Nondestructive Quantification of Soluble Solid Content in ‘Red Fuji’ Apples Using Near-Infrared Diffuse Reflectance Spectroscopy with a Low-Cost Embedded Spectrometer

Tianhao Wang , Chengcong Ma , Xiangjun Xu , Xuanbing Qiu * , Ye Teng

Posted Date: 28 February 2026

doi: 10.20944/preprints202602.2044.v1

Keywords: apple; near-infrared diffuse reflectance spectra; soluble solid content; nondestructive detection; embedded spectrometer; partial least squares regression

 doi:10.20944/preprints202602.2044.v1

Article

Nondestructive Quantification of Soluble Solid Content in 'Red Fuji' Apples Using Near-Infrared Diffuse Reflectance Spectroscopy with a Low-Cost Embedded Spectrometer

Tianhao Wang [1], Chengcong Ma [2], Xiangjun Xu [2], Xuanbing Qiu [2,*] and Ye Teng [1]

[1] College of Science, Shihezi University

[2] School of Applied Science, Taiyuan University of Science and Technology

* Correspondence: qiuxb@tyust.edu.cn

Abstract

Soluble solid content (SSC) is a critical indicator of 'Red Fuji' apple quality, directly governing fruit grading and maturity assessment processes. Conventional SSC measurement by refractometry is destructive and time-consuming, rendering near-infrared diffuse reflectance spectroscopy (NIR-DRS) a promising nondestructive alternative. In this study, a low-cost and compact embedded spectrometer named as DLP NIR-scan Nano EVM was used to acquire NIR-DRS spectra of 'Red Fuji' apples for SSC prediction. To improve prediction accuracy, we combined spectral preprocessing with machine learning methods. The dataset was cleaned using Monte Carlo outlier detection, and samples were divided into calibration and validation sets via Kennard–Stone (KS) and joint X-Y distance (SPXY) algorithms. Among preprocessing methods tested, a 12-point second derivative performed best when paired with KS partitioning. For feature-wavelength selection on the preprocessed KS data, competitive adaptive reweighted sampling, Monte Carlo uninformative variable elimination, and Random Frog were applied to the second-derivative spectra. Partial least squares regression (PLSR) models were then built using both full-spectrum data and four sets of selected wavelengths. The best preprocessed PLSR model achieved R^2c = 0.916, RMSEC = 0.4093%, R^2p = 0.8632, and RMSEP = 0.537%. These results demonstrate that NIR-DRS, combined with appropriate preprocessing and modeling strategies, offers a reliable, rapid, and nondestructive method for apple SSC quantification, paving the way for portable, cost-effective instruments for commercial fruit quality monitoring.

Keywords: apple; near-infrared diffuse reflectance spectra; soluble solid content; nondestructive detection; embedded spectrometer; partial least squares regression

1. Introduction

Apples are among the most widely consumed fruits globally and are often hailed as the "king of fruits". The quantification of soluble solid content (SSC) is indispensable for evaluating apple quality, playing a pivotal role in quality classification and maturity monitoring throughout the supply chain. Currently, the refractive index method serves as the gold standard for determining sugar content in apples. However, this conventional SSC measurement technique is inherently destructive, cumbersome to perform, and time-consuming, limiting its applicability for large-scale, real-time quality inspection[1]. In recent years, various spectral techniques have immerged for application in the field of food analysis and detection, including visible and near-infrared

 doi:10.20944/preprints202602.2044.v1

spectroscopy (VIS-NIRS)[2-5], near-infrared spectroscopy (NIRS)[6-8], and Raman spectroscopy[9-12].

As a rapid, low-cost, compact, and versatile nondestructive technique, NIRS has been widely applied for the nondestructive analysis of internal quality parameters in a broad range of products, including fruits, pharmaceuticals, vegetables, meat, milk, coffee, tea, and chocolate[13-16]. Grabska et al. reviewed the diverse VIS-NIRS testing methodologies and strategies employed to address challenges in authenticity verification, provenance determination, identification, anti-counterfeiting, and quality control[17]. Fodor conducted another comprehensive review focusing on the role of NIRS in food quality assurance, with specific emphasis on apple SSC measurement[16]. Liu et al. successfully developed a universal prediction model for apple sugar content by leveraging NIRS diffuse transmission spectra (DTS) and integrating the uninformative variable elimination (UVE) method with partial least squares regression (PLSR). resulting model achieved a prediction correlation coefficient (R^2p) of 0.80 and a root mean square error of prediction (RMSEP) of 0.61%[18]. Yuan et al. utilized a Maya2000Pro fiber-optic spectrometer to develop a portable analyzer for the nondestructive assessment of fruit internal quality[19]. By applying various wavelength selection algorithms to reduce input variable dimensionality and adopting multiple linear regression (MLR) for modeling, they achieved a remarkable R^2p of 0.951 and an RMSEP of 0.39%. Nevertheless, the high cost of the system and limited model robustness posed significant challenges for its commercialization. Guo et al. implemented four distinct variable selection methods to enhance model performance, among which the PLSR model combined with competitive adaptive reweighted sampling (CARS) exhibited the optimal results, with an R^2p of 0.9808 and an RMSEP of 0.327%[20]. However, the reliance on Monte Carlo (MC) sampling in the CARS method introduced instability in the selection of characteristic variables, compromising model reproducibility.

In contrast, Wang et al. developed a low-cost, online, portable NIRS system using a digital light processing NIR-scan Nano spectrometer (DLP-NIRS) for the nondestructive detection of SSC in sweet cherries, achieving an R^2p of 0.83 and an RMSEP of 1.56%[21]. Unfortunately, this accuracy level failed to meet the stringent requirements of commercial market applications. Lanjewar et al. acquired reflectance spectra in the 900–1700 nm range using a compact DLP-NIR module and analyzed turmeric samples adulterated with starch via the Savitzky-Golay (S-G) filter[22]. A series of machine learning models were tested, with extra tree regression demonstrating superior performance, demonstrating superior performance ($R^2 = 0.995$, RMSEV = 1.056 mg). Yao et al. developed a portable NIR diffuse reflectance instrument for monitoring the SSC of intact 'Fuji' apples[23]. They extracted optimal feature wavelengths from VIS-NIRS (400–1100 nm) via S-G smoothing and the successive projections algorithm (SPA), and the multivariate nonlinear regression (MNLR) model yielded high SSC prediction accuracy (R^2p = 0.953, RMSEP = 0.391%), outperforming the back-propagation artificial neural network (BP-ANN) model. This work confirmed the efficacy of combining NIRS with MNLR for SSC monitoring in apples using custom-built portable instrumentation.

In the Recently, Li et al. developed a fusion SSC prediction for Dangshan pears by combining NIR spectrometer and hyperspectral imaging sensor employing one-dimensional convolutional neural network enhanced with global context block (1D GC-CNN)[24]. Their experimental results shown that the 1D GC-CNN model achieved the highest R^2p (0.9012) and the lowest RMSEP (0.2788). To measure Fuji apple SSC outdoors using a handheld NIRS (650–950nm) device, Sun et al. implemented orthogonalization (EPO) and generalized least square weighting to correct for ambient light interference[25]. The experiment results demonstrated that EPO is more efficient, requiring minimal samples for spectral correction, and excellent predictive performance for unknown samples. Elamshity et al. proposed a nondestructive method to assess date fruit quality using VIS-NIRS (410–990 nm) and convolutional neural networks (ANNs)[26]. ANNs, outperforming PLSR on second-derivative preprocessed spectra and successfully modeled a composite quality index, achieving strong correlations (R^2p up to 0.944). Tian et al. proposed a multi-attention convolutional neural network (MA-CNN) combined with hyperspectral imaging for apple SSC detection[27]. By integrating channel and spatial attention mechanisms, and optimized via the Bayesian optimization

 doi:10.20944/preprints202602.2044.v1

algorithm, the MA-CNN effectively extracts features. It achieved superior SSC prediction performance (R^2p = 0.9602), significantly outperforming mainstream models in accuracy. To address the cumbersome manual operation of existing hyperspectral imaging (HSI) grape SSC nondestructive detection, Junhong Zhao et al. developed an improved method for bunch-harvested Shine-Muscat grapes using Deep learning and HSI[28]. Validated with 35 selected characteristic wavelengths, the method showed high feasibility and efficiency (F1=95.34%, R^2p=0.8755). However, these studies focus on multi-sensor fusion algorithms or high-performance prediction models, achieving high R^2p values and ultra-low RMSEP. Therefore, to address the limitations of expensive, bulky, and power-intensive commercial spectrometers, Zhang et al. developed a novel low-cost, handheld IoT multispectral detection device[29]. This innovation integrates AS7265X sensors to collect 18 spectral channels across the 410–940 nm range, facilitating on-site applications. Their SSC prediction model, established using MC sampling, PCA, and S-G preprocessing, demonstrated promising performance with a maximum R^2p value of 0.809, providing an accessible solution for spectral analysis.

DTS is a commonly employed method for apple SSC measurement, as it captures intrinsic physiological information of the fruit. However, DTS requires high-power light sources, which consume substantial energy and may adversely affect apple quality upon prolonged exposure. In contrast, NIR diffuse reflectance spectroscopy (DRS) employs low-power light sources, which reduces energy consumption and avoids compromising apple quality. Additionally, DRS offers advantages such as compact instrumentation and cost-effectiveness. In this study, we developed a simple and low-cost PLSR prediction model based on NIR DRS for the accurate quantification of SSC in 'Red Fuji' apples by using an embedded spectrometer. Sample sets were partitioned using the Kennard-Stone (KS) and sample set partitioning based on joint X-Y distance (SPXY) algorithms, and the model was optimized through a comprehensive evaluation of spectral preprocessing methods and feature variable selection techniques. The resulting PLSR prediction model lays a solid foundation for the development of a portable, cost-effective, and high-precision nondestructive testing instrument for apple sugar content analysis.

2. Materials and Methods

2.1. Experimental Sample

Twenty-nine defect-free 'Red Fuji' apples were purchased from a large supermarket in Taiyuan, Shanxi Province. The apples had equatorial diameters ranging from 75 mm to 95 mm. To ensure the accuracy of spectral acquisition and SSC determination, surface impurities and dust were carefully removed[22]. The fruits were then acclimatized in a laboratory environment at 20 °C for 24 hours prior to experimentation. For each apple, three measurement points were selected near the equatorial region, spaced 120° apart and oriented perpendicular to the stem axis. Both spectral data collection and sugar content measurement were conducted at each of these points, yielding a total of 87 independent samples. Each measurement position was treated as a distinct and independent sample in subsequent data analysis.

2.2. Spectral Acquisition and Soluble Solid Content Determination

A portable DLP NIR-scan Nano EVM spectrometer (Texas Instruments, USA, Figure 1(a)) was utilized to collect near-infrared diffuse reflectance spectra (NIR-DRS) in absorption mode across the wavelength range of 900-1700 nm. During measurements, the apple sample was placed in tight contact with the spectrometer's acquisition window, with the target measurement position precisely aligned with the window aperture. Illumination was provided by two integrated tungsten infrared lamps within the reflection sampling module. Each position was scanned three times, and the average spectrum was retained as the representative experimental spectrum for that sample. The lightweight spectrometer (weighing <100 g) was operated via the DLP NIR-scan Nano software installation in Android system, which enabled pre-scanning configuration and calibration to ensure measurement consistency.

 doi:10.20944/preprints202602.2044.v1

Following spectral data acquisition, a 1 cm×1 cm×1 cm tissue block was excised from each measurement position. Juice was extracted from the tissue block using a press juicer and applied to the measuring window of a digital fruit sugar meter (PAL-1, Atago Co., Japan, Figure 1(b)). After a 10-second stabilization period, the SSC value was recorded. This measurement process was repeated three times for each tissue block, and the mean value was used as the reference SSC for the corresponding position.

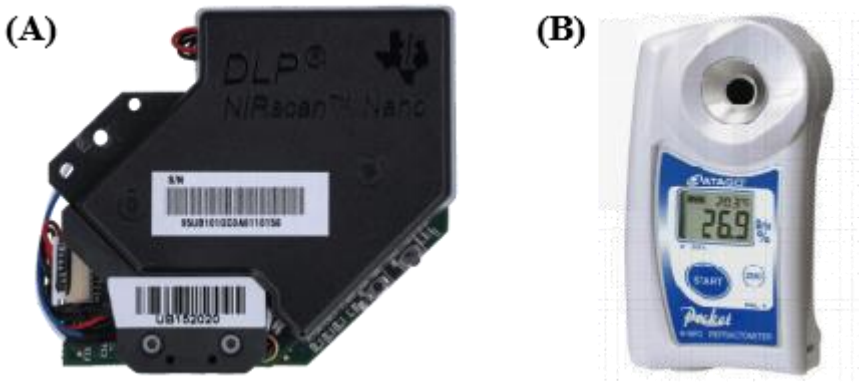


Figure 1. Photos of the portable spectrometer (DLP NIR-scan Nano EVM, USA) and digital fruit sugar meter (PAL-1, Japan).

2.3. Sample Set Division

The dataset was split into calibration and validation sets at a 4:1 ratio using two partitioning algorithms: the KS and SPXY algorithm. The KS algorithm was employed to ensure uniform spatial distribution of samples within the calibration set, thereby enhancing the model's generalizability. However, a limitation of KS is that it only considers spectral data when calculating Euclidean distances, neglecting the influence of physicochemical parameters (e.g., SSC values) on sample distribution. In contrast, the SPXY algorithm incorporates both spectral data and physicochemical values into distance calculations, enabling a multidimensional characterization of sample distribution and improving model robustness. It should be noted that the SPXY algorithm involves a more complex and computationally intensive spatial distance calculation process compared to KS. The sample set division strategy adopted a weighted scoring mechanism to balance model performance and data representativeness: 70% of the weight was assigned to the test set R^2 to evaluate predictive accuracy, while 30% was allocated to assessing the distributional similarity between the training and test sets. This design enabled the selection of sample splitting methods that ensured both high prediction precision and strong generalization capability, effectively mitigating the risks of overfitting and sampling bias.

2.4. Spectra Preprocess

To enhance the performance of the predictive model and improve the signal-to-noise ratio of the spectral data, seven spectral preprocessing techniques were applied: Savitzky-Golay (S-G) convolution smoothing, first and second derivatives, mean normalization, baseline offset correction (BLO), standard normal variable transformation (SNV), and multiple scattering correction (MSC).

The S-G smoothing method, combined with the first and second derivative calculations, effectively reduces system noise. Mean normalization is utilized to minimize the impact of spectral scattering during measurement. BLO was applied to adjust spectral baseline offsets. Both SNV and MSC were used to eliminate the influence of particle-induced surface scattering on diffuse reflection and to further correct baseline variations. These preprocessing steps were implemented to reduce the influence of external factors on spectral data, thereby improving the accuracy and stability of the predictive model.

 doi:10.20944/preprints202602.2044.v1

2.5. Feature Variable Selection

Prediction models based on NIRS can exhibit instability due to the relatively small sample size and the large number of spectral variables[23] [10]. To address this challenge, characteristic variable selection was employed to identify spectral variables with high interpretability, reduce model dimensionality, enhance computational efficiency, and improve model robustness. In this study, four feature variable selection algorithms were utilized: the successive projections algorithm (SPA), the competitive adaptive reweighting sampling (CARS), Monte Carlo uninformative variable elimination (MC-UVE), and the Random Frog (RF) algorithm. These algorithms were applied to process the full-band spectrum and evaluate their effectiveness in feature variable selection for the prediction model. Each method has its own advantages and limitations; a comprehensive comparison was conducted to determine the most suitable approach for the objectives of the study.

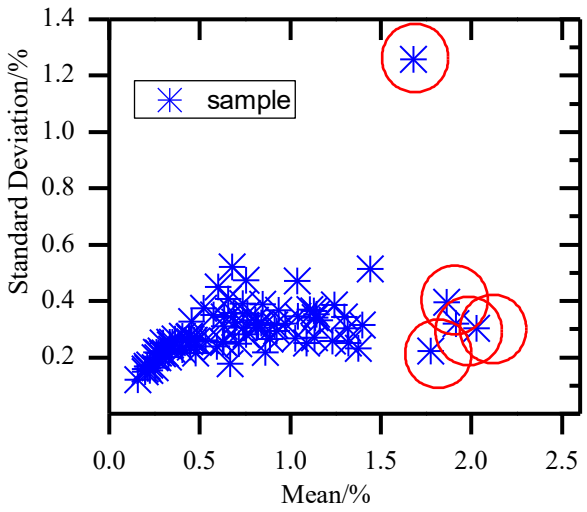
2.6. Model and Evaluation

Partial least squares regression (PLSR) is a standard method for quantitative analysis when predictor variables are collinear. In this study, PLSR was used to extract up to 20 latent variables (LVs) from the spectra to reduce dimensionality. Model performance for predicting SSC from diffuse reflectance spectra was evaluated using four metrics: the calibration correlation coefficient (R^2c), the prediction correlation coefficient (R^2p), the root mean square error of calibration (RMSEC), and the root mean square error of prediction (RMSEP). R^2c and R^2p reflect model fit and predictive ability (values closer to 1 indicate better performance), while RMSEC and RMSEP quantify calibration and prediction errors (lower values are better). A robust PLSR model therefore exhibits R^2c and R^2p near 1 and low, closely matched RMSEC and RMSEP[30]. .

3. Results

3.1. Sample Elimination and Dataset Division

Figure 2 shows the mean prediction errors for SSC with their standard deviations. To improve model accuracy, five samples (Nos. 43, 46, 55, 59 and 67) with unusually large mean errors or standard deviations were identified as outliers and excluded from further analysis. Table 1 reports the PLSR results from leave-one-out cross-validation for the full dataset (87 samples) and for the dataset after removing these five outliers. Excluding the outliers had a clear impact: the Rc rose from 0.795 to 0.873, indicating better stability and fit, while RMSEC fell from 0.6548% to 0.5221%, reflecting reduced calibration error. These findings highlight the importance of outlier detection and removal for improving the robustness and reliability of PLSR models for SSC prediction from diffuse reflectance spectra.



 doi:10.20944/preprints202602.2044.v1

Figure 2. The distributions of mean values and standard deviations of SSC prediction error.

Table 1. PLSR model before and after removing outliers.

Sample number	R^2_c	RMSEC/%	R^2_p	RMSECV/%
87	0.795	0.6548	0.6343	0.8886
82	0.873	0.5221	0.6992	0.8105

Sample set division involves partitioning the raw dataset into subsets based on specific characteristics, such as training and testing in machine learning, cross-validation, or stratified sampling. In this study, the KS algorithm and SPXY algorithm are employed to divide all samples into calibration and validation sets at a 4:1 ratio. Figure 3 illustrates the distribution of SSC reference values for all samples. The SSC values of 82 samples followed a normal distribution, with a mean of 13.17% and a standard deviation of 1.43%, ranging from 10.2% to 17.3%, thereby covering a wide interval. Table 2 compares the SSC values of the calibration and validation sets, obtained using two partitioning methods. The calibration set derived by the KS method shows minimal differences among samples but has a relatively narrow range. In contrast, the SPXY method yields a broader calibration set range, though the corresponding validation set is limited, which may restrict a comprehensive evaluation of model performance.

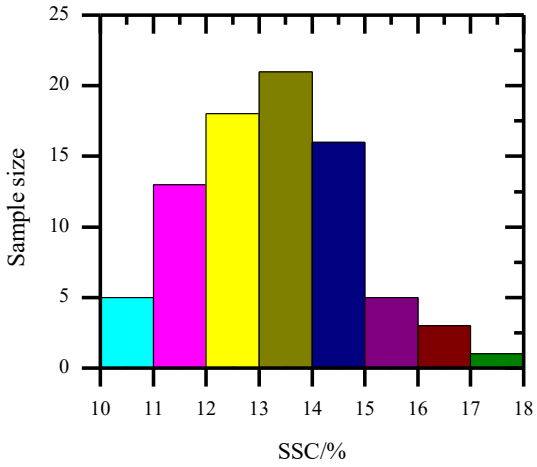


Figure 3. The SSC distribution of all samples.

Table 2. The Calibration set and validation set SSC (%) distribution.

Approach	Calibration Set				Validation Set			
	Number	Range	Mean	Standard deviation	Number	Range	Mean	Standard deviation
KS	65	10.2-16.3	13.19	1.42	17	10.8-17.3	13.1	1.49
SPXY	65	10.2-17.3	13.26	1.48	17	10.8-14.5	12.79	1.20
ALL	65	10.2-17.3	13.17	1.43	-	-	-	-

 doi:10.20944/preprints202602.2044.v1

3.2. Raw Spectral Feature

Figure 4 presents the absorption spectra of the samples in the range of 900 to 1700 nm. The spectra display a consistent pattern with three distinct absorption peaks located near 970 nm, 1200 nm, and 1450 nm[31]. The band at 1450 nm corresponds to the first O–H stretching overtone, so the peak at 1450 nm is the most pronounced and corresponds to the absorption band of water, while the other two peaks are associated with O-H and C-H groups, respectively as shown in Table 3[16].

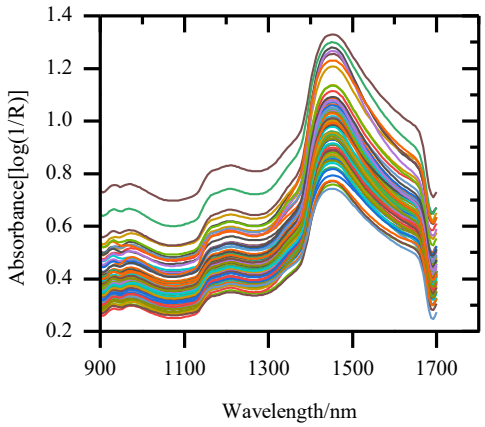


Figure 4. The near infrared diffuse reflectance spectra of all samples.

Table 3. Correlation between characteristic wavelengths and apple sugar functional groups.

Characteristic wavelength (nm)	Corresponding functional group	Associated sugar type	Vibration mode
970	O-H	Glucose	Bending vibration
1200	C-H	Sucrose	Stretching vibration
1450	O-H	Water	Stretching vibration

As shown in Table 3, the 970 nm wavelength corresponds to the O-H bending vibration of glucose, which is one of the primary monosaccharide components of apple soluble solid content (SSC)[16,17]. The 1200 nm wavelength corresponds to the characteristic peak of the C-H stretching vibration of sucrose. As the most abundant disaccharide in apples, sucrose contributes to the strongest correlation between this wavelength and SSC, thereby rendering it the core characteristic wavelength for SSC prediction in this study.

The 1450 nm wavelength corresponds to the absorption peak of the O-H stretching vibration of water in apple tissue and exerts a negative effect on the accuracy of SSC determination. This is because an increase in apple SSC is typically accompanied by a decrease in tissue water content, which in turn leads to the attenuation of the absorption signal at 1450 nm. The aforementioned analysis of spectral absorption mechanisms verifies the rationality of the characteristic wavelengths selected by the model and also explains the variability in the model's adaptability to samples with different sugar compositions, which is consistent with the findings reported in the literature[23]. Accordingly, these specific wavelength components were selected as key markers for characterizing apple SSC.

3.3. Comparison and Analysis of Pre-Processioning Methods

Spectral preprocessing plays a crucial role in reducing systematic noise in the raw spectra and enhancing the performance of prediction models. Utilizing the KS sample set partitioning method, the PLSR model preprocessed with the second derivative preprocessing achieved the best

 doi:10.20944/preprints202602.2044.v1

performance. The predictive performance of PLSR models with different preprocessing methods is listed in Table 4.

Table 4. PLSR model performance for apple sugar content after spectral preprocessing.

Sample set		SPXY					KS			
pre-processing	LVs	R^2_c	RMSEC /%	R^2_p	RMSEP /%	LVs	R^2_c	RMSEC /%	R^2_p	RMSEP /%
Raw	11	0.8758	0.5161	0.6974	0.6921	13	0.9307	0.3718	0.8459	0.5700
S-G	11	0.8494	0.5682	0.7053	0.6829	12	0.8942	0.4594	0.8408	0.5794
1st derivative	9	0.8725	0.5228	0.763	0.6125	14	0.8889	0.4708	0.8629	0.5377
2nd derivative	13	0.9411	0.3623	0.8095	0.5492	9	0.9160	0.4094	0.8632	0.5370
Mean normalization	9	0.8140	0.6315	0.7128	0.6743	13	0.9433	0.3363	0.7707	0.6917
BLO	12	0.8805	0.5062	0.7161	0.6704	14	0.9463	0.3272	0.8387	0.5831
SNV	11	0.8704	0.5272	0.6610	0.7325	14	0.9408	0.3437	0.7650	0.7083
MSC	11	0.8628	0.5426	0.6202	0.7753	13	0.9298	0.3743	0.7426	0.7366

According to Table 4, the 12-point second derivative preprocessing method yielded the most significant improvement in prediction model performance under both SPXY and KS sample set divisions compared to other methods. It is worth noting that while SPXY has superior model stability, KS has better prediction ability and significantly lower computational cost. Therefore, according to the results obtained, it seems that the PLSR model preprocessed with the second derivative (12 points) performs better than other models when using the KS sample set division mode, with R^2c and RMSEC of 0.916 and 0.4094%, and R2p and RMSEP of 0.8632 and 0.5370%, respectively.

3.4. Analysis of Feature Selection Methods

NIRS signals contain numerous spectral variables, including both informative variables and uninformative ones. Incorporating irrelevant variables not only increases the computational complexity of the calibration model but also reduces prediction accuracy. Therefore, the selection of characteristic wavelengths is essential to eliminate non-informative variables, thereby simplifying the model and enhancing its predictive performance. In this section, characteristic wavelength extraction methods are applied to reduce dimensionality and improve the accuracy of SSC prediction.

3.4.1. SPA

SPA is a deterministic and reproducible variable selection strategy for multivariate calibration. By performing simple vector space operations, it effectively minimizes collinearity among variables. Compared with the genetic algorithm, SPA is regarded as more robust[32]. In this study, the PLSR model was constructed using the leave-one-out cross-validation method, and the subset of variables that achieved the lowest RMSECV was identified as optimal. Figure 5 illustrates the variation of RMSECV during cross-validation using the SPA-PLS method. As the number of selected wavelengths

 doi:10.20944/preprints202602.2044.v1

increases, the RMSECV initially rises, then rapidly decreases, and eventually stabilizes. The best performance was obtained when SPA selected 11 characteristic variables to construct the PLSR model, yielding an RMSECV of 0.6911. The characteristic wavelengths selected by SPA are shown in Figure 6.

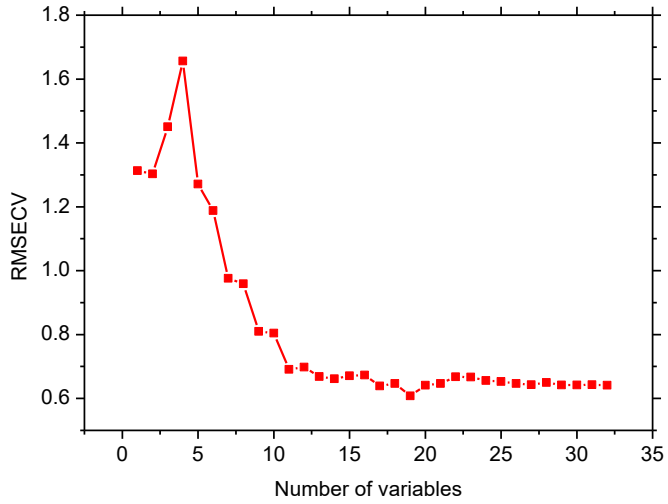


Figure 5. RMSECV using the SPA-PLS method.

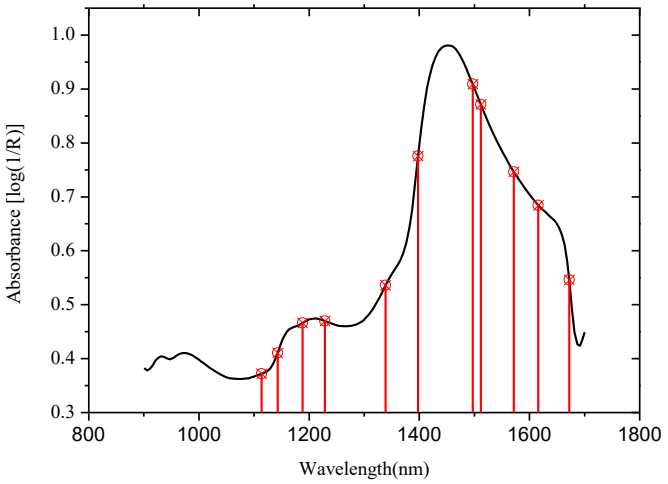


Figure 6. Characteristic wavelengths selected by SPA.

3.4.2. CARS

The CARS method was designed to extract the optimal combination of wavelengths from the full spectrum in combination with PLSR[33]. Inspired by Darwin's theory of evolution, this method applies the principle of natural selection to variable selection. In this study, the MC sampling count was set to 100, and the results of the CARS algorithm are presented in Figure 7. As shown in Figure 7(a), the number of variables decreases rapidly at first and then more gradually within the initial 10 MC samplings. This trend results from the exponential decreasing function, which performs variable selection in two stages: rapid elimination followed by refined selection. Figure 7(b) shows that the RMSECV value decreases as the number of MC samplings increases, then rises again, reaching a minimum value of 0.6949 at the 41st sampling, where 18 characteristic variables were selected. The characteristic wavelengths identified by CARS are displayed in Figure 8.

 doi:10.20944/preprints202602.2044.v1

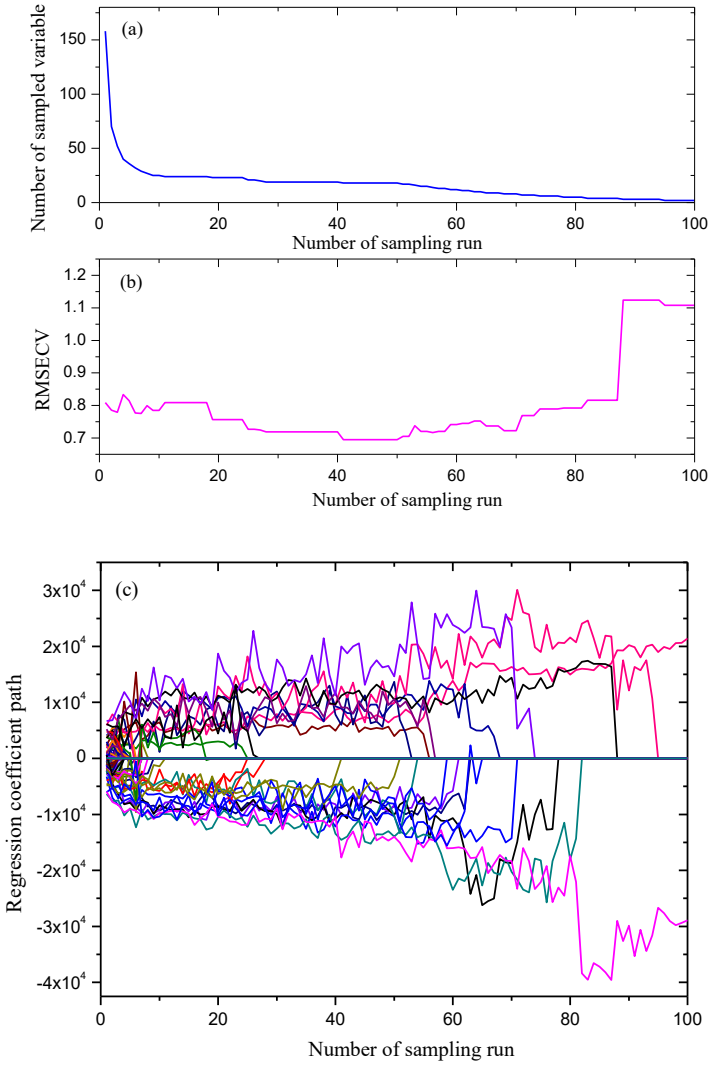


Figure 7. the MC results of the CARS algorithm.

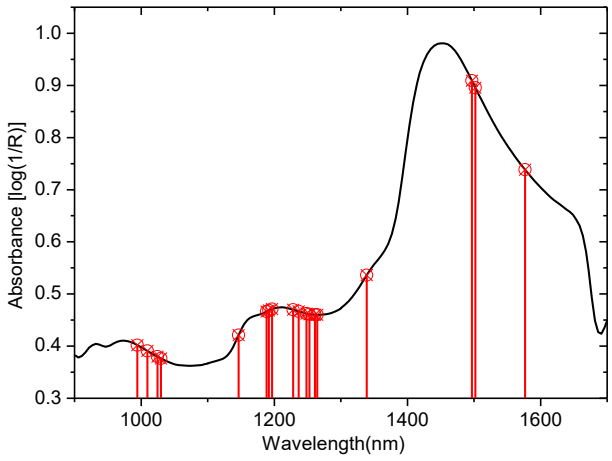


Figure 8. Characteristic wavelengths selected by CARS.

 doi:10.20944/preprints202602.2044.v1

3.4.3. MC-UVE Algorithm

MC-UVE is a variable selection method that integrates the principles of MC with uninformative variable elimination[30]. In this method, the MC strategy replaces the traditional leave-one-out strategy in the UVE-PLS process. Compared with UVE, which estimates the cutoff threshold by introducing random noise variables into the original data matrix, MC-UVE provides a more efficient strategy for wavelength selection.

In this study, the MC sampling number was set to 500, with 80% of the calibration set samples randomly selected to form calibration subsets. A key step in the algorithm is the determination of the stability threshold: variables exceeding this threshold are regarded as characteristic wavelengths. Figure 9 illustrates the variation of RMSECV in cross-validation of the PLSR model, with Nj increased in increments of 20 up to 140. The minimum RMSECV of 0.5911 was achieved at Nj=40, corresponding to a stability threshold of 4.3918. Figure 10 presents the stability distribution of each spectral variable, where those surpassing the threshold were identified as characteristic variables, resulting in 40 selected wavelengths.

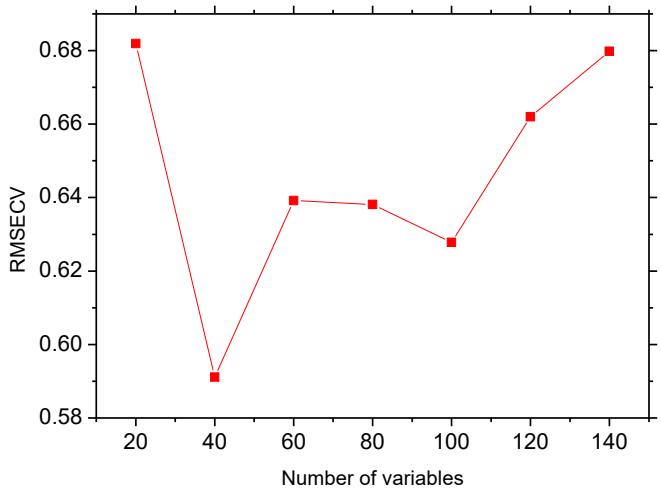


Figure 9. RMSECV for MC-UVE algorithm.

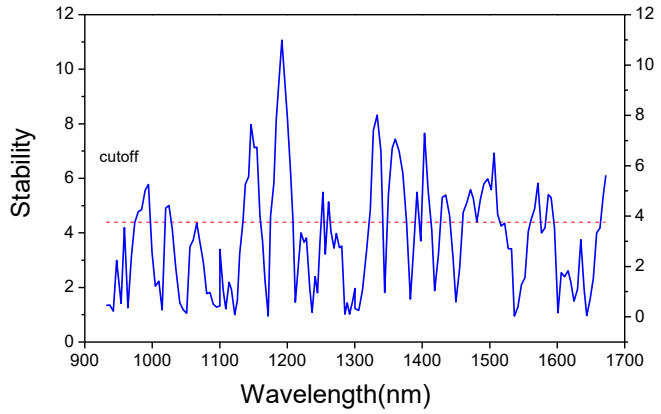


Figure 10. Stability with wavelength ranging for 900 to 1700 nm.

3.4.4. RF

To overcome the challenges of extensive computation and complex mathematics, the authors propose the RF as an efficient and user-friendly method for selecting disease-associated features[34].

 doi:10.20944/preprints202602.2044.v1

In this study, the number of variable subsets q was set to 2 and the number of loops to 10,000 for the RF algorithm. The algorithm calculates the probability of each wavelength being selected, with a higher probability indicating greater importance. A key step is determining the probability threshold: variables with probabilities exceeding this threshold are regarded as characteristic wavelengths. As shown in Figure 11, the variation in RMSECV was obtained by cross-validating the PLSR model with Nj set in increments of 20 up to 120. The minimum RMSECV of 0.7070 occurred at Nj=120, corresponding to a probability threshold of 0.0007. Based on this criterion, 120 wavelengths were identified as characteristic variables.

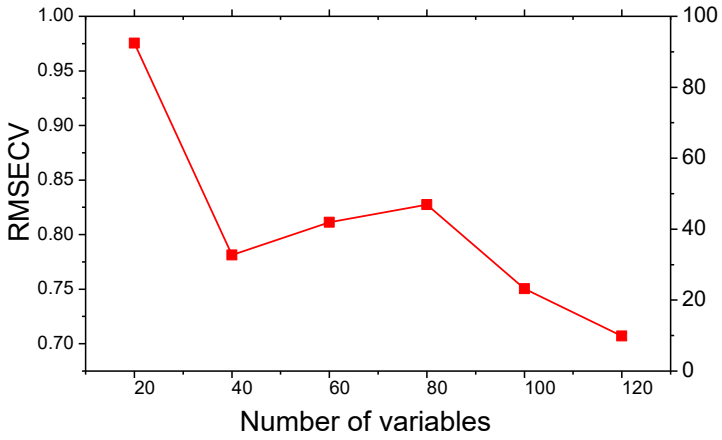


Figure 11. RMSECV by using cross-validating the PLSR model.

3.4.5. Comparison Results

Table 5 summarizes the results of the four wavelength selection methods utilized in this study. Each method employed leave-one-out cross-validation to establish the PLSR model, with the number of characteristic wavelengths determined by the minimum RMSECV. Among them, the number of significant wavelengths selected by SPA was the lowest, with only 11 wavelengths, accounting for 6.96% of the entire spectrum. In contrast, MC-UVE selected 40 characteristic wavelengths and achieved the lowest RMSECV of 0.5911%, demonstrating superior performance compared with the other three methods.

Table 5. The Results of different characteristic wavelength selection methods.

Approach	Number of characteristic wavelengths	RMSECV/%
SPA	11	0.6911
CARS	18	0.6949
MC-UVE	40	0.5911
RF	120	0.7070

The PLSR models were established using characteristic spectral variables obtained from preprocessed spectra by employing different characteristic wavelength selection methods. Table 6 compares and evaluates the model performance. Among these methods, preprocessing alone displayed the most significant spectral effect, with R^2c, RMSEC, R^2p, and RMSEP of 0.916, 0.4093 %, 0.8632, and 0.537 %, respectively, as illustrated in Figure 12. Moreover, the MC-UVE algorithm produced a more computationally efficient model, achieving R^2c = 0.878 and RMSEC = 0.4933%. Although prediction metrics were slightly lower (R^2p = 0.8159 and RMSEP = 0.623%) compared with the preprocessing-only approach, MC-UVE reduced the number of latent variables and simplified the model, thereby lowering computational cost.

 doi:10.20944/preprints202602.2044.v1

Table 6. The PLSR model performance based on different characteristic wavelength selection methods.

Model	Feature selection	Variables	LVs	R^2_c	RMSEC/%	R^2_p	RMSEP/%
PLSR	**NONE**	**168**	**9**	**0.916**	**0.4093**	**0.8632**	**0.5370**
	SPA	11	8	0.8381	0.5683	0.7702	0.6961
	CARS	18	6	0.8283	0.5853	0.5839	0.9365
	MC-UVE	40	6	0.8780	0.4933	0.8159	0.623
	RF	120	9	0.9060	0.4329	0.8346	0.5905

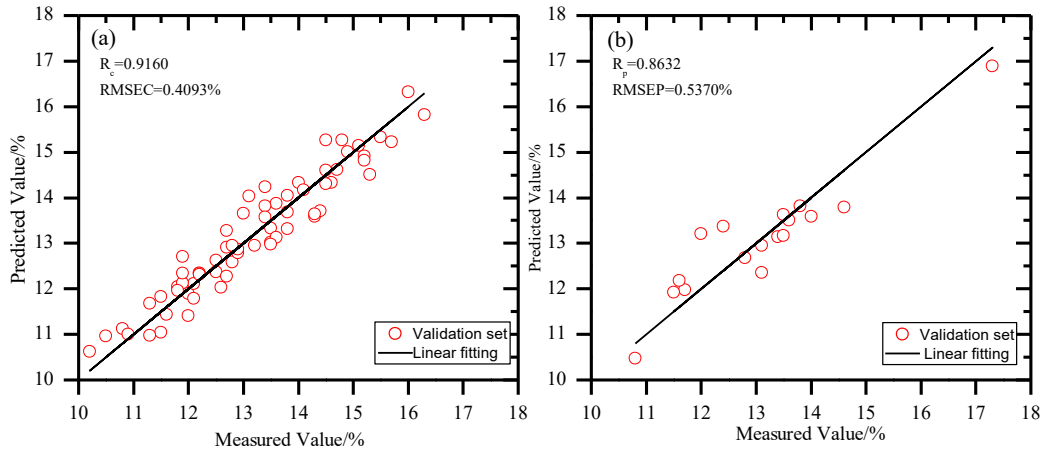


Figure 12. PLSR calibration model (a) and prediction results (b) after second derivative preprocessing..

4. Conclusions

In this study, near-infrared diffuse reflectance spectroscopy was applied to predict the SSC of 'Red Fuji' apples using a PLSR model. Outlier elimination through the Monte Carlo method effectively improved model stability. Among different preprocessing techniques, the second derivative (12 points) yielded the best performance for both models. The preprocessed PLSR model achieved R^2c = 0.916 and RMSEC = 0.4093%, with R^2p = 0.8632 and RMSEP = 0.537%. For PLSR combined with variable selection, the MC-UVE algorithm reduced the number of latent variables to six while maintaining acceptable prediction accuracy. Although variable selection did not improve predictive accuracy, it reduced model complexity and enhanced computational efficiency. Overall, these results confirm that near-infrared spectroscopy provides an accurate and non-destructive approach for detecting apple sugar content. Future research should focus on two key directions: developing portable, cost-effective instruments for field applications and exploring advanced neural network architectures to further improve predictive accuracy.

Author Contributions: Conceptualization, Xuanbing Qiu. and Ye Teng .; methodology, Chencong MA and Tianhan Wang; formal analysis and original draft writing, Tianhan Wang, review and editing, Xiangjun Xu ; supervision, Xuanbing Qiu., Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant/award numbers: 52076145 and 12304403), the Key research plan of Shanxi Province (Number: 202402150301012), Fundamental Research Program of Shanxi Province (Number: 202203021222204 and 202303021212224).

 doi:10.20944/preprints202602.2044.v1

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest..

References

1. Tian, X.; Li, J.; Wang, Q.; Fan, S.; Huang, W.; Zhao, C. A multi-region combined model for non-destructive prediction of soluble solids content in apple, based on brightness grade segmentation of hyperspectral imaging. *Biosyst. Eng.* **2019**, *183*, 110-120.
2. Rong, Y.; Zareef, M.; Liu, L.; Din, Z.U.; Chen, Q.; Ouyang, Q. Application of portable Vis-NIR spectroscopy for rapid detection of myoglobin in frozen pork. *Meat Sci.* **2023**, *201*, 109170.
3. Wu, J.; Zareef, M.; Chen, Q.; Ouyang, Q. Application of visible-near infrared spectroscopy in tandem with multivariate analysis for the rapid evaluation of matcha physicochemical indicators. *Food Chem.* **2023**, *421*, 136185.
4. Ouyang, Q.; Rong, Y.; Wu, J.; Wang, Z.; Lin, H.; Chen, Q. Application of colorimetric sensor array combined with visible near-infrared spectroscopy for the matcha classification. *Food Chem.* **2023**, *420*, 136078.
5. Li, S.; Li, J.; Wang, Q.; Shi, R.; Yang, X.; Zhang, Q. Determination of soluble solids content of multiple varieties of tomatoes by full transmission visible-near infrared spectroscopy. *Front. Plant Sci.* **2024**, *15*, 1324753.
6. Liu, L.; Zareef, M.; Wang, Z.; Li, H.; Chen, Q.; Ouyang, Q. Monitoring chlorophyll changes during Tencha processing using portable near-infrared spectroscopy. *Food Chem.* **2023**, *412*, 135505.
7. Wu, X.; Fang, Y.; Wu, B.; Liu, M. Application of near-infrared spectroscopy and fuzzy improved null linear discriminant analysis for rapid discrimination of milk brands. *Foods* **2023**, *12*, 3929.
8. Li, Q.; Wu, X.; Zheng, J.; Wu, B.; Jian, H.; Sun, C.; Tang, Y. Determination of pork meat storage time using near-infrared spectroscopy combined with fuzzy clustering algorithms. *Foods* **2022**, *11*, 2101.
9. Li, H.; Zhang, W.; Nunekpeku, X.; Sheng, W.; Chen, Q. Investigating the change mechanism and quantitative analysis of minced pork gel quality with different starches using Raman spectroscopy. *Food Hydrocolloids* **2025**, *159*, 110634.
10. Jiang, H.; Wang, Z.; Deng, J.; Ding, Z.; Chen, Q. Quantitative detection of heavy metal Cd in vegetable oils: A nondestructive method based on Raman spectroscopy combined with chemometrics. *J. Food Sci.* **2024**, *89*, 8054-8065.
11. Monago-Maraña, O.; Afseth, N.K.; Knutsen, S.H.; Wubshet, S.G.; Wold, J.P. Quantification of soluble solids and individual sugars in apples by Raman spectroscopy: A feasibility study. *Postharvest Biol. Technol.* **2021**, *180*, 111620, doi:https://doi.org/10.1016/j.postharvbio.2021.111620.
12. Zahidah, I.; Bölek, S.; Terzioğlu, Ö.T.; Adıgüzel, S. Determination of the effects of novel paraprobiotic supplement of Lactobacillus plantarum on soy dairy-free beverage by physicochemical, antioxidant, sensory analyses, and Raman spectroscopy technique. *J. Food Sci.* **2024**, *89*, 7189-7202.
13. Huang, Y.; Dong, W.; Chen, Y.; Wang, X.; Luo, W.; Zhan, B.; Liu, X.; Zhang, H. Online detection of soluble solids content and maturity of tomatoes using Vis/NIR full transmittance spectra. *Chemometrics and Intelligent Laboratory Systems* **2021**, *210*, 104243.
14. Choi, J.-H.; Chen, P.-A.; Lee, B.; Yim, S.-H.; Kim, M.-S.; Bae, Y.-S.; Lim, D.-C.; Seo, H.-J. Portable, non-destructive tester integrating VIS/NIR reflectance spectroscopy for the detection of sugar content in Asian pears. *Sci. Hortic.* **2017**, *220*, 147-153.
15. Zontov, Y.; Balyklova, K.; Titova, A.; Rodionova, O.Y.; Pomerantsev, A. Chemometric aided NIR portable instrument for rapid assessment of medicine quality. *J. Pharm. Biomed. Anal.* **2016**, *131*, 87-93.
16. Fodor, M.; Matkovits, A.; Benes, E.L.; Jókai, Z. The role of near-infrared spectroscopy in food quality assurance: A review of the past two decades. *Foods* **2024**, *13*, 3501.
17. Grabska, J.; Beć, K.B.; Ueno, N.; Huck, C.W. Analyzing the quality parameters of apples by spectroscopy from Vis/NIR to NIR region: A comprehensive review. *Foods* **2023**, *12*, 1946.
18. Yan-de, L.; Hai, X.; Xu-dong, S.; Xiao-gang, J.; Yu, R.; Yu, Z. Development of multi-cultivar universal model for soluble solid content of apple online using near infrared spectroscopy. *Spectroscopy and Spectral Analysis* **2020**, *40*, 922-928.

 doi:10.20944/preprints202602.2044.v1

19. Yuan, L.-m.; Cai, J.-r.; Sun, L.; Han, E.; Ernest, T. Nondestructive measurement of soluble solids content in apples by a portable fruit analyzer. *Food Anal. Methods* **2016**, *9*, 785-794.
20. Guo, Z.; Wang, M.; Agyekum, A.A.; Wu, J.; Chen, Q.; Zuo, M.; El-Seedi, H.R.; Tao, F.; Shi, J.; Ouyang, Q. Quantitative detection of apple watercore and soluble solids content by near infrared transmittance spectroscopy. *J. Food Eng.* **2020**, *279*, 109955.
21. Wang, T.; Chen, J.; Fan, Y.; Qiu, Z.; He, Y. SeeFruits: Design and evaluation of a cloud-based ultra-portable NIRS system for sweet cherry quality detection. *Comput. Electron. Agric.* **2018**, *152*, 302-313.
22. Lanjewar, M.G.; Morajkar, P.P.; Parab, J.S. Portable system to detect starch adulteration in turmeric using NIR spectroscopy. *Food Control* **2024**, *155*, 110095.
23. Yao, Y.-n.; Ma, K.; Zhu, J.; Huang, F.; Kuang, L.; Wang, X.; Li, S. Non-destructive determination of soluble solids content in intact apples using a self-made portable NIR diffuse reflectance instrument. *Infrared Physics & Technology* **2023**, *132*, 104714.
24. Li, Z.-Y.; Huang, X.; Yang, J.-X.; Luo, S.-H.; Wang, J.; Fang, Q.-L.; Hui, A.-L.; Liang, F.-X.; Wu, C.-Y.; Wang, L. An improved 1D CNN with multi-sensor spectral fusion for Detection of SSC in pears. *J. Food Compos. Anal* **2025**, *144*, 107732.
25. Sun, X.; Du, Y.; Nawaz, M.A.; Abobatta, W.F.; Lyu, Q.; Liu, J.; Chen, Z.; Feng, S. Apple SSC estimation using hand-held NIRS instrument for outdoor measurement with ambient light correction. *Postharvest Biol. Technol.* **2024**, *217*, 113101.
26. Elamshity, M.G.; Alhamdan, A.M. Development and Prediction of a Non-Destructive Quality Index (Qi) for Stored Date Fruits Using VIS–NIR Spectroscopy and Artificial Neural Networks. *Foods* **2025**, *14*, 3060.
27. Tian, Y.; Sun, J.; Zhou, X.; Cong, S.; Dai, C.; Shi, L. Nondestructive Detection of Soluble Solids Content in Apples Based on Multi-Attention Convolutional Neural Network and Hyperspectral Imaging Technology. *Foods* **2025**, *14*, 3832.
28. Zhao, J.; Hu, Q.; Li, B.; Xie, Y.; Lu, H.; Xu, S. Research on an Improved Non-Destructive Detection Method for the Soluble Solids Content in Bunch-Harvested Grapes Based on Deep Learning and Hyperspectral Imaging. *Applied Sciences* **2023**, *13*, 6776.
29. Zhang, X.; Qin, Z.; Zhao, R.; Xie, Z.; Bai, X. A Handheld IoT Vis/NIR Spectroscopic System to Assess the Soluble Solids Content of Wine Grapes. *Sensors* **2025**, *25*, 4523.
30. Li, J.; Huang, W.; Chen, L.; Fan, S.; Zhang, B.; Guo, Z.; Zhao, C. Variable selection in visible and near-infrared spectral analysis for noninvasive determination of soluble solids content of 'Ya'pear. *Food Anal. Methods* **2014**, *7*, 1891-1902.
31. Li, X.; Ma, L.; Bi, S.; Shen, T. Internal Quality Classification of Apples Based on Near Infrared Spectroscopy and Evidence Theory. In Proceedings of the Proceedings of the 11th International Conference on Computer Engineering and Networks, 2021; pp. 321-330.
32. Araújo, M.C.U.; Saldanha, T.C.B.; Galvao, R.K.H.; Yoneyama, T.; Chame, H.C.; Visani, V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and intelligent laboratory systems* **2001**, *57*, 65-73.
33. Li, H.; Liang, Y.; Xu, Q.; Cao, D. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* **2009**, *648*, 77-84.
34. Li, H.-D.; Xu, Q.-S.; Liang, Y.-Z. Random frog: An efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification. *Anal. Chim. Acta* **2012**, *740*, 20-26