ONLINE DETECTION OF SOLUBLE SOLID CONTENT IN FRESH JUJUBE BASED ON VISIBLE / NEAR-INFRARED SPECTROSCOPY
/
基于可见/近红外光谱的鲜枣可溶性固形物在线检测

Bin WANG [1)]; Lili LI [*1)] [1]
[1)] College of Information Science and Engineering, Shanxi Agricultural University, Taigu 030800/China
Tel: 18306832356; E-mail: lilycqdxyjs@163.com
DOI: https://doi.org/10.35633/inmateh-72-27

Keywords: *near infrared spectroscopy, soluble solid content, characteristic wavelengths, partial least squares regression, nondestructive testing*

ABSTRACT
Soluble solid content (SSC) is one of the important evaluation indexes of the internal quality and taste of fresh jujube. In order to realize the online nondestructive detection of SSC of fresh jujube, this paper took Huping jujube as the research object, adopted self-constructed nondestructive online testing system to collect the spectral information of jujubes (350~2500 nm), and studied the influence of the rotational speed of 4 r/min on the online prediction model of SSC of jujube. Kennard-Stone (KS) algorithm was used to divide the sample into correction set and prediction set. Six commonly used preprocessing methods such as SG smoothing (S-G), multiplicative scatter correction (MSC), standard normal variate (SNV), orthogonal signal correction (OSC), first derivative (FD), and second derivative (SD) were applied to the spectral data, and the regression coefficient (RC) algorithm and the successive projections algorithm (SPA) were utilized to select informative wavelengths, and a quantitative prediction model for the SSC of Huping jujube was established using partial least squares regression (PLSR). The results indicate that the PLSR prediction model established by preprocessing the original spectrum with OSC and combining it with RC algorithm to select characteristic wavelengths was optimal. Therefore, when predicting the SSC of Huping jujube, the optimal model was OSC-RC-PLSR, and the correlation coefficients of the correction set and prediction set were 0.846 and 0.782, respectively, and the corrected root mean square error (RMSEC) and predicted root mean square error (RMSEP) were 1.962 and 2.247, respectively. The results show that non-destructive detection of soluble solid content of jujube can be achieved by combining visible-near-infrared spectroscopy and appropriate regression model, which provides an innovative way for online sorting and identifying fresh jujube.

摘要
可溶性固形物(Soluble Solid Content, SSC)是鲜枣内部品质与口感的重要评价指标之一。为实现鲜枣SSC的在线无损检测，本文以壶瓶枣为研究对象，采用自行搭建的无损在线检测系统采集壶瓶枣的光谱信息（350～2500nm），研究了旋转速度为4 r/min条件下对壶瓶枣SSC在线预测模型的影响。采用Kennard-Stone (KS)算法将样本划分为校正集和预测集，对原始光谱使用SG平滑(S-G)、乘多元散射校正(MSC)、标准正态变量(SNV)、正交信号校正(OSC)、一阶导数(FD)和二阶导数(SD)等6种方法进行预处理，采用RC回归系数法和连续投影算法对原始光谱降维，结合偏最小二乘回归（PLSR）建立壶瓶枣SSC的定量预测模型。结果表明，原始光谱经OSC预处理，再结合RC算法筛选特征波长建立的PLSR预测模型最优。因此在预测壶瓶枣SSC时，最优模型为OSC-RC-PLSR，其校正集和预测集相关系数分别为0.846和0.782，均方根误差分别为1.962和2.247。该研究表明，结合可见-近红外光谱和合适的回归模型，可实现对壶瓶枣可溶性固形物含量的无损检测，为鲜枣品质在线检测提供新的途径。

INTRODUCTION
Huping jujube is not only a nutritious food, but also a natural health product. Its meat is crispy, its taste is sweet, and its flavor is unique, which is deeply loved by consumers *(Fan et al., 2003)*. The soluble solid content (SSC) is an important indicator of the quality of Huping jujube, and the rapid non-destructive testing of SSC helps to detect and classify the quality of Huping jujube *(Fan et al., 2015)*.

[1] *Bin Wang, Lec. Ph.D.; Lili Li*, Lec. Ph.D.*

The traditional methods for measuring SSC content have drawbacks such as destructive, inefficient, and time-consuming, and are difficult to meet the needs of online detection of large batches of samples. The classification of fresh jujube according to quality is indispensable in the post-production process of fresh jujube, which has important significance for the storage of fresh jujube products and the "price according to quality" in the sales process. Therefore, how to realize the fast and nondestructive online detection of fresh jujube SSC is particularly important.

In recent years, near infrared spectroscopy has been used to detect the quality of apples, strawberries, oranges, pears, watermelons, which is a very effective and economical sorting method *(Agulheiro et al., 2022; Song et al., 2020; Wang et al., 2020)*. Scholars at home and abroad have conducted many studies on online detection of near-infrared red spectroscopy. *Jiang et al. (2023)* studied the influence of different parameters (motion speed, integration time, and light intensity) on the prediction of the apple SSC model by near infrared spectrum based on the online device of near infrared. The study showed that the standard normal variate (SNV) and competitive adaptive reweighting sampling (CARS) and partial least squares (PLS)
(SNV-CARS-PLS) model established had the best performance when the motion speed of the device was 0.5 m/s, integration time was 120 ms, and light intensity was 6.5 A. Its predicted correlation coefficient (Rp) and the prediction root mean square error (RMSEP) were 0.991 and 0.149, respectively. *Liu et al. (2022)* established an online non-destructive testing equipment for apples using near-infrared spectroscopy and established a prediction model for apple SSC. When the detection speed and integration time were 0.5 m/s and 100 ms, respectively, the Rp reached 0.919 and the RMSEP was 0.477. *Ding et al. (2020)* established a non-destructive online detection system for potatoes using visible/near-infrared diffuse transmission spectroscopy (with a detection speed of approximately 4 per second). The correlation coefficient between the predicted starch value and the standard physicochemical value was 0.893, and the partial least squares regression (RMSE) was 0.713%. *Jiang et al. (2023)* established an online detection system for navel orange SSC using near-infrared transmission method. The research results showed that when the detection speed was 0.5 m/s, the partial least squares regression (PLSR) prediction model established had the best performance, with RMSEP and residual prediction error (RPD) of 0.442% and 2.77%, respectively. However, there are few reports on establishing an online detection model using visible/near-infrared spectroscopy technology to detect the internal quality of Huping jujube.

This study selected Huping jujube in Jinzhong City as the research object, dynamically collected its visible/near-infrared spectral data, and the actual SSC values were obtained by handheld refractometer. Comparing the optimal selection of modeling results of different pretreatment methods, the regression coefficient (RC) and successive projections algorithm (SPA) algorithms were used to reduce the dimensionality of the pre-treated full-band spectral data, further discussing the influence of spectral variable selection methods on the accuracy of SSC online detection model, and evaluating the corresponding model prediction effect. The feasibility of applying RC-PLSR model to the prediction of SSC of Jujube was verified, and the reference was provided for the realization of fast, scientific and accurate online SSC detection of fresh jujube.

MATERIALS AND METHODS

Experimental Sample

The experimental samples selected for this study were picked from a jujube garden in Beizhan Village, Taigu County on October 4, 2023, and 120 jujube dates were picked and shipped to the laboratory on the same day, as shown in Fig. 1. The samples with no external defects, rot and injury, and basically the same physical properties were selected as the research objects.

Fig. 1 - Intact samples of Huping jujube

Experimental System and Data Acquisition

This study used the Field Spec3 spectrometer produced by ASD (Analytical Spectral Device) in the United States and a self-developed dynamic spectral collection system to achieve spectral collection of fresh jujube samples. The schematic diagram of the dynamic spectrum acquisition system is shown in Fig. 2. The spectral data collection interval is 1 nm, the wavelength range is 350-2500 nm, the resolution is 3.5 nm, the probe field of view angle is 20°, and the light source is a 14.5 V halogen lamp. The probe of the spectrometer is perpendicular to the upper surface of the sample, 90 mm away from the upper surface of the sample, and the sample is placed between two rollers. To minimize errors, the system configuration optimization and whiteboard calibration were performed 0.5 hours after the spectrometer was turned on. After passing the performance test, the sample spectrum was sampled using diffuse reflection method. During data collection, the fresh jujube sample was rotated at the speed of 4 r/min, and the spectral data of the sample was collected once every 120°, for a total of 3 times, and the average value was obtained as the final spectral data of the test sample.

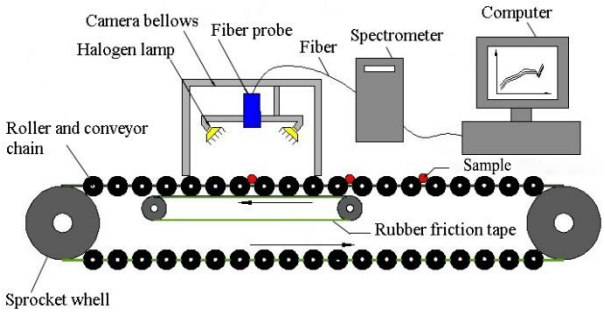


Fig. 2 - Schematic diagram of online Vis-NIR spectroscopy detection device

Determination of Soluble Solid Content

A handheld digital refractometer was used to measure the SSC value of the sample immediately after the collection of the visible/near infrared spectrum. A piece of flesh with skin was cut from the spectral scanning position, and the juice was manually squeezed and filtered to the center of the mirror of the refractometer to read the SSC value. The refractometer needs to be calibrated with distilled water for each measurement, and the average value of each sample was repeated 3 times as the true value of SSC.

Kennard-Stone (KS) method can ensure the uniform distribution of the samples in the training set according to the spatial distance, and improve the stability and accuracy of the prediction model *(Galvao et al., 2005)*. The K-S method was used to divide the correction set and prediction set according to the ratio of 3:1, and the results of sample set division was shown in Table 1.

Table 1

True value tables of intact Huping jujube sample soluble solids for calibration and prediction sets

Sample set	No. of sample	Min. value	Max. value	Average value	Standard deviation	Coefficient of variation (%)
Correction set	90	17.3	37.2	22.15	2.21	9.98
Prediction set	30	18.6	34.6	21.82	1.85	8.48

From Table 1, it can be seen that the SSC range of the correction set was 17.3~37.2, and the SSC range of the prediction set was 18.6~34.6. The SSC range of the correction set covers the prediction set well, which was helpful to establish a stable and effective prediction model.

Data Analysis and Processing Methods

Spectral Pretreatment

In addition to the internal structure and component information of the measured sample, the original NIR spectral data contains interference information caused by light scattering, sample heterogeneity, temperature change, instrument noise and other factors. Therefore, it is necessary to preprocess the original spectrum to improve the stability of the prediction model. In the present work, six spectral pre-processing methods, including S-G smoothing (S-G), multiplicative scatter correction (MSC), standard normal variate (SNV), orthogonal signal correction (OSC), first derivative (FD), and second Derivative (SD) to preprocess the

original spectra *(Liu et al., 2022)*, and PLSR was used to evaluate the performance of different spectral pre-processing data.

Method for Selecting Characteristic Wavelengths

Due to the large amount of wavelength information, continuous information and redundant information contained in the original full spectrum data collected, the correlation between adjacent bands is very strong. The optimal selection of effective wavelength can eliminate redundant information in spectral data, simplify the calculation of spectral data, and improve the efficiency and stability of modeling. In this study, the RC and SPA algorithm were used to reduce the dimensionality of the data to simplify the model.

The PLSR method is based on the PLS algorithm principle and selects the best sensitive band by selecting the local extreme value in the RC *(Wang et al., 2023)*. The greater the absolute value of the regression coefficient corresponding to each wavelength point, the greater the influence of the wavelength on the prediction performance of the model. Therefore, the characteristic wavelength can be extracted according to the local extreme value of the regression coefficient corresponding to the wavelength.

The SPA algorithm is a forward variable selection algorithm that can eliminate the collinearity effect between wavelength data, extract wavelength subsets with the lowest redundancy and collinearity, select a small number of SSC sensitive information variable groups from a large amount of spectral information, reduce model input, and improve modeling speed and efficiency *(Wang et al., 2021)*.

Modeling Methods

PLSR is the most commonly used multivariate statistical method in stoichiometric modeling, which is based on principal component regression. The main idea of PLSR algorithm is to first perform principal component operations on the spectral data matrix, obtain principal factors or hidden variables, eliminate non-useful information in the data matrix, and select useful spectral data information for parameter regression operations.

Model Evaluation

Corrected correlation coefficient (Rc), predicted correlation coefficient (Rp), corrected root mean square error (RMSEC), and predicted root mean square error (RMSEP) were selected as the evaluation criteria. A model with excellent performance has Rc and Rp closer to 1, and RMSEC and RMSEP closer *(Cozzolino et al., 2007)*.

RESULTS AND DISCUSSION

Spectral Characteristics and Analysis

Due to the influence of external adverse factors in the process of collecting NIR spectral data, there will be large noise, baseline drift and other irrelevant information at both ends of the spectral curve, which will directly affect the stability and prediction ability of the built model. Therefore, only the spectral data in the range of 400~2450 nm spectral band were used for analysis in this experiment, with a total of 2051 spectral bands. Fig. 3 shows the spectral diagram of Huping jujube samples with a rotation speed of 1.5 r/min.

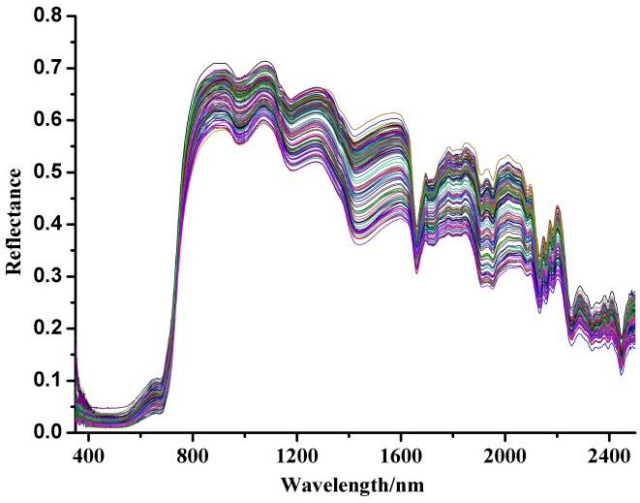


Fig. 3 - Vis-NIR spectral curves of Huping jujube at 1.5 r/min

As can be seen from Fig. 3, the trend of NIR spectral curves of 120 exact jujube samples was basically similar, with no significant differences. From the figure, it can be seen that the absorption peaks at the wavelengths of 960 nm, 1180 nm, 1300 nm, 1590 nm, and 2000 nm, and these absorption peaks at these wavelengths were caused by the absorption of the water contained inside the fresh dates. Among them, the wavelength of 2000 nm was the combined frequency absorption peak of O-H bond, the wavelength of 1300 nm was the primary frequency doubling peak of O-H bond, and the wavelengths of 960 nm and 1180 nm were the secondary frequency doubling peaks of O-H bond *(Liu et al., 2010)*. The wavelength of 2000~2500 nm was the near infrared spectral band of C-H, N-H, and O-H bonds in the soluble solids molecules within the sample of fresh dates, and there was an obvious absorption peak at 680 nm for jujube, which was caused by the spectral absorption of chlorophyll in the fruit pulp cells of jujube.

Spectral Data Preprocessing

From Table 2, it can be seen that the original spectrum uses 13 principal factor numbers, with Rc of 0.647, RMSEC of 2.176, Rp of 0.578, and RMSEP of 1.878. After using OSC processing on the original spectrum, compared with other preprocessing methods, the Rc value of the modeling set using this preprocessing method was 0.876, and the RMSEC was 1.846, which is the smallest. The Rp was 0.763, and the difference between RMSEC and RMSEP was the smallest. The model had good predictive performance and stability, and the OSC was selected as the optimal preprocessing method for comprehensive comparison.

Table 2

Calibration results of PLSR model based on different preprocessing methods

Pretreatment methods	Calibration set		Prediction set		Factors
	Rc	RMSEC	Rp	RMSEP	
Original spectra	0.647	2.176	0.578	1.878	13
S-G	0.742	2.136	0.622	1.864	13
SNV	0.726	2.009	0.584	1.910	9
MSC	0.840	2.068	0.694	2.003	7
OSC	0.876	1.846	0.763	1.792	9
FD	0.752	2.183	0.698	2.013	7
SD	0.801	2.106	0.652	1.896	7

Characteristic Wavelength Extraction

For the full spectral data preprocessed by OSC, the RC and SPA algorithm were used to select the characteristic wavelength, and corresponding 13 characteristic wavelengths (796 nm, 915 nm, 993 nm, 1362 nm, 1463 nm, 1662 nm, 1816 nm, 1953 nm, 1984 nm, 2134 nm, 2198 nm, 2255 nm, and 2415 nm) and 3 characteristic wavelengths (416 nm, 1362 nm, and 2323 nm) were obtained. The analysis results were shown in Fig. 4 and Fig. 5.

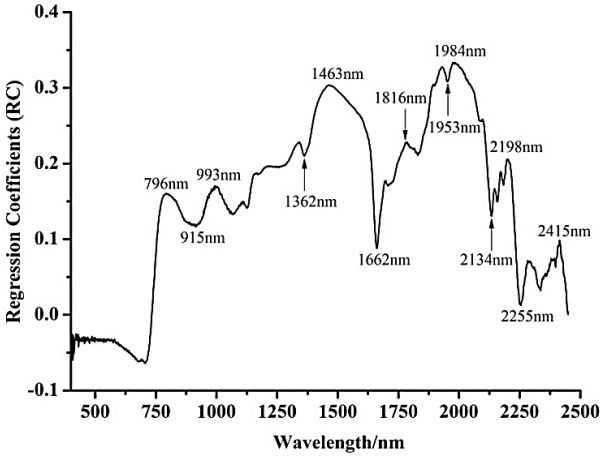
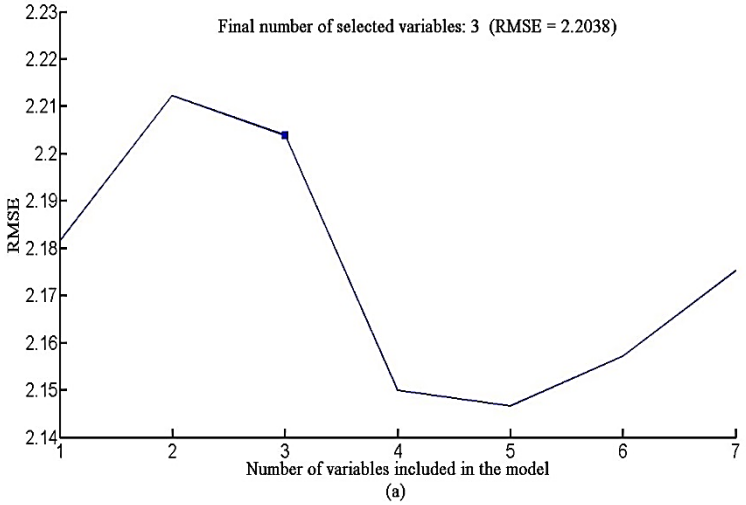
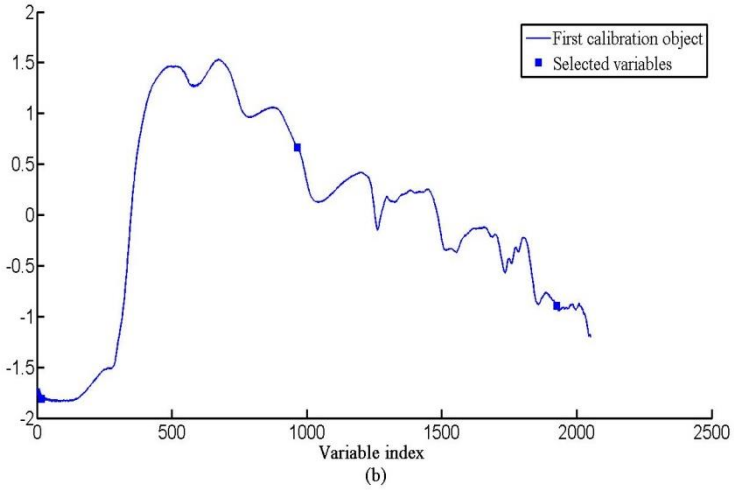


Fig. 4 - Regression coefficients of PLSR model and selected effective wavelength



(a) RMSE distribution of SPA with different variables



(b) Distribution of preferable characteristic wavenumber for SPA

Fig. 5 - Characteristic wavelengths selected by SPA algorithm

Regression model

The characteristic wavelength spectral data optimized by the RC and SPA algorithm were used as input for establishing the PLSR quantitative analysis model, respectively. And obtain the RC-PLSR and SPA-PLSR prediction models. The relevant parameters of the two established prediction models were shown in Table 3.

Table 3

Results of PLSR models based on different characteristic wavelengths

Model	Calibration set		Prediction set		Factors	No. of variables
	Rc	RMSEC	Rp	RMSEP		
RC-PLSR	0.846	1.962	0.782	2.247	11	13
SPA-PLSR	0.739	2.136	0.673	1.762	3	3

As can be seen from Table 3, the number of factors for the RC-PLSR and SPA-PLSR models were 11 and 3, respectively. According to the evaluation of the merits and stability of the model, compared to the SPA-PLSR model with the corrected set (Rc=0.739, RMSEC=2.136) and predicted set (Rp=0.673, RMSEP=1.762), the RC-PLSR model with the corrected set (Rc=0.846, RMSEC=1.962) and predicted set (Rp=0.782, the RMSEP=2.247) had better performance. This may be due to the fact that important spectral information contained in the original spectrum was removed by SPA. Therefore, the prediction performance of the proposed RC-PLSR prediction model for unknown samples was better than the SPA-PLSR prediction model. Fig. 6 shows the scatter plots of SSC predicted and true values of the optimal RC-PLSR prediction model for 30 samples of the prediction set.

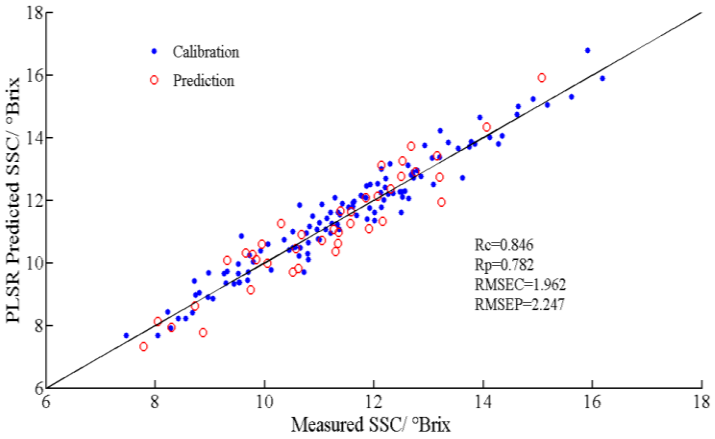


Fig. 6 - Scatter plot of measured and predicted values of SSC index based on the optimal model

CONCLUSIONS

This study used visible/near-infrared spectroscopy combined with RC-PLSR to establish an online detection model for SSC indicators of Huping jujube. By combining multiple preprocessing methods with band selection methods, the prediction accuracy of the model can be effectively improved, and accurate prediction of fresh jujube SSC indicators can be achieved. The results indicate that:

(1) The evaluation of different preprocessing results (Original spectra, S-G, SNV, MSC, OSC, FD, and SD) using PLSR showed that the preprocessed data had higher model accuracy than those directly using raw spectral data as input, and using OSC for preprocessing had the most ideal effect.

(2) Compared with the results of full band modeling, extracting feature wavelengths and establishing a prediction model could reduce computational complexity and improve the prediction accuracy of the model to a certain extent. The RC-PLSR model achieved the optimal prediction performance, with Rp and RMSEP of 0.782 and 2.247, respectively. Research has shown that the RC method can effectively enhance the stability and accuracy of feature wavelength selection in Huping jujube spectral data, improve the prediction accuracy of the model, and dynamically detect the SSC index of fresh jujube based on PLSR and visible/near-infrared spectroscopy was an effective method. At the same time, this study has a reference for the development of more accurate visible/near infrared online nondestructive testing equipment.

ACKNOWLEDGEMENT

This work was funded by Basic Research Project of Shanxi Province (Free Exploration) (Project No. 202203021212426, No. 202303021212120), "Introduction of Talents and Scientific Research Initiation Project" of Shanxi Agricultural University (Project No. 2023BQ42, No. 2023BQ114).

REFERENCES

[1] Agulheiro-Santos, A. C., Ricardo-Rodrigues, S., Laranjo, M., Melgão, C., & Velázquez, R. (2022). Non-destructive prediction of total soluble solids in strawberry using near infrared spectroscopy. *Journal of the Science of Food and Agriculture*, 102(11), 4866-4872.

[2] Cozzolino, D., Kwiatkowski, M. J., Waters, E. J., & Gishen, M. (2007). A feasibility study on the use of visible and short wavelengths in the near-infrared region for the non-destructive measurement of wine composition. *Analytical and bioanalytical chemistry*, *387*(6), 2289-2295.

[3] Ding, J., Han, D., Li, Y., Qi, W., & Xi, H. (2020). Simultaneous non-destructive on-line detection of potato black-heart disease and starch content based on visible/near infrared diffuse transmission spectroscopy. *Spectroscopy and Spectral Analysis*, 40(6), 1909-1915.

[4] Fan, J., Lv, L., & Shang, H. (2003). Progress in research and development of jujube. *Food Science*, *4*(1), 161-163.

[5] Fan, S., Huang, W., Guo, Z., Zhang, B., Zhao, C., & Qian, M. (2015). Assessment of influence of origin variability on robustness of near infrared models for soluble solid content of apple. *Chinese Journal of Analytical Chemistry*, 43(2), 239-244.

[6] Galvao, R. K. H., Araujo, M. C. U., José, G. E., Pontes, M. J. C., Silva, E. C., & Saldanha, T. C. B. (2005). A method for calibration and validation subset partitioning. *Talanta*, 67(4), 736-740.

[7] Jiang, X., Zhu, M., Yao, J., Li, B., & Liao, J. (2023). Research on parameter optimization of apple sugar model based on near-infrared on-line device. *Spectroscopy and Spectral Analysis*, 43(1), 116-121.

[8] Jiang, Z., Ying, J., Wan, Y., Wang, C., Lin, X., & Liu, B. (2023). Non-destructive evaluation of soluble solids content in navel orange by an on-line visible near-infrared system with four parallel spectrometers. *Journal of Food Measurement and Characterization*, 1-11.

[9] Liu, Y., Hu, X., Zhu, M., Yao, J., & Jing, H. (2022). Influence of near-infrared on-line detection device parameters on the applicability of apple soluble solid content model. *Journal of South China Agricultural University*, 43(5), 108-114.

[10] Liu, Y., Sun, X., Zhang, H., & Aiguo, O. (2010). Nondestructive measurement of internal quality of Nanfeng mandarin fruit by charge coupled device near infrared spectroscopy. *Computers and Electronics in Agriculture*, *71*, S10-S14.

[11] Liu, Z., Zhang, R., Yang, C., Hu, B., Luo, X., Li, Y., & Dong, C. (2022). Research on moisture content detection method during green tea processing based on machine vision and near-infrared spectroscopy technology. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 271, 120921.

[12] Song, J., Li, G., Yang, X., Liu, X., & Xie, L. (2020). Rapid analysis of soluble solid content in navel orange based on visible-near infrared spectroscopy combined with a swarm intelligence optimization method. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 228, 117815.

[13] Wang, B., He, J. L., Zhang, S. J., & Li, L. L. (2020). Non-destructive testing of soluble solids content in cerasus humilis using visible/near-infrared spectroscopy coupled with wavelength selection algorithm. *INMATEH Agricultural Engineering*, 61(2), 251-262.

[14] Wang, B., He, J., Zhang, S., & Li, L. (2021). Nondestructive prediction and visualization of total flavonoids content in Cerasus Humilis fruit during storage periods based on hyperspectral imaging technique. *Journal of Food Process Engineering*, 44(10), e13807.

[15] Wang, B., Yang, H., Zhang, S., & Li, L. (2023). Detection of Defective Features in Cerasus Humilis Fruit Based on Hyperspectral Imaging Technology. *Applied Sciences*, 13(5), 3279.