Non-destructive Determination of the Soluble Solid Content in Winter Jujubes Using Hyperspectral Technology and the SCARS-PLSR Prediction Model

Aoran Liu, Yufei Song*, Zheng Xu, Xi Meng, Zhiguo Liu

***Abstract*—Soluble solid content (SSC) is an important index for evaluating the quality of winter jujubes. The purpose of this study was to detect the soluble solid content of winter jujubes in a non-destructive manner. Spectra were collected from mature samples, with SSC values determined by refractometry. Several spectral preprocessing methods, including Multivariate Scattering Correction (MSC), First Derivative (FD), Second Derivative (SD), and Savitzky–Golay (SG), were compared to assess their impact on model accuracy. The MSC – FD – SG preprocessing approach yielded superior results (R = 0.713, RMSE = 1.445). Feature wavelengths were extracted using four methods: Stable Competitive Adaptive Reweighted Sampling (SCARS), Successive Projections Algorithm (SPA), Uninformative Variable Elimination (UVE), and Iteratively Retained Informative Variables (IRIV). Partial Least Squares Regression (PLSR) was employed to establish a prediction model. The SCARS feature selection method exhibited the best performance, with RMSECV values ranging from 0.374 to 0.525 lower than those from the other methods. The SCARS–PLSR model outperformed the full-spectrum model and the SCARS–Random Forest Regression (RFR) model in terms of prediction accuracy, achieving a correlation coefficient (Rp) of 0.939 and RMSEP of 0.587. These results demonstrate that the SCARS – PLSR model is highly effective for non-destructive SSC prediction in winter jujubes and provides valuable theoretical and technical insights for rapid quality assessment.**

***Index Terms*—winter jujube, soluble solid content, feature extraction, regression forecasting model**

I. Introduction

WINTER jujube is a late-maturing fresh jujube variety, also known as Yanlai red, apple jujube, and rock sugar jujube. The nutritional value of winter jujubes is very high, as they contain large amounts of vitamin A, vitamin E, potassium, sodium, iron, copper and other trace elements, and they are known as "natural vitamin pills"[1]. The quality of winter jujubes affects their flavour, taste, production and sales. The soluble solids content (SSC) is an important index for evaluating the quality of winter jujubes [2]. Therefore, SSC detection is particularly important. Soluble solids mainly refer to various soluble sugars including glucose, sucrose, and fructose. Due to evaluate the sugar content of fruits because of the significant positive correlation between SSC and soluble sugar content of fruits [3]. The traditional detection method uses a refractometer to detect SSC in fruit juice, which destroys the fruit, is time-consuming, requires considerable effort, and is difficult to scale. Therefore, it is highly important to explore a simple, rapid, and non-destructive SSC detection method to evaluate the quality of winter jujube for production and marketing.

With the continuous development of spectral technology, the non-destructive detection of fruit quality using spectral technology has gradually become a research hotspot in recent years. Many scholars at home and abroad have successfully applied it to the detection of fruit SSC [4]. Lu[5] used near-infrared technology to collect spectral data in the range of 800~1700 nm to estimate the hardness and sugar content of cherries and established a prediction model using PLSR. Zhang Dongyan et al.[6] obtained tomato spectral data using Vis/NIR spectroscopy and constructed tomato SSC prediction models using PLSR and LS-SVM. The results indicated that the experimental PLSR model achieved the best performance. Pan Tian et al.[7] used hyperspectral technology to obtain mango spectral data, selected variable intervals and feature variables from the spectral data, and established a PLS prediction model of the mango SSC.The R_p of the SNV-CARS-PLS prediction model was 0.9001, which was greater than the prediction effect without variable selection. Zhang De et al. [8] established the PLS prediction model of apple SSC based on variable interval selection and feature variable selection of apple spectral data. The results show that the R_p and *RMSEP* of the PLS model established on the variable set, which was selected by features, were 0.907 and 0.479, respectively. The above studies show that the PLSR model can effectively predict the SSC in various fruits, and feature extraction is an important link for improving the prediction accuracy and efficiency of the model.

For characteristic extraction methods, the Successive Projection Algorithm (SPA), Competitive Adaptive

Manuscript received July 19, 2024; revised December 26, 2024.

This work was supported in part by the Agricultural Science and Technology Achievements Transformation Fund Project of Hebei Province (202460104030028) and Scientific Research Project of Higher Education in Hebei Province, China（BJ2025097）

Aoran Liu is a postgraduate student in the College of Computer and Cyber Security at Hebei Normal University, Shijiazhuang, 050035, China(e-mail:347354547@qq.com).

Yufei Song is a professor of Shijiazhuang University, Shijiazhuang, 050035, China (corresponding author to provide e-mail:songyufei0311@163.com).

Zheng Xu is a lecturer of Langfang Polytechnic Institute, Langfang, China (E-mail:xuzheng@163.com).

Xi Meng is a lecturer of Shijiazhuang University, Shijiazhuang, 050035, China (E-mail: mengxi19930511@163.com).

Zhiguo Liu is a professor of Shijiazhuang University, Shijiazhuang, 050035, China (E-mail: 66344844@qq.com).

Reweighted Sampling (CARS) and other methods are often used for spectral feature extraction[9]. Jie Dengfei et al.[10] obtained the spectral data of navel orange using hyperspectral diffuse transmission technology, adopted the SPA feature extraction method, and established the SSC prediction model of PLSR for navel orange. The results showed that SPA-PLSR predicted a correlation coefficient of Rp=0.889. Gao Sheng et al.[11] obtained the spectral data of red extract using near-infrared spectroscopy, applied CARS, SPA, UVE and other feature extraction methods, and established the SSC prediction model of red extract based on PLSR; the SG-CARS-SPA-PLSR model predicted a correlation coefficient of R_p=0.9787. To improve feature extraction methods, Zheng Kaiyi et al. [12] proposed a competitive adaptive weighting algorithm based on the variable stability. SCARS considers the volatility of variable regression coefficients with sampling and uses stability as an index to select variables. Gao Sheng and Wang Qiaohua et al. [13] applied SCARS and other methods to extract the characteristic wavelengths of red extract, established PLSR prediction models of the Vc content and sugar content of red extracts, and verified their effects.

In summary, domestic and foreign scholars have applied PLSR modelling to many studies on internal fruit quality detection and have achieved good results. Various feature extraction methods have also been proposed to simplify the model and improve its prediction accuracy. However, studies on SSC of winter jujubes are limited to classification studies. Therefore, in this study, the spectral data of winter jujubes were obtained, SCARS and PLSR were combined to construct a prediction model for the SSC of winter jujubes, and a new non-destructive detection method for winter jujubes SSC was proposed. Fig. 1 illustrates this research route in detail.

II. Experiment

A. Experimental materials and instruments

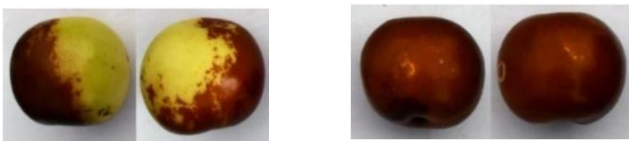
The experimental research object was "winter jujube 103" from the Cangxian National Jujube Breeding Base, Hebei Province, which was collected at 38°16' N and 114°54' E. In total, 986 samples were collected in this study. In reference to GB/T32714 “Winter Jujube”, GB/T22345 “Fresh Jujube Quality Grade”, other national standard documents, and the guidance of jujube experts, 60 winter jujubes in the crisp ripening stage without diseases or pests were selected as the research objects. To facilitate follow-up work, the winter jujube samples were refrigerated (0~5℃).

The main instruments in the experiment were a Nikon D7500 SLR digital camera, a USB2000+micro fibre spectrometer (spectral range: 350~1000 nm; optical resolution: 0.3~10 nm, American Ocean Optics Corporation) and an LC-DR-53B digital refractometer (range: 0.0~53.0% Brix; accuracy: ±0.2%; Shanghai Lichen Instrument Technology Co. Ltd.).

B. Experimental methods

1) Sample image acquisition

To avoid poor light, clear weather was selected for image acquisition. The Nikon D7500 SLR digital camera was selected as the image acquisition device. The shooting parameters of the camera were set as follows: Select the default settings for the aperture priority, auto white balance, auto focus, shutter speed and sensitivity and other parameters of the camera; set the image storage format to JPEG to prevent environmental interference during shooting; place the samples on a white background board; position the camera lens 30 cm directly above the sample and at an angle of 90° from the white background plate. Take part of the sample image in Fig. 2.



(a) Seventy percent ripe (b) Ninety percent ripe

Fig. 2. Examples of some samples

2) Spectral data acquisition

Since the band in the spectral edge range has more noise interference, the 400~1000 nm part was intercepted as the original spectrum for research and analysis. Before the spectral data of the sample were scanned, the halogen tungsten light source was preheated for 30 min to ensure the accuracy of the experiment. To reduce the interference of noise, the experiment was performed in a closed dark box, and the spectral probe was at an angle of 90° from the surface of the target sample with a fixed distance of 2 cm. To obtain a suitable range of signal strength, the integration time was set to 100 ms. The sliding average width can reduce the detection error of neighbouring pixels, and it was set to 3. To reduce the error of random spectral jitters, the average number of scans was set to 100. Five sampling points were selected near the "equator" of each winter jujube and scanned in turn. The Oceanview software was used to monitor the data collection in real time and record the experimental data. Finally, the arithmetic mean of the spectral data of the five sampling points was calculated as the reference spectrum of a single winter jujube sample. This set of spectral systems does not need to manually calculate the spectral data after the black and white correction after scanning, but in the experimental parameter setting, the use of the matching calibration whiteboard can complete the black and white correction operation. Before the formal scanning, the spectral data of the calibration whiteboard and background after the light source had been turned off were recorded; Then, the automatically corrected spectral data were obtained when the winter jujube samples were scanned.

The black and white correction formula is as follows:

$$\boldsymbol{X} = \frac{\boldsymbol{x}-\mathbf{B}}{\boldsymbol{W}-\mathbf{B}} \tag{1}$$

where x is the spectral data of the scanned winter jujube, W is the calibration data of the calibration whiteboard (theoretically, the reflectivity is the maximum value), B is the calibration data after the light source has been turned off (theoretically, the reflectivity is 0), and X is the spectral data of winter jujube after the black and white correction.Fig. 3 shows the original spectrum of the sample.

The original spectrum of winter jujubes shows that the spectral reflectance of the 400~600nm wavelengths was generally low, and a trough appeared at approximately 480nm. Probably due to the carotenoid relationship, when the wavelength increased, the reflectance continued to increase. A trough reappeared at 680 nm, which may be related to chlorophyll, after which the reflectance increased again with increasing wavelength. The change in the upward trend near 840nm may be related to the content of soluble solids, and the

small fluctuation at 980nm may be related to water absorption.

3) SSC determination of winter jujubes

The SSC of each winter jujube sample was measured using a refractometer. Referring to NY/T 2637-2014 “Fruit and vegetable soluble intangible substance content determination Refractometer method”, before the determination, the refractometer was calibrated to zero; then, the jujube pulp juice was extracted using a juicer. After filtration, 3 ml was absorbed through a calibrated Pasteur straw and titrated on the surface of the refractometer prism. The readings were taken after 3 s of rest, and three readings were taken for each sample. The average value was calculated to ensure the stability of the experimental data. The SSC results of the measured samples were 21.1 to 29.7% with an average value of 25.4%. The standard deviation was 1.71, and Table I shows the results.

TABLE I
Measured soluble solid contents of winter jujubes

Sample number	SSC(%)
1	26.3
2	26.3
3	26.0
4	25.9
…	…
56	27.0
57	25.8
58	27.7
59	28.4
60	26.7

III. Data Processing and Modelling Methods

A. Spectral data preprocessing

Spectral data preprocessing can eliminate the influence of noise, background interference, stray light interference and other factors on the data generated by the test instrument during spectral data acquisition. It can also eliminate redundant information in the spectral data to improve the model accuracy [14]. To avoid the influence of stray light, noise, baseline shifts and other factors on the final quantitative analysis results [15], it is necessary to preprocess the spectra. The wavelength with a large amount of noise was removed, and the region of 400~1000 nm was analyzed as the original spectrum.

Scattering correction, including multivariate scattering correction (MSC), is mainly used to eliminate the effects of the uneven particle size and distribution. MSC is a commonly used algorithm in spectral data preprocessing. This method uses a "standard spectrum" to correct the deviation caused by the baseline drift of the spectral data of winter jujubes, but it is difficult to perfectly obtain the "standard spectrum" in practice, and the average of the spectral data is often used instead.

Therefore, the average value of all winter jujube spectral data was selected as the "standard spectrum" in the experiment. The winter jujube spectral data contained 60 samples, denoted as n, and each sample had 1814 wavelength points, denoted as p. The winter jujube spectral data were set as an n×p matrix. The specific steps of the MSC are as follows:

Step 1: Calculate the average spectrum of the winter jujube samples, as shown in Formula (4-3):

$$\overline{\boldsymbol{X}} = \frac{\sum_{i=1}^{n} \boldsymbol{X_{i,j}}}{\boldsymbol{n}} \tag{2}$$

Step 2: Calculate the baseline offset coefficient and baseline translation of each winter jujube spectral data point; i.e., use the average spectrum of the winter jujubes to fit the spectral data of each winter jujube sample through linear regression as follows:

$$\boldsymbol{X_i} = \boldsymbol{k_i}\overline{\boldsymbol{X}} + \boldsymbol{b_i} \tag{3}$$

Step 3: Correct each spectrum in the original spectrum of the winter jujubes to eliminate the deviation error as follows:

$$\boldsymbol{X_{i\ (MSC)}} = \frac{\mathbf{X}_i - \boldsymbol{b_i}}{\boldsymbol{k_i}} \tag{4}$$

Here, $\overline{\boldsymbol{X}}$ is the average spectral vector of the original spectral data of the winter jujube sample, $\boldsymbol{X_i}$ is the spectral vector of the $\boldsymbol{i}$th winter jujube sample, $\boldsymbol{k_i}$ is the offset coefficient, and $\boldsymbol{b_i}$ is the translation.

The derivative method can remove the interference of background noise in the spectrum, such as the first- and second-order derivatives, and can enhance the imperceptible change trend in the spectrum.

First derivative formula:

$$\boldsymbol{y(n)}' = \frac{\boldsymbol{y(n+l)} - \boldsymbol{y(n)}}{\boldsymbol{l}} \tag{5}$$

Second derivative formula:

$$\boldsymbol{y(n)}'' = \frac{\boldsymbol{y(n+l)} - \boldsymbol{2y(n)} + \boldsymbol{y(n+l)}}{\boldsymbol{l^2}} \tag{6}$$

where $\boldsymbol{n}$ is the $\boldsymbol{N}$th wavelength, and $\boldsymbol{l}$ is the interval between spectra.

Smoothing is used to eliminate the effect of random errors on the spectral data. Savitzky–Golay(SG) is a type of weighted average algorithm that performs a polynomial least squares fit to the data within a moving window.

In this experiment, a combination of MSC−FD−SG pretreatment was proposed, which was combined with multiple scattering correction (MSC), Savitzky−Golay (SG), first derivative, second derivative and four single pretreatment methods to preprocess the spectral data. Compared with the original spectral data, Figure 4 shows the preprocessed spectral data.

The PLSR model of jujube SSC was established using partial least square regression on the processed spectral data (400~1000 nm). The effect of preprocessing was evaluated based on the root mean square error and correlation coefficient. Table II shows the results. The regression model based on the preprocessed spectrum performed better that that based on the original spectrum. In the regression model of the preprocessed spectra, the results of the MSC−FD−SG preprocessing combination had RMSEP=1.446 and R=0.713, which are better than those of the single preprocessing

method. The interference of the baseline drift, noise and background was effectively eliminated, and the trend change characteristics of the spectral curve were more obvious, which proves that the model fitting is more accurate. The MSC–FD–SG combination was selected as the optimal preprocessing method.

B. Characteristic wavelength extraction algorithm: Stable Competitive Adaptive Reweighted Sampling (SCARS)

The high spectral resolution in the experiment resulted in more wavelength points, which yielded more comprehensive spectral information, but there was also a lot of redundant information. This redundancy would increase the modelling time and reduce the correlation of the model. Thus, to simplify the model, remove redundant information, and obtain more efficient prediction model effect, we extracted the characteristic wavelength of the spectral data. In this paper, Stability Competitive Adaptive Reweighted Sampling (SCARS) was used as the main method of feature wavelength extraction, and SPA, UVE and IRIV were used as the control group.

Competitive Adaptive Reweighted Sampling (CARS) is a variable selection method that mimics Darwin's survival of the fittest evolutionary theory[16]. The Stability Competitive Adaptive Reweighted Sampling improves the CARS algorithm and takes the stability of variables as the measurement index [17]. Then, the variables are screened according to the process of the CARS algorithm to improve the stability, accuracy and selection efficiency of the selected variable subset [18].

Let matrix $X_{m\times n}$ be the spectral data of the winter jujube samples, *m* be the number of samples, and *n* be the number of variables, i.e., the number of wavelengths. The steps of SCARS are as follows:

Step 1: Calculate the stability value for each variable:

$$c_i = \left|\frac{\bar{b}_i}{s(b_i)}\right| \quad i = 1,2,\ldots,n \tag{7}$$

where c_j is the stability value of the j TH variable in M Monte Carlo samples, $\bar{b}_j$ is the mean value of the regression coefficient of the j TH variable sampled for this round, and $s(b_j)$ is its standard deviation.

Step 2: Use the exponential decay function (EDF) to force the retention of wavelength variables with greater stability and use the adaptive reweighted sampling (ARS) to select a set of variable subsets with relatively greater stability.

Step 3: Repeat steps 1-2 to obtain N subsets of variables, i.e., N cycles of the SCARS algorithm to establish its PLS model, Perform K-fold cross validation, Finally, select the smallest subset of RMSECV as the optimal variable subset, i.e., screen the characteristic wavelength.

C. Regression algorithm–Partial Least Squares Regression(PLSR)

Partial Least Squares Regression (PLSR) is a regression algorithm commonly used in spectral analysis [13]. PLSR combines factor and regression analysis and considers the influence of the independent variable matrix *X* and dependent variable matrix *Y* on the modelling effect. PLSR can address the problems of multicollinearity, non-normal distribution and uncertainty of factor results[20]. It combines MLR, PCA[27] and canonical correlation analysis to maximize the winner component, variance of the winner component and response. It avoids the disadvantage of only *X* decomposition in principal component regression. In this work, the independent variable matrix *X* is the spectral reflectance matrix of winter jujube, and the dependent variable matrix *Y* is the SSC measurement matrix of winter jujube. The established PLSR model has high prediction stability and is suitable for the analysis of small samples.

D. Model evaluation metrics

The prediction model evaluation considers the accuracy and stability of the model. The correlation coefficient (R), root mean square error of prediction (RMSEP), and root mean square error of cross validation (RMSECV) are important indices to evaluate prediction models[21].

The related coefficient represents the relationship between prediction result and standard result.
The calculation formula is as follows:

$$R = \sqrt{1 - \frac{\sum_{n=1}^{m}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{m}(y_n - \bar{y}_n)^2}} \tag{8}$$

where $\hat{y}_n$ is the predicted SSC value of the *n*th winter jujube sample, $\bar{y}_n$ is the average SSC value of the *n*th winter jujube sample, y_n is the true SSC value of the *n*th winter jujube sample, and *m* is the corresponding total number of samples. The maximum value of *R* is 1. A larger *R* indicates that the model prediction result is closer to the measured standard value; i.e., the prediction result is more accurate.

The Root Mean Square Error of Prediction is calculated as follows:

$$RMSEP = \sqrt{\frac{\sum_{n=1}^{m_p}(\hat{y}_{n,p} - y_n)^2}{m_p}} \tag{9}$$

where m_p is the number of samples in the prediction set, and $\hat{y}_{n,p}$ is the prediction value of the *n*th sample in the prediction set. A smaller RMSEP value corresponds to a smaller apparent error and a more concentrated prediction value of the prediction model.

The Root Mean Square Error of Cross Validation is calculated as follows:

$$RMSECV = \sqrt{\frac{\sum_{n=1}^{m_{cv}}(\hat{y}_{n,cv} - y_n)^2}{m_{cv}}} \tag{10}$$

where m_{cv} is the number of cross validation samples, and $\hat{y}_{n,cv}$ is the *n*th sample predictive value of cross validation.

TABLE II
Evaluation results of different preprocessing methods

Pretreatment method	*RMSEP*	R_p
Raw spectrum	1.937	0.342
SG	1.932	0.354
VN	1.567	0.694
MSC	1.449	0.708
FD	1.483	0.702
SD	1.312	0.487
MSC–FD–SG	1.446	0.713

IV. Experimental results and analysis

A. Analysis of feature wavelength extraction results

SCARS introduces a stability value to improve the stability of variable selection and uses the ARS, EDF and RMSECV

to select the optimal wavelength variable subset. The experimental spectrum (400~1000 nm) contained 1814 wavelength points, and the characteristic wavelengths were extracted by SCARS. The number of Monte Carlo samples M=50 and number of cross validation k=10 were set. Fig. 6 shows the feature extraction results of SCARS.

The figure shows that the number of retained wavelengths rapidly decreased when the number of runs increased, and the RMSECV continuously decreased, which indicates that this process eliminates many redundant variables. Afterwards, when the number of runs increased again, the number of filtered wavelengths slowly decreased, which indicates the process from rough screening to precise screening. When running to the 31st time, RMSECV=0.374, and the number of variables in the selected variable subset was 28. Then, the running times continued to increase, and the RMSECV showed an increasing trend, which indicates that the wavelengths strongly correlated with the SSC of winter jujubes in the spectral data were eliminated in this process.

Taking the spectral curve of one of the samples as an example, Fig. 7 shows the distributions of 28 feature variables extracted from the SCARS features. The number of selected characteristic wavelengths accounted for only 1.54% of the number of full-band wavelengths, which greatly reduced the required input variables and simplified the input variables of the model.

Except for the wavelength points at approximately 450 nm and 700 nm, most of the other selected feature bands concentrated in the range of 800~1000 nm, which is basically consistent with the range of SSC spectral characteristic variables of other fruits, such as pears and citrus. Note that 709~759 nm and 789~999 nm are important spectral regions for predicting the SSC [8]. The reason may be that the frequency doubling peaks of OH and CH bonds are related to the spectrum in the range of 750~980 nm.

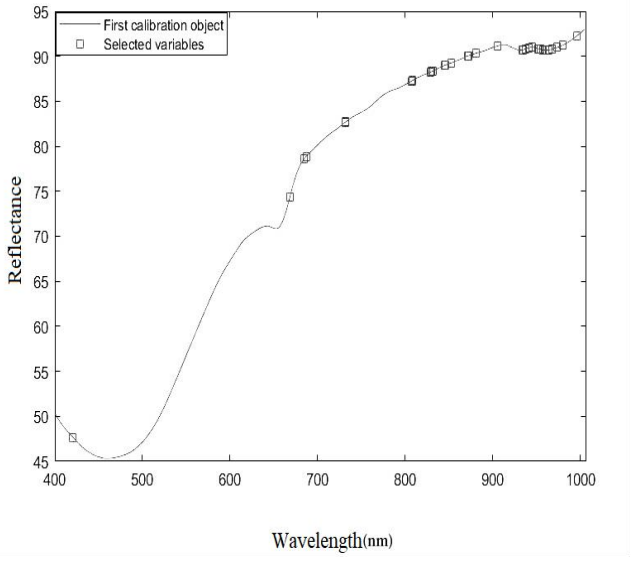


Fig. 7. Distribution of the characteristic variables selected by SCARS

B. Comparison of SCARS with SPA, UVE and IRIV feature extraction methods

In this work, SPA, UVE and IRIV were used as the control group to compare with the SCARS feature extraction method, and the feature extraction results are shown in Table III.

As shown in Table III, all four feature extraction methods significantly reduced the input variables of the prediction model.

Among them, the number of wavelengths selected by SPA was 14, which accounted for only 0.77% of the full wavelengths and most of the collinearity among variables was eliminated [22]. However, this method follows the principle of small collinearity in the process of variable selection and cannot guarantee that the selected wavelength point is an effective wavelength point; therefore, the stability is not good, and its *RMSECV*=0.897 is relatively high.

UVE can eliminate the wavelength points that do not contribute to the model to reduce the dimension of the data [23], and 26 wavelength points were retained in the final screening, i.e., 1.43% of the full wavelengths. However, the leave-one-out method is used in the sampling, so the final parameter reliability must be improved, with an *RMSECV* of 0.874.

IRIV has the characteristic of soft shrinkage. In the screening process, higher weights are assigned to the points with a high frequency of wavelength points in the excellent subset to ensure a greater probability of retention in the next iteration[24]. In total, 78 characteristic wavelengths were retained, i.e., 4.3% of all the wavelengths. Compared with other methods in this paper, the RMSECV had a small decrease. Although IRIV can retain the feature wavelength reliably, it must build a large quantum model, which leads to a large computational amount, and its process has a significantly longer running time than other feature extraction methods.

SCARS obtained 28 feature wavelengths, i.e., 1.54% of the total wavelengths, which is a more condensed subset of feature variables than IRIV screening. The *RMSECV* decreased by 0.424~0.525 compared with those of the first three methods, which indicates high stability and requires a relatively small number of calculations.

C. Establishment of the SSC prediction model for winter jujubes

1) PLSR prediction model

TABLE III
Comparison of characteristic wavelength extraction results

Feature selection	Number of variables	Characteristic wavelength	*RMSECV*
SPA	14	422, 533, 631, 639, 643, 646, 657, 695, 700, 702, 705, 724, 711, 713	0.897
UVE	26	499, 631, 682, 713, 799, 822, …, 873, 949, 977, 993	0.874
IRIV	78	438, 498, 631, 657, 699, 712, …, 822, 873, 947, 977	0.796
SCARS	28	437, 453, 458, 499, 682, 822, …, 948, 954, 977, 993	0.374

Different PLSR models were established based on the feature bands screened by SCARS, SPA, UVE, IRIV feature extraction methods and the full wavelengths (FW). Table IV shows the prediction results of the established FW-PLSR, SPA-PLSR, UVE-PLSR, IRIV-PLSR and SCARS-PLSR models, where the extracted feature wavelength was the

independent variable, and the measured SSC of winter jujubes was the dependent variable.

As Table IV shows, the PLSR prediction model established after the extraction of four types of features only used 0.77~4.3% of the entire wavelengths, which simplified the model and improved the running speed. Compared with the FW-PLSR model established in the entire wavelengths, the PLSR prediction effect of the four models significantly improved, RMSEP significantly decreased, and the correlation coefficient increased by 0.121~0.228. Compared with other feature extractions, the PLSR model based on SCARS feature extraction achieved the best results with R_p=0.939 and *RMSEP*=0.587.

TABLE IV
PLSR prediction model based on feature screening

Prediction model	Number of variables	*RMSEP*	R_p
FW-PLSR	1814	1.473	0.711
SPA-PLSR	14	0.788	0.832
UVE-PLSR	26	0.753	0.804
IRIV-PLSR	78	0.803	0.831
SCARS-PLSR	28	0.587	0.939

2) Random forest regression prediction model

The random forest (RF) algorithm is an integrated learning algorithm based on a decision tree[24]. Random Forest Regression (RFR) is an application of random forests to regression problems. In this work, the mean square error (MSE) was selected as the impurity function of the random forest regression model. Then, the average of the prediction results of each individual tree was used as the prediction result. The feature wavelengths and full wavelengths (FW) extracted by SPA, UVE, IRIV and SCARS were used as input variables to establish random forest regression models with different feature variables, and Table V shows the results.

The correlation coefficient R_p of the five models established by random forest was 0.755-0.888. Compared with the RFR prediction model established by the full wavelengths, the RFR prediction model constructed by the characteristic wavelength simplified the model and slightly improved the overall prediction accuracy. Among them, R_p=0.888 and RMSEP=0.774 for the SCARS-RFR model, which indicates that SCARS is slightly better than other feature extraction methods in predicting the SSC of winter jujubes.

The experimental data are presented in TABLES IV and V, respectively. Compared with the prediction model of RFR, the prediction model of winter jujube SSC established by PLSR is superior to that of random forest under identical characteristic variables except for the prediction model established by all wavelengths. SCARS-PLSR has the best prediction effect among the 10 models, which indicates that SCARS-PLSR reduces the model complexity and greatly improves the model accuracy. The prediction model based on ridge regression had poor performance, with correlation coefficients of 0.672 to 0.797, and the prediction results of the prediction models based on PLSR and RFR were quite different, so a detailed comparison was not conducted.

TABLE V
RFR prediction model based on feature screening

Prediction model	Number of variables	*RMSEP*	R_p
FW-RFR	1814	1.374	0.755
SPA-RFR	14	0.972	0.766
UVE-RFR	26	1.023	0.773
IRIV-RFR	78	0.932	0.801
SCARS-RFR	28	0.774	0.888

V. Conclusion

To explore a non-destructive testing method to determine the soluble solid content of winter jujubes, mature winter jujubes were used as the research object. Spectral data were obtained using hyperspectral technology, and the feature wavelengths were extracted using SCARS, SPA, UVE and IRIV feature wavelength extraction algorithms. Finally, PLSR and RFR prediction models were established based on the characteristic wavelengths and full wavelengths. The results showed that SCARS performed the best among the four feature wavelength extraction methods. The *RMSECV* was 0.374, and the number of feature wavelengths decreased from 1 814 to 28, accounting for only 1.54 % of all wavelengths. Among the PLSR and RFR models, the PLSR and RFR models based on the feature wavelengths were superior to those based on the entire wavelengths. SCARS-PLSR had the best prediction effect(R_p=0.939 and RMSEP=0.587). The SSC model of winter jujubes based on hyperspectral SCARS-PLSR ensures the integrity of winter jujube samples and improves the detection speed and accuracy, satisfying the requirements of non-destructive testing of winter jujube. This study provides a theoretical basis and technical support for rapid and non-destructive testing of the winter jujube quality. However, this work mainly studied mature winter jujubes, so jujubes in the other growth stages must be further discussed.

References

[1] Yong H, Du Jiao-jun, Shu-min Z, et al. Research on construction of visible-near infrared spectroscopy analysis model for soluble solid content in different colors of jujube[J]. SPECTROSCOPY AND SPECTRAL ANALYSIS, 2021, 41(11): 3385-3391.

[2] Zhang D, Xu L, Wang Q, et al. The optimal local model selection for robust and fast evaluation of soluble solid content in melon with thick peel and large size by Vis-NIR spectroscopy[J]. Food Analytical Methods, 2019, 12: 136-147.

[3] Monago-Maraña O, Afseth N K, Knutsen S H, et al. Quantification of soluble solids and individual sugars in apples by Raman spectroscopy: A feasibility study[J]. Postharvest Biology and Technology, 2021, 180: 111620.

[4] Fuxian H, Qinghua M, Liu T, et al. Research progress of hyperspectral imaging technology in fruit quality detection [J]. Journal of fruit trees, 2021, 38(09): 1590-1599.

[5] Lu R. Predicting firmness and sugar content of sweet cherries using near–infrared diffuse reflectance spectroscopy[J]. Transactions of the ASAE, 2001, 44(5): 1265.

[6] Zhang D, Yang Y, Chen G, et al. Nondestructive evaluation of soluble solids content in tomato with different stage by using Vis/NIR technology and multivariate algorithms[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2021, 248: 119139.

[7] Tian P, Meng Q, Wu Z, et al. Detection of mango soluble solid content using hyperspectral imaging technology[J]. Infrared Physics & Technology, 2023, 129: 104576..

[8] Zhang D, Xu Y, Huang W, et al. Nondestructive measurement of soluble solids content in apple using near infrared hyperspectral

imaging coupled with wavelength selection algorithm[J]. Infrared Physics & Technology, 2019, 98: 297-304.
[9] Wen-Li X U , Lin-Tao Y , Tong S ,et al.CARS-SPA baesd Visble/near Infraed spectroscopy on-line detection of apple soluble solids content[J].Science and Technology of Food Industry, 2014.[12]
[10] Deng-fei, Jie, Li Ze-hai, Zhao Jun-wei, Lian Yu-xiang and Wei Xuan. "Visualized Detection of Soluble Solid Content Distribution of Navel Orange Based on Hyperspectral Diffuse Transmittance Imaging." Chinese Journal of Luminescence 38 (2017): 685-691.
[11] Sheng G A O, Jianhua X U. Non-destructive Detection of the Internal Quality of Red Globe Grapes Based on Near Infrared Spectroscopy[J]. 2022.
[12] Zheng K, Li Q, Wang J, et al. Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra[J]. Chemometrics and Intelligent Laboratory Systems, 2012 (112): 48-54.
[13] Gao S H, WANG Q H, Li Q X, et al. Non-destructive detection of vitamin C, sugar content and total acidity of red globe grape based on near-infrared spectroscopy[J]. Chinese Journal of Analytical Chemistry, 2019, 47(6): 941-949.
[14] Meng-ru L, Shu-juan Z, Rui R E N, et al. Nondestructive detection of moisture content in fresh fruit corn based on hyperspectral technology[J]. Food and Machinery, 2021, 37(9): 127-132.
[15] Liu J, ** S, Bao C, et al. Rapid determination of lignocellulose in corn stover based on near-infrared reflectance spectroscopy and chemometrics methods[J]. Bioresource Technology, 2021, 321: 124449.
[16] Tao Z, Ning L, Hong S. Visualization of chlorophyll distribution of potato leaves based on hyperspectral imaging technology[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(S1): 153-159.
[17] Jiang Z, Du Y, Cheng F, et al. A simple multiple linear regression model in near infrared spectroscopy for soluble solids content of pomegranate arils based on stability competitive adaptive re-weighted sampling[J]. Journal of Near Infrared Spectroscopy, 2021, 29(3): 140-147.
[18] Liu G H, **a R S, Jiang H, et al. A wavelength selection approach of near infrared spectra based on SCARS strategy and its application[J]. Guang pu xue yu Guang pu fen xi= Guang pu, 2014, 34(8): 2094-2097..
[19] Meacham-Hensold K, Montes C M, Wu J, et al. High-throughput field phenoty** using hyperspectral reflectance and partial least squares regression (PLSR) reveals genetic modifications to photosynthetic capacity[J]. Remote Sensing of Environment, 2019, 231: 111176.
[20] Xu Y, Liu J, Sun Y, et al. Fast detection of volatile fatty acids in biogas slurry using NIR spectroscopy combined with feature wavelength selection[J]. Science of The Total Environment, 2023, 857: 159282.
[21] Zhang B H, Qian C Q, Jiao J K, et al. Rice moisture content detection method based on dielectric properties and SPA-SVR algorithm[J]. Transactions of the CSAE, 2019, 35(18): 237-244.
[22] Centner V, Massart D L, de Noord O E, et al. Elimination of uninformative variables for multivariate calibration[J]. Analytical chemistry, 1996, 68(21): 3851-3858.
[23] Wang S, Sun J, Fu L, et al. Identification of red jujube varieties based on hyperspectral imaging technology combined with CARS-IRIV and SSA-SVM[J]. Journal of Food Process Engineering, 2022, 45(10): e14137.
[24] Mishra P, Woltering E, Brouwer B, et al. Improving moisture and soluble solids content prediction in pear fruit using near-infrared spectroscopy with variable selection and model updating approach[J]. Postharvest Biology and Technology, 2021, 171: 111348.
[25] Amira M. Idrees, Nermin Samy Elhusseny, and Shimaa Ouf, "Credit Card Fraud Detection Model-based Machine Learning Algorithms," IAENG International Journal of Computer Science, vol. 51, no. 10, pp1649-1662, 2024
[26] Panduranga Vital Terlapu, U D Prasan, T. Ravi Kumar, Vijaya Bendalam, Sasibhushana Rao Pappu, M. Jayanthi Rao, Maddula Ratna Mohitha, and Jayaram. D, "Rice Category Identification through Deep Transfer Learning Features and Machine Learning Classifiers: An Intelligent Approach," IAENG International Journal of Computer Science, vol. 51, no. 7, pp765-784, 2024
[27] Abdul Rahaman Shaik, and P. Rajesh Kumar, "Performance Evaluation of Machine Learning Algorithms on Skin Cancer Data Set Using Principal Component Analysis and Gabor Filters," IAENG International Journal of Computer Science, vol. 51, no. 7, pp831-841, 2024

Aoran Liu is a postgraduate student in the College of Computer and Cyber Security at Hebei Normal University, He is also a researcher at the Hebei Key Laboratory of Network and Information Security, and a researcher at the Hebei Provincial Engineering Research Center for Supply Chain Big Data Analytics and Data Security.
Yufei Song is a professor of Shijiazhuang University,She is also a researcher at the Hebei Key Laboratory of IoT Blockchain Integration.

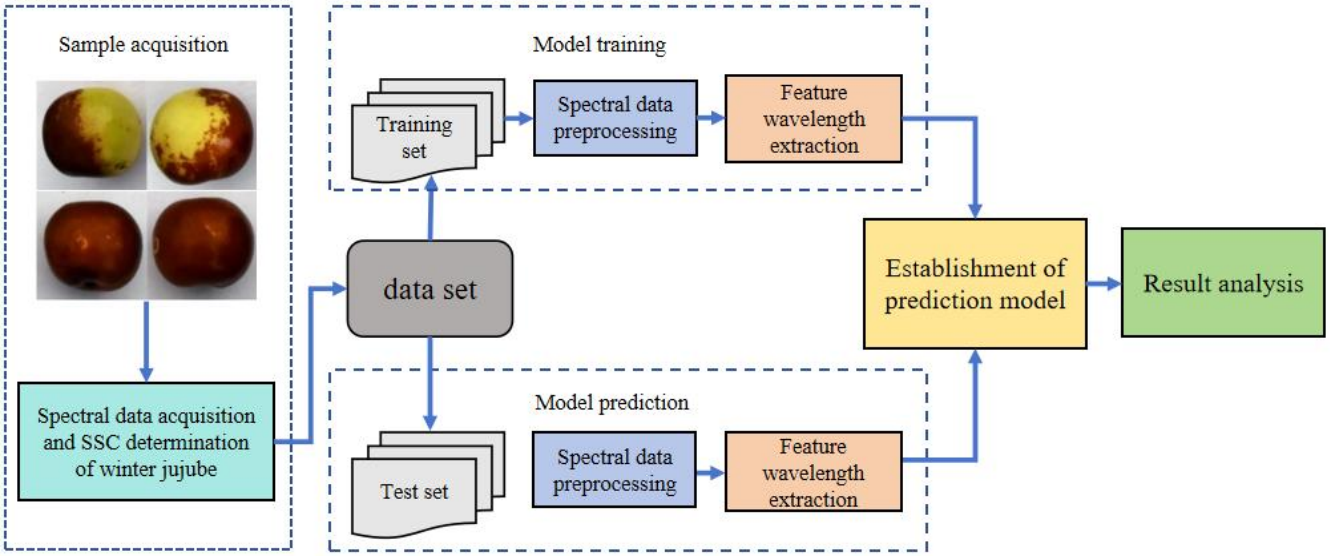


Fig1.research route

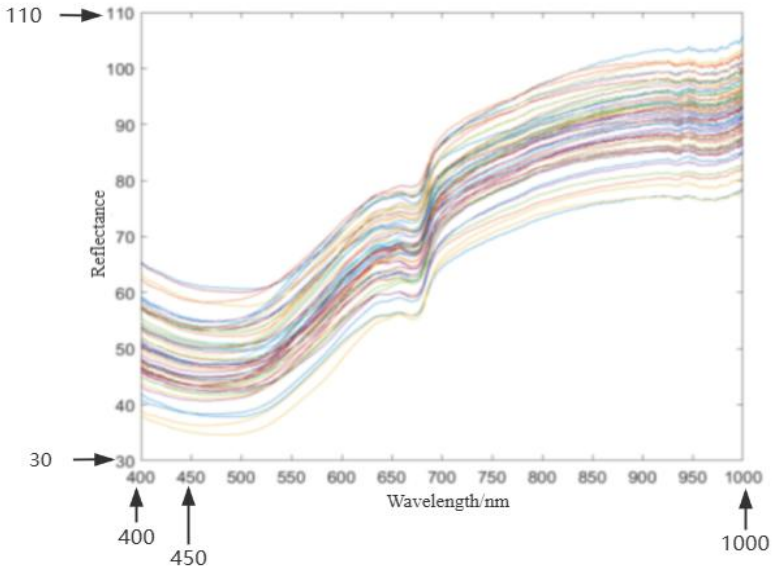
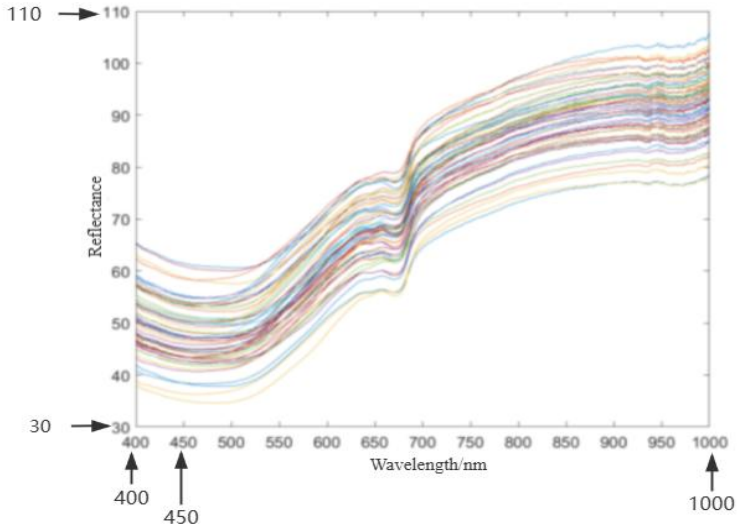
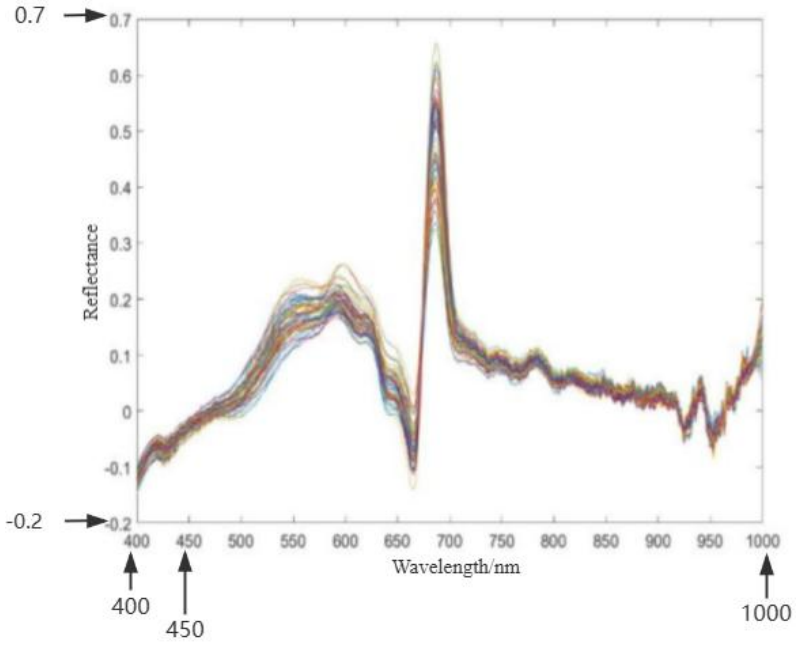


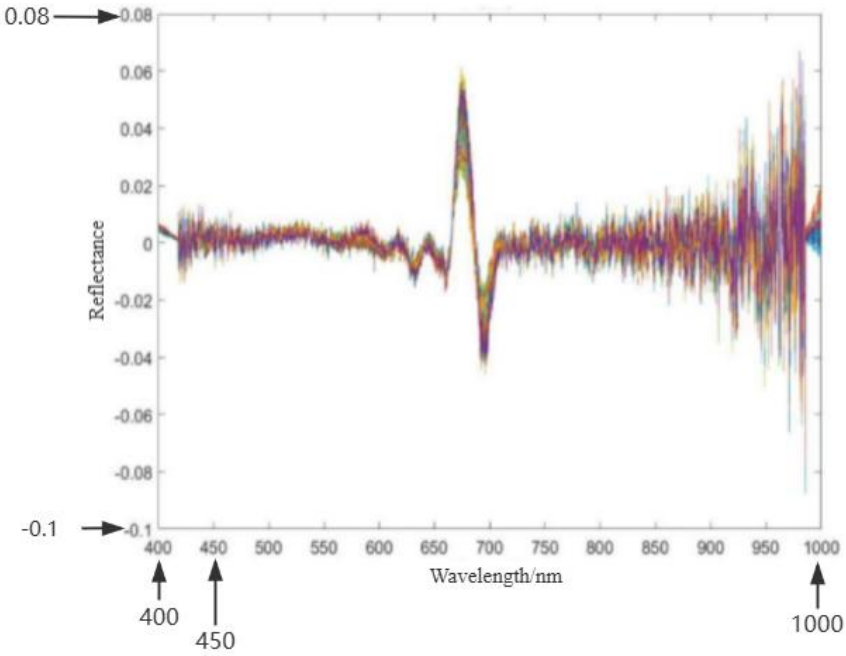
Fig. 3. Original spectrum



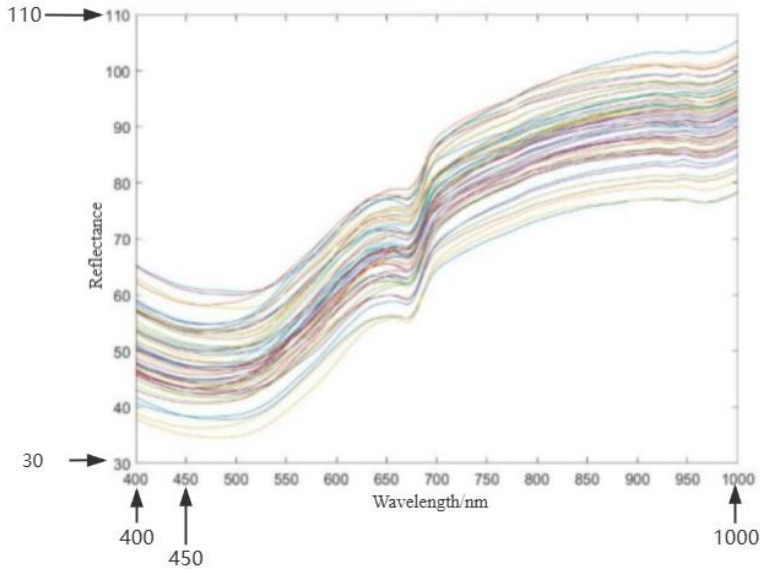
(a)Original spectrum



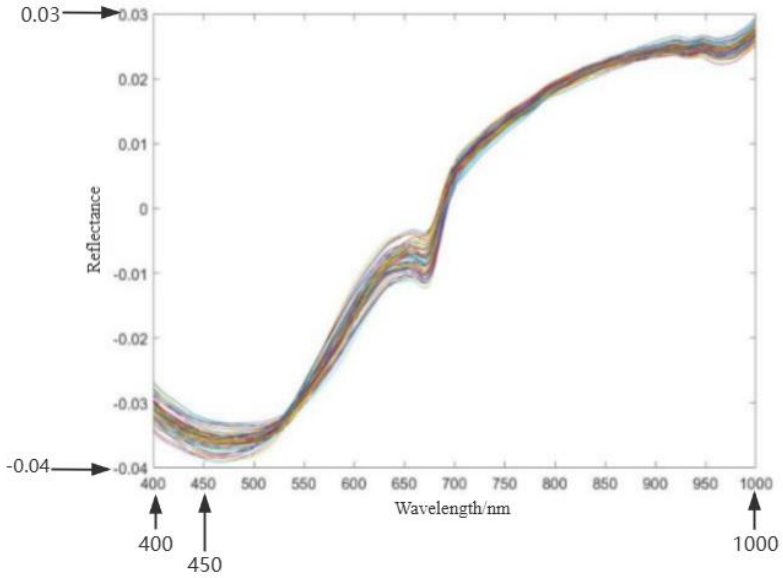
(b) First Derivative



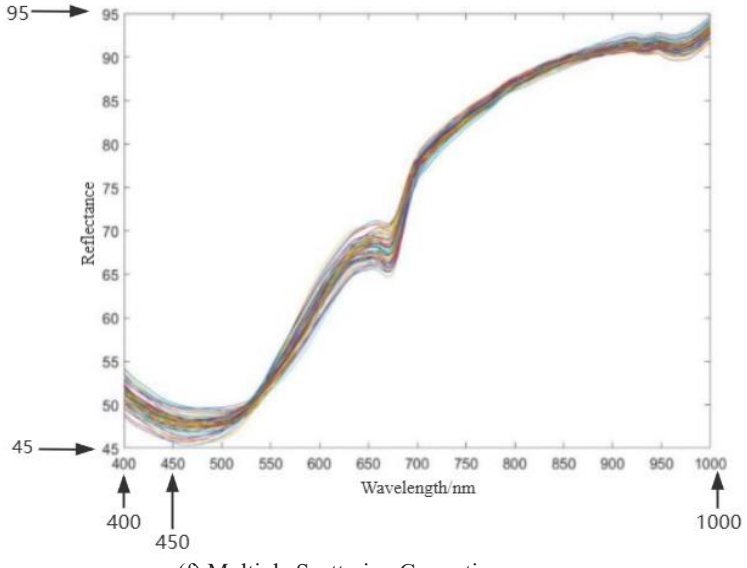
(c) Second Derivative



(d) Savitzky-Golay



(e) Vector Normalization



(f) Multiple Scattering Correction

Fig.4. Comparison of different preprocessing resultsestablished by the pretreated spectrum is improved.

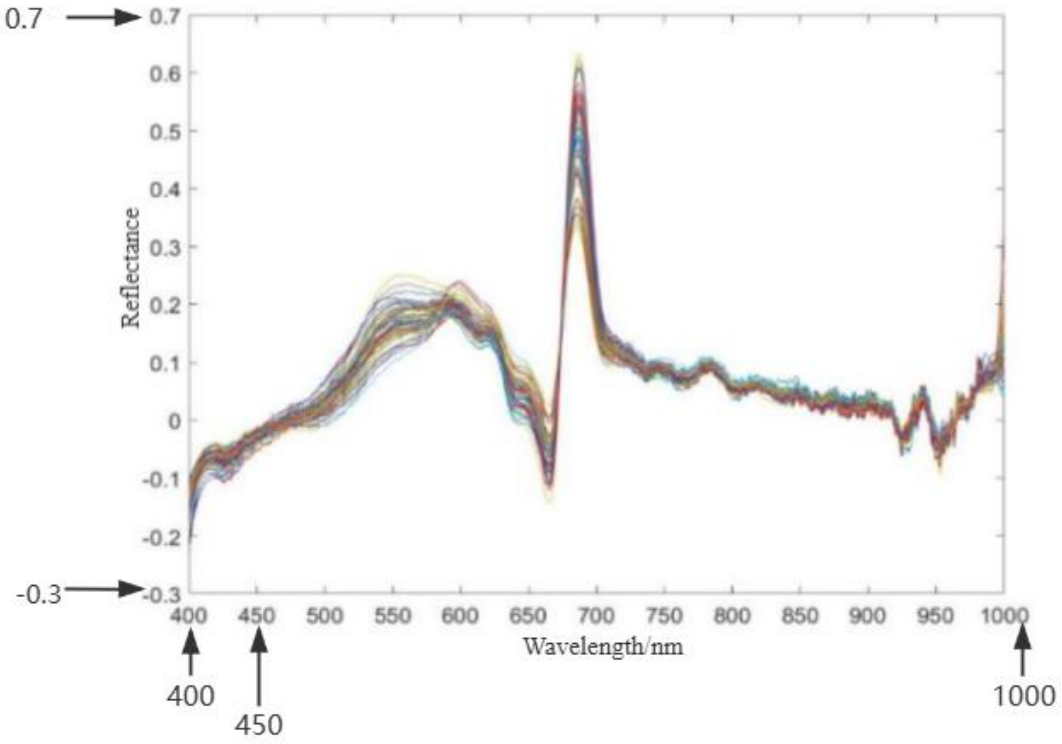


Fig.5. Spectral preprocessing result of MSC-FD-SG

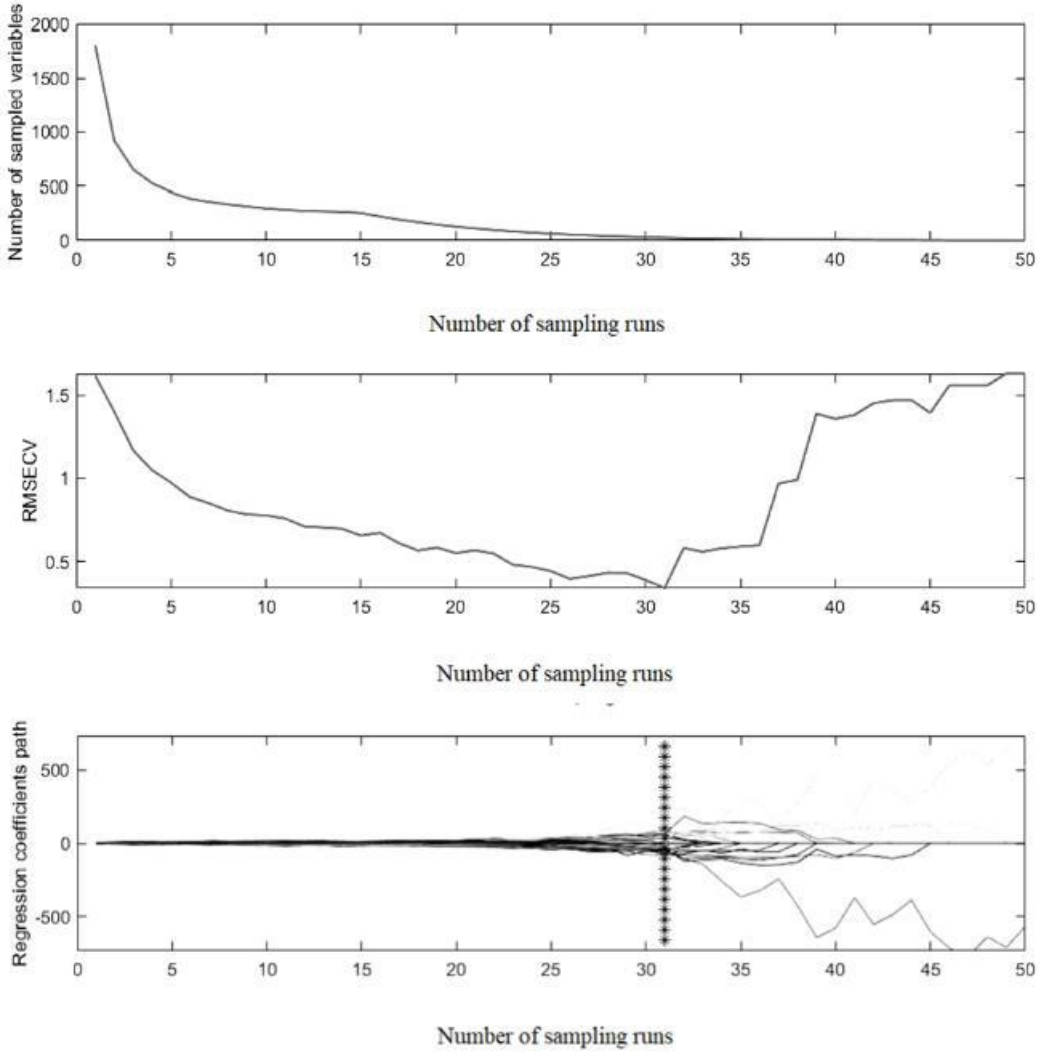


Fig.6.Stability competitive adaptive reweighted sampling