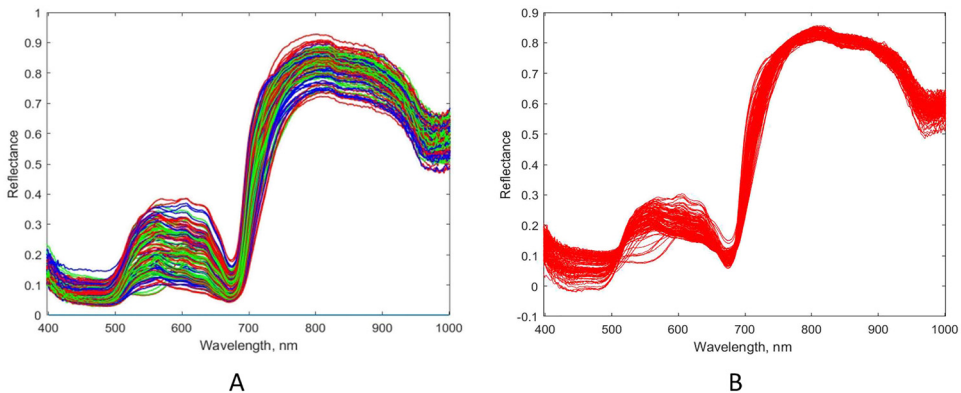


Fig. 3. Spectra of all mangoes in the wavelength range of 400–1,000 nm: A: original reflectance spectra and B: spectra after MSC preprocessing

bands at approximately 680 nm and 740 nm in Baise mangoes may be related to changes in scattering caused by changes in water and carbohydrates or changes in tissue structure (Pu et al., 2016). The absorption peak of soluble solids is at the band 940–947 nm, which is a tertiary multiplicative characteristic absorption peak of the C–H group (Li, P. et al., 2020). The absorption peak appearing at 970–980 nm is mainly related to the water content of the mango, and this band is a secondary multiplicative characteristic absorption peak of the O–H group (Cortés et al., 2016). Since each compound contains multiple chemical bonds, we could not directly observe the specific wavelengths associated with the SSC values. The hidden relationships between SSC and spectra must be mined and expressed using data analysis methods.



3.2. PLS models based on selected wavelengths

3.2.1. Effective variable selection by competitive adaptive reweighted sampling (CARS). CARS is a sampling algorithm for solving optimisation problems. It is based on genetic algorithms and adaptive reweighting strategies for generating excellent sets of sampled solutions for further analysis and optimisation. We use interactive verification to select the subset with the lowest root mean square error cross validation (RMSECV) index.

The number of characteristic spectral variables rapidly decreased with increasing sampling times and subsequently smoothly decreased, as shown in Fig. 4A. When the number of samples increased, the RMSECV slowly decreased and subsequently steeply increased, as shown in Fig. 4B. Figure 4C shows the path diagram of the regression coefficients of the characteristic spectral variables with the number of samples. When the RMSECV value reaches the minimum value in Fig. 4B, the regression coefficients of each characteristic spectral variable are located at the position of the vertical line marked by “*” in Fig. 4C. When RMSECV = 0.5482, which was the lowest value, 16 characteristic spectral variables were extracted, which accounted for 4.9% of the full band.

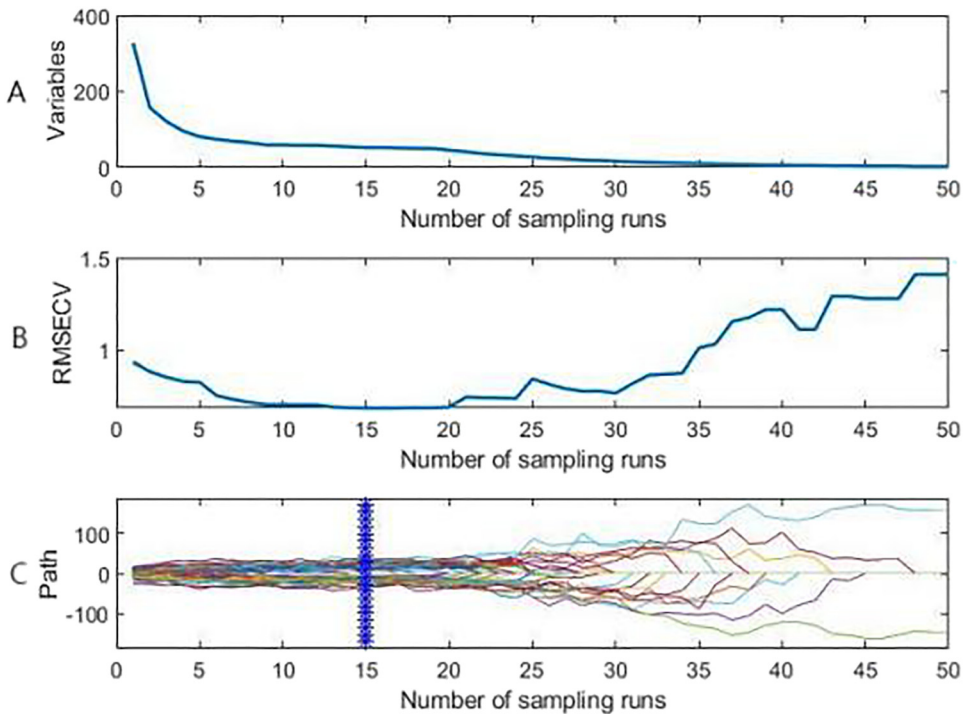


Fig. 4. CARS characteristic wavelength selection figure of mango soluble solids: A: number of sampling variables; B: RMSECV values; C: regression coefficient path



3.2.2. Effective variable selection by genetic algorithm (GA). Genetic algorithm is an optimisation algorithm that simulates the natural evolutionary process. By simulating the mechanisms of genetics and natural selection, optimal solutions or near-optimal solutions are evolved generation by generation from an initial set of candidate solutions. It can be used to optimise the topological weight threshold and has strong robustness.

During the GA operation, the initial population was set to 30, the crossover rate was 50%, the variation rate was 1%, and the number of iterations was 100. Using the minimum RMSECV value as the criterion, the RMSECV variation graph is shown in Fig. 5B. The wavelength points that more frequently appeared in the iterative process were selected as the optimal wavelength points, and 38 characteristic wavelength points were finally selected, as shown in Fig. 5A, which accounted for 11.6% of the original spectrum.

3.2.3. Effective variable selection by uninformative variable elimination (UVE). The goal of the UVE method is to exclude irrelevant variables from a given feature set that do not contribute to the target variable. Reducing irrelevant variable variables improves the effectiveness and explanatory power of the model and reduces the risk of overfitting.

In the UVE algorithm, 99% of the absolute value of the maximum stability at the noise matrix is used as the rejection threshold. As shown in Fig. 6, the left curve represents the stability values of the spectral variables, the right curve represents the stability values of the noise variables, and the two horizontal dashed lines are the selection thresholds of the variables (± 28.74). The inside of the dashed line is useless information, and the outside is useful information. Finally, 36 characteristic wavelengths were selected, which accounted for 11.0% of the original spectrum.

The characteristic wavelengths extracted by the three methods have common wavelengths at around 500, 680, 840, and 960 nm. Among them, 500 nm is associated with the presence of carotenoids; the low reflectance around 680 nm indicates high absorbance in this region, absorbing red pigments, mainly due to chlorophyll, which gives the fruit its characteristic green colour (Pu et al., 2016); there is a smaller absorption peak at 840 nm and a significant absorption peak out of 960 nm, which is caused by the absorption effect of the O–H group vibration of carbohydrates and water at all levels (Mishra et al., 2021).

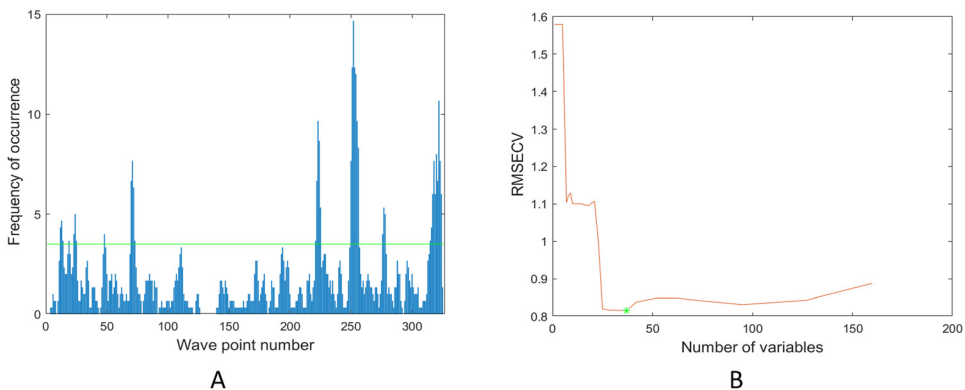


Fig. 5. GA characteristic wavelength selection diagram of mango soluble solids: A: GA screening diagram; B: RMSECV change diagram



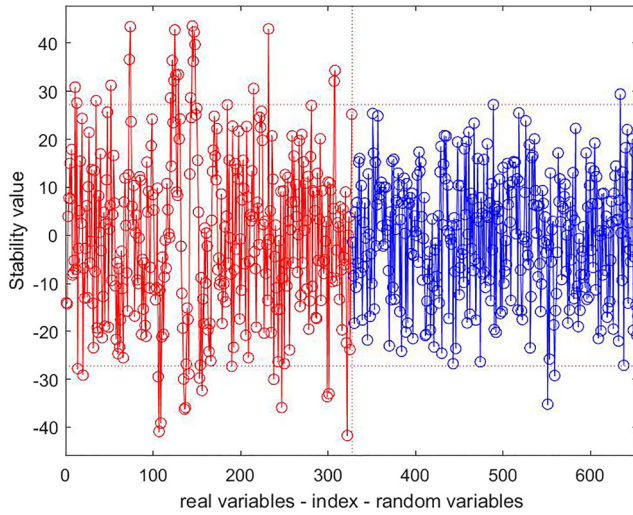


Fig. 6. Selection of UVE features of mango soluble solids

3.3. Effective variable selection by multiple combinations

CARS performed the best among the three feature wavelengths and selected the lowest number of feature wavelengths with $R^2_p = 0.8504$, $RMSEP = 0.7611$, and $RPD = 2.5854$ for the prediction set when only one feature wavelength extraction method was used, respectively, as shown in Table 3. Among them, the multiple combination feature wavelength extraction method CARS + GA-SPA-PLS model had the best performance with $R^2_p = 0.8798$, $RMSEP = 0.6822$, and $RPD = 2.8843$ for the SSC prediction set. The selected feature wavelengths decreased from 54 to 10, which improved the model fitting speed and efficiency. In summary, the CARS + GA-SPA algorithm has certain

Table 3. Effects of the PLS model established based on different feature wavelength extraction methods

Variable selection methods	No. of variables	LVs	Calibration set		Prediction set		
			R^2_c	RMSEC	R^2_p	RMSEP	RPD
CARS	16	9	0.8730	0.5482	0.8504	0.7611	2.5854
GA	38	9	0.8044	0.6415	0.7145	1.0501	1.8715
UVE	36	9	0.8852	0.4320	0.7477	1.0539	1.9908
CARS-SPA	10	10	0.7968	0.6934	0.5569	1.3097	1.5022
GA-SPA	10	9	0.8047	0.6797	0.6735	1.1243	1.7501
UVE-SPA	10	12	0.7784	0.7240	0.7866	0.9088	2.1647
UVE + CARS-SPA	10	9	0.8964	0.4951	0.8747	0.6964	2.8250
GA + CARS-SPA	10	9	0.8991	0.4886	0.8798	0.6822	2.8843
GA + UVE-SPA	10	11	0.8814	0.5297	0.8258	0.8213	2.3959
UVE + CARS	52	9	0.8860	0.5194	0.8739	0.6988	2.8160
GA + CARS	54	9	0.8885	0.5137	0.8116	0.8540	2.3038
GA + UVE	74	9	0.8724	0.5495	0.8257	0.8214	2.3953



advantages over CARS + GA in feature wavelength selection. It is good to help simplify the prediction model and apply HSI to the online quality inspection of mangoes.

4. CONCLUSIONS

Combining other studies we found that using one feature wavelength algorithm may ignore information from other wavelengths, thus causing the problem of incomplete information. Also it may not be able to distinguish between different substances that may be spectrally very similar at certain wavelengths. Even the use of inappropriate feature wavelength algorithms may lead to inaccurate feature wavelength extraction, which affects the prediction accuracy. Multi-combination feature wavelength extraction fuses information from different wavelengths, which can improve the prediction accuracy, especially for complex sample data, such as the prediction of sugar content of mangoes. It can also reduce the overfitting phenomenon of the model, which can improve the generalization ability of the model and make the model can be better adapted to new data.

The results showed that the multiple combination feature wavelength extraction method MSC-CARS + GA-SPA-PLS was the optimal prediction method for detecting soluble solids in mangoes, and the number of feature wavelengths selected was 10, with $R^2_p = 0.8798$, RMSEP = 0.6822, and RPD = 2.8843 for the SSC prediction set. Therefore, the multi-feature wavelength fusion algorithm is a novel and pioneering algorithm, which can improve the accuracy and generalisation of hyperspectral NDT and has a wide application prospect. Meanwhile, the results show that it is feasible to predict the soluble solids content of mangoes based on hyperspectral imaging, which can provide a reference for the real-time monitoring of mango quality using spectral imaging technology. The model proposed in this paper is based on hyperspectral images of the top, middle, and bottom of the mango. Since the mango is an inhomogeneous ellipsoid with nonuniform illumination throughout the sample, visualising the SSC distribution on the surface remains challenging.

ACKNOWLEDGEMENT

This work is supported by the Guangxi Scientific and Technological Project (No. Guike AD20238059), Baise Hi tech Industrial Zone Guidance Project (No. K-YS-ST-2018-01), Guangxi Degree and Postgraduate Education Reform Project (No. JGY2022220) and Demonstrative Modern Industrial School of Guangxi University – Smart Logistics Industry School Construction Project, Nanning Normal University (No. 6020303891823).

REFERENCES

- Cortés, V., Ortiz, C., Aleixos, N., Blasco, J., Cubero, S., and Talens, P. (2016). A new internal quality index for mango and its prediction by external visible and near-infrared reflection spectroscopy. *Postharvest Biology and Technology*, 118: 148–158.



- Fan, S., Huang, W., Guo, Z., Zhang, B., and Zhao, C. (2015). Prediction of soluble solids content and firmness of pears using hyperspectral reflectance imaging. *Food Analytical Methods*, 8(8): 1936–1946.
- Feng, Y.Z. and Sun, D.W. (2013). Near-infrared hyperspectral imaging in tandem with partial least squares regression and genetic algorithm for non-destructive determination and visualization of *Pseudomonas* loads in chicken fillets. *Talanta*, 109: 74–83.
- Gao, Q., Wang, P., Niu, T., He, D.J., Wang, M.L., Yang, H.J., and Zhao, X.Q. (2021). Soluble solid content and firmness index assessment and maturity discrimination of *Malus micromalus Makino* based on near-infrared hyperspectral imaging. *Food Chemistry*, 370: 131013.
- He, Y., Xiao, Q.L., Bai, X.L., Zhou, L., Liu, F., and Zhang, C. (2021). Recent progress of nondestructive techniques for fruits damage inspection: a review. *Critical Reviews in Food Science and Nutrition*, 62(20): 5476–5494.
- Helland, I.S., Nas, T., and Isaksson, T. (1995). Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, 29(2): 233–241.
- Li, L.S., Jang, X.G., Li, B., and Liu, Y.D. (2020). Wavelength selection method for near-infrared spectroscopy based on standard-sample calibration transfer of mango and apple. *Computers and Electronics in Agriculture*, 190: 106448.
- Li, P., Li, S.K., Du, G.R., Jiang, L.W., Liu, X., Ding, S.H., and Shan, Y. (2020). A simple and nondestructive approach for the analysis of soluble solid content in citrus by using portable visible to near infrared spectroscopy. *Food Science & Nutrition*, 8(5): 2543–2552.
- Liu, Y.D., Xiao, H.C., Sun, X.D., Zhu, D.N., Han, R.B., Ye L.Y., Wang, J.G., and Ma, K.R. (2018). Spectral feature selection and discriminant model building for citrus leaf Huanglongbing. *Transactions of the Chinese Society of Agricultural Engineering*, 34(3): 180–187. (In Chinese with English abstract).
- Mishra, P., Karami, A., Nordon, A., Rutledge, D.N., and Roger, J.M. (2019). Automatic denoising of close-range hyperspectral images with a wavelength-specific shearlet based image noise reduction method. *Sensors and Actuators B: Chemical*, 281: 1034–1044.
- Mishra, P., Woltering, E., Brouwer, B., and van Hogeveen, E.E. (2021). Improving moisture and soluble solids content prediction in pear fruit using near-infrared spectroscopy with variable selection and model updating approach. *Postharvest Biology and Technology*, 171: 111348.
- Pu, H.B., Liu, D., Wang, L., and Sun, D.W. (2016). Soluble solids content and pH prediction and maturity discrimination of lychee fruits using visible and near infrared hyperspectral imaging. *Food Analytical Methods*, 9(1): 235–244.
- Rungpichayapichet, P., Nagle, M., Yuwanbun, P., Khuwijitjaru, P., Mahayothee, B., and Müller, J. (2017). Prediction mapping of physicochemical properties in mango by hyperspectral imaging. *Biosystems Engineering*, 159: 109–120.
- Sun, X.D., Subedi, P., and Walsh, K.B. (2020). Achieving robustness to temperature change of a NIRS-PLSR model for intact mango fruit dry matter content. *Postharvest Biology and Technology*, 162: 111117.
- Wang, L.L., Lin, Y.W., Wang, X.F., Xiao, N., Xu, Y.D., Li, H.D., and Xu, Q.S. (2018). A selective review and comparison for interval variable selection in spectroscopic modeling. *Chemometrics and Intelligent Laboratory Systems*, 172: 229–240.
- Weingerl, V. and Unuk, T. (2015). Chemical and fruit skin colour markers for simple quality control of tomato fruits. *Croatian Journal of Food Science and Technology*, 7(2): 76–85.
- Weng, S., Guo, B., Tang, P., Yin, X., Pan, F., Zhao, J., Huang, L., and Zhang, D. (2020). Rapid detection of adulteration of minced beef using Vis/NIR reflectance spectroscopy with multivariate methods. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy*, 230: 118005.



- Zhang, C., Liu, F., and He, Y. (2018). Identification of coffee bean varieties using hyperspectral imaging: influence of preprocessing methods and pixel-wise spectra analysis. *Scientific Reports*, 8(1): 2166.
- Zhang, H.L., Zhan, B.S., Pan, F., and Luo, W. (2020). Determination of soluble solids content in oranges using visible and near infrared full transmittance hyperspectral imaging with comparative analysis of models. *Postharvest Biology and Technology*, 163: 111148.

