agriculture

MDPI

Article

Detection of Soluble Solids Content (SSC) in Pears Using Near-Infrared Spectroscopy Combined with LASSO–GWF–PLS Model

Baishao Zhan, Peng Li, Ming Li, Wei Luo and Hailiang Zhang *

College of Electrical and Automation Engineering, East China Jiaotong University, Nanchang 330013, China; 3050@ecjtu.edu.cn (B.Z.); 15270030556@163.com (W.L.)
* Correspondence: hailiang.zhang@163.com

Abstract: The soluble solids content (SSC) of pears is mainly composed of sugars, organic acids, and other soluble substances and is one of the important indices used to measure the sweetness and quality of pear juice. The SSC of pears is mainly composed of sugars, organic acids, amino acids, esters, alcohols, phenols, flavonoids, and other compounds, and different groups within these compounds have different characteristic absorption peaks corresponding to different characteristic wavelengths. Traditional methods such as genetic algorithm (GA) and competitive adaptive reweighted sampling (CARS) models used for screening characteristic wavelengths are mainly based on statistical methods, and characteristic wavelengths are selected by finding the wavelengths related to the changes in the concentration of the target analytes. By ignoring the molecular structure and chemical properties of the target analytes and disregarding the influence of the groups of the compounds in the target analytes on the spectral characteristics, wavelengths that are not related to the target analytes may be selected, thus affecting the accuracy of the analytical results. In this paper, a partial least squares (PLS) model was established based on the characteristic wavelengths of CARS, GA, and LASSO algorithms, and the best least absolute shrinkage and selection operator (LASSO) was selected and compared with the characteristic wavelengths selected by group weighted fusion (GWF). The LASSO regression was validated by 10-fold cross-validation to select the appropriate regularization parameter, and the 33 characteristic wavelengths correlated with the SSC of pears were selected in the full spectral range, and the 9 characteristic wavelengths corresponding to the group response were weighted and fused and input into the PLS regression model. Using an established model, the coefficient of determination (R^2) and the root mean square error (RMSE) of the calibration set were 0.992 and 0.177%, respectively, and the R^2 and RMSE of the test set were 0.998 and 0.128%, respectively. The R^2 of our LASSO–GWF–PLS prediction model was improved from 0.975 to 0.998, indicating that the LASSO–GWF–PLS method has very good prediction ability for detection of SSC in pears.

Keywords: near-infrared spectroscopy; soluble solids content; pear; fusion variable selection algorithm; modeling

Citation: Zhan, B.; Li, P.; Li, M.; Luo, W.; Zhang, H. Detection of Soluble Solids Content (SSC) in Pears Using Near-Infrared Spectroscopy Combined with LASSO–GWF–PLS Model. *Agriculture* **2023**, *13*, 1491. https://doi.org/10.3390/agriculture13081491

Academic Editors: Baohua Zhang, Zhiming Guo and Jiangbo Li

Received: 9 July 2023
Revised: 24 July 2023
Accepted: 25 July 2023
Published: 27 July 2023

1. Introduction

Soluble solids content (SSC) is an important index used to measure the commercial quality of pear fruit. SSC refers to the content of soluble solids in juice, including sugars, organic acids, amino acids, and other soluble substances [1–3]. The level of SSC directly affects the taste, sweetness, nutritional value, and quality of pears. The traditional methods used to determine SSC are the extraction method and the density method, and these methods require a great deal of time and effort and are easily affected by environmental and operational factors, resulting in less accurate and less stable results [4]. In recent years, with the development of near-infrared spectroscopy (NIRS), more and more studies have started to explore the use of NIR spectroscopy for the rapid and accurate determination of SSC. This method has the advantages of being non-destructive, rapid, efficient, accurate, and

reproducible, and has become one of the most important tools in the evaluation of the commercial quality of pear fruit. Vis–NIRS is a rapid and non-destructive detection technique which integrates visible and near-infrared spectroscopy. It obtains spectral information from a sample by detecting its reflectance spectra in the visible and near-infrared bands. This spectral information allows a quantitative relationship to be established between the sample spectrum and the soluble solids content [5,6]. In the determination of SSC of pears, the NIRS technique can be used to achieve a rapid and accurate determination by using an established quantitative relationship model.

A large number of studies have improved the model's accuracy by picking feature wavelengths and model combinations, etc. Chen et al. [7] optimized the feature wavelengths and parameters of the corresponding models using a competitive adaptive reweighted sampling method and a hybrid wind-driven differential evolutionary algorithm, and the wavelength variables of the optimized partial least squares (PLS) and least squares support vector machine (LS-SVM) models were reduced to 8.67% and 67.80% of the full spectrum, respectively. The calibration coefficient of determination (R_c^2), predicted coefficient of determination (R_P^2), root mean square error of calibration (RMSEC), root mean square error of prediction (RMSEP), and ratio of performance to deviation (RPD) of the two models were 0.9708, 0.9542, 0.2586, 0.2628, 5.91, and 0.9873; and 0.9830, 0.1705, 0.1734, and 8.96, respectively. Guo et al. [8] used NIR spectroscopy combined with data dimensionality reduction to develop a prediction model (competitive adaptive reweighted sampling–support vector machine, CARS–SVM) for the soluble solids content of cantaloupe. The correlation coefficient of the calibration set was 0.9814, and the correlation coefficient of the prediction set was 0.9002. The model was able to predict the soluble solids content of cantaloupe accurately. Zhang et al. [9] improved the network structure of a back propagation (BP) neural network by using the extraction of feature bands by CARS, and the performance of the optimized BP neural network model was significantly improved. The R^2 of the test set was improved by 0.4193 and the RMSE was reduced by 0.516 after the extraction of feature bands by CARS. These studies indicated that the accuracy and stability of NIR spectroscopy for detecting soluble solid content in fruits can be significantly improved through feature wavelength selection or model optimization methods based on spectral data. However, the existing methods only improve the accuracy of the model through mathematical analysis and other optimization methods, without considering the chemical structure information of the measured substance, which may lead to unfavorable optimization results.

The specific objectives of this study were to: (1) select a set of functional response characteristic wavelengths corresponding to the soluble solids content of pears based on the infrared characteristic absorption peaks of the functional groups; (2) weight the selected characteristic wavelengths and fuse with the characteristic wavelengths selected by mathematical analysis (LASSO algorithm); (3) establish the PLS model using the fused characteristic wavelength group to predict the SSC of pears by considering both the dimension reduction and optimization idea of mathematical analysis and the chemical properties; and (4) assess the performance of all models to obtain the best one for the determination of SSC in pears.

2. Materials and Methods

2.1. Samples

In this study, all pear samples were purchased from the local fruit market in Nanchang (China). Two varieties of pears including 'Cuiguan' and 'Huanghua' are used. On the one hand, they are widely cultivated in China; on the other hand, they have a greater difference in soluble solids content. Ripe and well-preserved pears were selected as samples. The samples had a uniform mature yellow–green color and did not have any skin defects or damage upon visual observation. A total of 240 pears (120 for each variety) were ultimately selected. We placed all samples under laboratory conditions with an ambient temperature of 20 °C and a relative humidity of 60% for 24 h to eliminate the impact of temperature on the accuracy of the prediction model.

2.2. SSC Measurement

Each pear was cut into small pieces with the volume of each small piece not exceeding 20 mm^3, and the sample pieces were placed into a manual juicer (L18-Y915S, Joyoung, Jinan, China) to make the pulp and then filtered using a centrifuge (LC-LX-L40B, Shanghai, China) running at 4000 rpm for 10 min. Then, the supernatant was poured into a clean beaker as the measurement solution and, next, the supernatant was poured into a specific gravity meter, and the value of SSC was obtained by reading the scale of the specific gravity meter. All experiments were performed at 25 °C, and air was completely prevented from entering the mixture during the measurement, which made the measurement results more accurate.

2.3. Spectral Data Acquisition and Preprocessing

Vis–NIR spectral data of pears were obtained by a desktop infrared spectrometer (ASD, Collegeville, PA, USA) (parameter settings: wavelength range of 350–1800 nm, spectral sampling interval of 1 nm, 10 scans, resolution of 3 nm, detection field angle of 25°, halogen lamp light source of 12 V/45 W). The measurement system was arranged in reflectance mode for collecting diffuse reflectance spectra from pears. Pears were placed steadily upon the fruit holder, with the stem–calyx axis arranged horizontally. Three separate spectral measurements were performed on each sample around the equator (120°) to decrease the errors of operator and instrument. The average spectrum of these three measurements was used to build the model. The spectral data were exported as ASCII codes in ASD ViewSpecPro 5.6.8 and then imported into Unscrambler V9.7 and Matlab 2021a software for data analysis and processing. Specifically, Unscrambler V9.7 software was used for spectral preprocessing and modeling. The statistics and machine learning toolbox in Matlab 2021a software was used for LASSO regression and wavelength selection and optimization.

Corrected spectral data are often disturbed and affected by various factors, such as instrument noise, baseline drift, light source fluctuation, etc. These factors can lead to noise and unnecessary fluctuations in the spectral data, which affect the quality of the spectral data and the analysis results. Therefore, the preprocessing of spectral data is required to improve the signal-to-noise ratio and analytical accuracy of spectral data. The SavitzkyGolay (SG) smoothing method is a commonly used signal-processing method used to smooth spectral data to remove noise and unnecessary fluctuations. In the SG preprocessing of spectral data, the raw spectral data were first baseline-corrected to remove background noise and baseline drift. Then, SG smoothing was performed on the baseline-corrected spectral data, using a window size of 21 and a polynomial order of 3.

2.4. Feature Wavelength Selection

2.4.1. Feature Wavelength Selection Using Different Algorithms

(1) Competitive adaptive reweighted sampling: The competitive adaptive reweighted sampling (CARS) algorithm is a method used for feature wavelength screening, and its basic idea is to select the optimal combination of wavelengths in the full spectrum by iterative competition [10–13]. Specifically, the CARS algorithm sequentially selects a subset of N wavelengths and uses an exponential decreasing function and adaptive reweighted sampling to determine the number of wavelengths selected for each sampling. Cross-validation is performed on each subset to calculate its minimum root mean square error of cross-validation, and the subset with the minimum RMSECV is selected as the optimal variable subset.

(2) Genetic algorithm: The genetic algorithm (GA) is an optimization algorithm based on the principle of biological evolution that continuously iterates to optimize the individuals in a population by performing operations such as selection, crossover, and mutation on the population and, finally, obtains the individual with the highest adaptation as the optimal solution [14–16]. The core idea of this algorithm is to pass on good individuals to the next generation while introducing randomness to ensure global search capability. The genetic algorithm has the advantages of global search capability, parallel computing capability, and self-adaptability and has a wide range of applications in solving complex

optimization problems. The basic process of this algorithm includes the steps of initializing the population, evaluating the fitness, performing a selection operation, a crossover operation, and a variation operation, and repeating iterations until the individual with the highest fitness is output as the optimal solution.

(3) Least absolute shrinkage and selection operator regression: Least absolute shrinkage and selection operator (LASSO) regression is a commonly used feature selection method that can compress the coefficients of some features to zero by L1 regularization of the coefficient matrix B [17–19]. The LASSO regression's least residual sum of squares expression is shown in Equation (1):

$$\underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{q}\left(y_i-\sum_{j=1}^{m}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{m}\left|\beta_j\right|\right\} \quad (1)$$

In the LASSO regression's least residual sum of squares expression, λ is a hyperparameter that is a 1-paradigm penalty coefficient for the regression coefficient j to control the complexity of the model. λ values serve to penalize the feature coefficients so that some feature coefficients become 0 for feature selection. For each λ value, a coefficient matrix B on the regression coefficients β_j is obtained, where each column corresponds to a λ value. A series of λ values are used to train the LASSO regression model, and the optimal λ value is selected by k-fold cross-validation. The data set was divided into k subsets and we used $k-1$ of them as the training set and the remaining subsets as the validation set each time, and then we calculated the performance of the model on the validation set and, finally, took the average value as the performance index of the model and selected the λ value with the minimum average MSE value as the optimal λ value. The optimization goal of LASSO regression is to minimize the sum of the absolute values of the regression coefficients. This enables the selection of features, i.e., the coefficients of some unimportant features are reduced or even become zero, thus simplifying the model and improving its generalization ability and interpretation.

2.4.2. Characteristic Wavelength Selection Using Chemical Group Response Spectra

The correlation between SSC and spectral characteristics is essentially caused by the response of chemical groups in SSC to the spectrum [20]. SSC in pear is mainly composed of monosaccharides and polysaccharides such as glucose, fructose, and sucrose, and these compounds absorb or reflect light in a specific wavelength range, producing a specific spectral signature [21,22]. Table 1 shows the wavelength ranges corresponding to the response spectra of the different groups within the SSC of pears and the types of compounds that may be present.

Table 1. Wavelengths corresponding to the response spectra of different compounds of soluble solids of pears.

Perssad	Chemical Compounds	Wavelength Range (cm^{-1})
O-H	saccharides, organic acid, amino acid	3200–3600
C-H	saccharides, organic acid, amino acid, delspray	2800–3000
C=O	saccharides, organic acid	1700–1750
C-O	saccharides, organic acid, amino acid	1000–1300
N-H	amino acid	3300–3500

These chemical groups include O-H, C-H, C=O, C-O, and N-H. The types of compounds in which these groups are present in pear fruit include sugars, organic acids, amino acids, fatty acids, and polyphenols. Correlation analysis of the SSC with the spectra can determine which bands are related to the soluble solids content of pears, and a correlation coefficient plot can be used to obtain the characteristic wavelengths corresponding to the response spectra of the chemical groups of the SSC. The following Equation (2) was used to calculate the correlation coefficient between SSC and spectra [23].

$$R = \frac{\sum_{k=1}^{n}(x_k - \overline{x})(c_k - \overline{c})}{\sqrt{\sum_{k=1}^{n}(x_k - \overline{x})^2 \sum_{k=1}^{n}(c_k - \overline{c})^2}} \quad (2)$$

where x denotes the original spectral wavelength, c denotes the concentration of SSC, and $\overline{x}$ and $\overline{c}$ denote the average values of x and c, respectively.

2.4.3. Feature Wavelength Selection for Group Weighted Fusion Methods

The method of group weighted fusion for selecting features is to weight and fuse a group of characteristic wavelengths selected by a mathematical analysis method (e.g., LASSO, CARS, GA, etc.) with a group of characteristic wavelengths obtained by chemical analysis of groups contained in the compounds of pear soluble solids. The specific steps are as follows: first, a characteristic wavelength group is obtained from the mathematical analysis method; then, another characteristic wavelength group is obtained from the chemical analysis of the moieties contained in the compounds of pear soluble solids; next, these two characteristic wavelength groups are weighted and fused to obtain a new wavelength group; subsequently, the new wavelength group is subjected to a weighting calculation for determining the importance of each wavelength; and finally, a final characteristic wavelength group is obtained by weighting analysis to obtain the final characteristic wavelength group, which is input into the PLS model as the final input variables for modeling and prediction.

2.5. Characterization Factor-Weighting Model for Pear SSC Modified by Contribution

The characterization factor-weighting model for the SSC of pears corrected by contribution is a prediction model based on feature selection and weighting. The core idea of the model is to automatically select the most useful features for the prediction task by calculating the contribution of each feature to the SSC and weighting them for training the prediction model [24]. This can effectively improve the accuracy and stability of the prediction model while also reducing the number of features and the complexity of the model. Assuming that the measured SSC of pears is related to J characterization factors, the SSC value can be expressed as:

$$c = X_j b_j + e_j \quad j = 1, 2, \ldots\ldots, J \quad (3)$$

where c represents the SSC value matrix; X_j represents the characterization factor matrix; b_j represents the correlation coefficient matrix; and e_j represents the residual matrix. Since each characterization factor contributes differently to the value of the SSC of pears, factor weights are introduced for further weighting calculation, and then the SSC value can be expressed as:

$$y = \sum_{j=1}^{J} w_j X_j b_j \quad (4)$$

where W_j is the weight of the jth characterization factor. Each factor weight can be calculated using the least squares method, which minimizes the squared deviation between the theoretical value and the calculated value, and it is expressed as:

$$\begin{gathered} \text{minimize} \left\| c - \sum_{j=1}^{j} w_j x_j b_j \right\|^2 \\ 1 \geq w_j \geq 0 \ (j = 1, 2, \ldots\ldots, j) \end{gathered} \quad (5)$$

When the calculated value of w_j is not 0, it means that the corresponding characterization factor is related to the SSC value of pears, and when the calculated value of w_j is 0, it means that the corresponding characterization factor is not related to the SSC value of pears.

2.6. Modeling Methods and Model Evaluation

2.6.1. PLS-Based Predictive Model Construction Method

Partial least square is a commonly used multivariate statistical method and mathematical optimization technique for building predictive models and solving multiple linear regression problems [25–30]. Compared with traditional multiple linear regression methods, PLS methods can effectively deal with high-dimensional data and multi-collinearity problems and also improve the predictive power of models. A PLS method was used for modeling in this study.

2.6.2. Model Performance Evaluation

The coefficient of determination (R^2) and the root mean square error (RMSE) were used as model evaluation criteria. They are calculated as follows:

$$R^2 = 1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}(y_i - \overline{y}_i)^2 \quad (6)$$

$$RMSE = \sqrt{(1/n) \times \sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (7)$$

where $\hat{y}_i$ is the predicted value of the ith sample, y_i is the measured value of the ith sample, $\overline{y}$ is the mean value of the calibration or prediction set, and n is the number of samples in the data set or the calibration or prediction set. Generally, a good model should have a higher R^2, a lower RMSE (RMSEC or RMSEP), and also a small difference between RMSEC and RMSEP.

3. Results and Discussion

3.1. Sample Division

The calibration and prediction sets were divided using the SPXY (sample set partitioning based on joint x–y distance) algorithm [31]. When calculating the inter-sample distance, the X variable (spectral data) and Y variable (SSC of pear) are taken into account simultaneously to calculate the inter-sample distance to ensure the maximum characterization of the sample distribution. Based on the SPXY method, the inter-sample distance $d_{xy}(i, j)$ can be expressed by Equation (8) as follows:

$$d_{xy}(i,j) = \frac{d_x(i,j)}{\max_{i,j\in(1,z)}[d_x(i,j)]} + \frac{d_y(i,j)}{\max_{i,j\in(1,z)}[d_y(i,j)]}, i,j \in [1,z] \quad (8)$$

where $d_x(i, j)$ and $d_y(i, j)$ denote the distance between each sample with only spectral or mass attributes as the characteristic parameter statistics, respectively. As shown in Table 2, a division ratio of 2:1 was set, 160 samples were used for modeling, and 80 samples were used for prediction.

Table 2. Summary statistics of SSC values of pears in calibration and prediction sets.

Sample Sets	No. of Samples	Mean (%)	Min. (%)	Max. (%)	S.D. (%)
Calibration set	160	10.899	9.411	12.670	0.817
Prediction set	80	10.316	9.549	11.945	0.667
Total sample	240	10.826	9.411	12.670	0.709

Table 2 provides an overview of SSC distributions of pears in the calibration and prediction sets. Statistical values include number of samples, range, mean, and standard deviation (S.D.). As can be seen, the SSC measurements of 240 samples were fairly normally distributed around the mean values (mean = 10.826%), with a standard deviation of 0.709. In this study, samples were divided into calibration and prediction sets. The range of the calibration set was from 9.441% to 12.670% and the range of the prediction set was from

9.549% to 11.945%. The SSC range of the calibration set is larger than that of the prediction set, which is helpful in developing a good model.

3.2. Preprocessing Spectral Data

Figure 1 shows the preprocessed spectral data of some typical samples from the two pear varieties. As shown in Figure 1, the smoothing of spectral data not only removes the noise and unnecessary fluctuations but also retains the trends and features of the original data. It can also be found that the two varieties of samples have similar spectral curve variation trends throughout the entire spectral region, but there is a very significant reflection difference around 680 nm, mainly due to differences in skin reflection of different varieties of pears. Compared to 'Cuiguan' pears, 'Huanghua' pears have a yellower surface color and higher reflection in the spectral range of 600–700 nm. The negative reflection phenomenon of pears in the 700–1100 nm wavelength range may be related to the absorption of components and pigments in pears. For example, some components (such as moisture and sugar content) in the tissue and pigments (such as chlorophyll) in the pear skin can better absorb the light in the visible and short-wave NIR region.

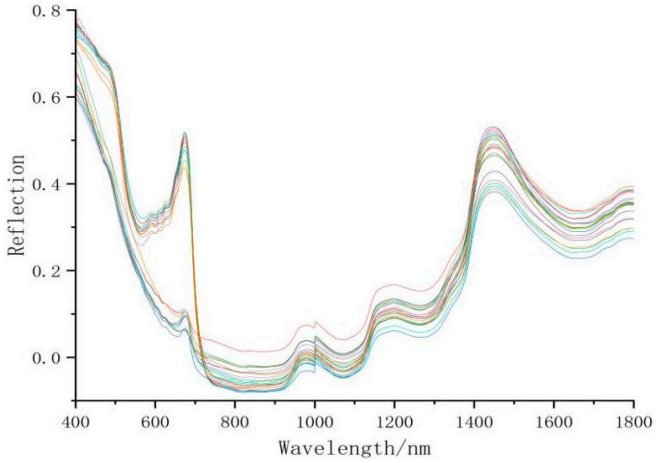

Figure 1. The preprocessed spectral data with SG smoothing.

3.3. Results Obtained by Different Characteristic Wavelength Selection Methods

3.3.1. CARS

The three plots (as shown in Figure 2) created in MATLAB in the process of feature wavelength selection using the CARS algorithm depict the number of variables, the ten-fold cross-validation RMSECV values, and the variation in regression coefficients for each variable. These three plots can be used to analyze the feature wavelength selection process and the results of the CARS algorithm. Figure 2a shows the variation in the number of variables as the number of samples increases. It can be seen that the CARS algorithm has two processes: 'coarse selection' and 'fine selection'. In the early stages of sampling, the number of variables decreases rapidly and then slowly as the number of samples increases. This indicates that the CARS algorithm can quickly filter out the characteristic wavelengths associated with the target variables. Figure 2b shows the change in the ten-fold cross-validated RMSECV values of a single PLS model as the number of samples increases, and it can be seen that the RMSECV values show a trend of decreasing and then increasing, where the subset of variables with the smallest number of samples for the RMSECV values can be identified as the key subset of variables associated with the target variables. This indicates that the CARS algorithm can effectively remove irrelevant information from the target variables and improve the prediction performance of the model. In Figure 2c, the position marked by the vertical line indicates that the subset of variables with the smallest RMSECV value is identified as the key subset of variables associated with the target variable. Before this subset of variables is identified, the CARS

algorithm removes a large amount of irrelevant information by quickly filtering out feature wavelengths that are not relevant to the target variable. Therefore, the RMSECV values show a trend of decreasing and then increasing in sampling operations 1–29, indicating that the CARS algorithm can effectively remove irrelevant information from the target variables and improve the prediction performance of the model. Starting from the 30th sampling operation, the RMSECV value increases, which indicates that some critical information is removed, and the model performance become worse. This may be due to the CARS algorithm being too strict in the selection process and removing some feature wavelengths related to the target variables, which leads to a decrease in the model's performance.

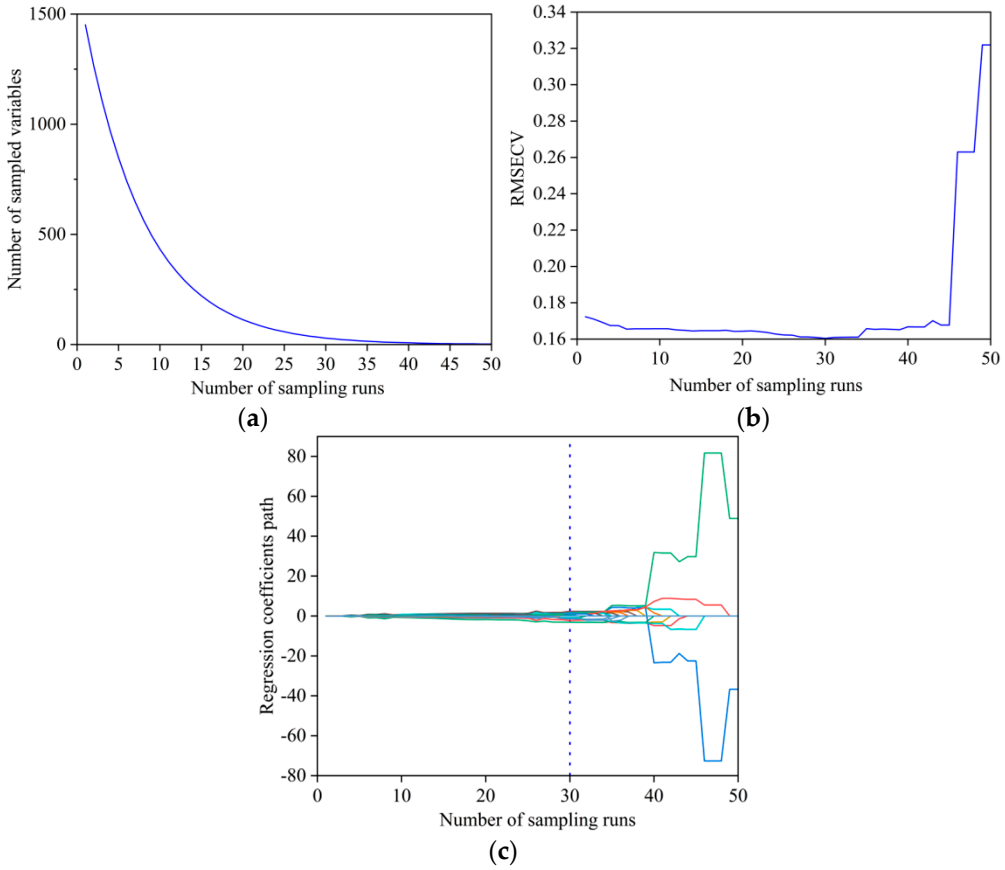

Figure 2. Variable extraction results by CARS. (**a**) Change trend of variables. (**b**) Change trend of RMSECV. (**c**) Path of regression coefficients.

3.3.2. GA

The genetic algorithm was used to determine the number of modeling variables by frequency values with an overall size of 30, a maximum number of iterations of 100, a crossover probability of 50%, and a variance probability of 1%. Figure 3a represents the frequency of the selected wavelengths; the red line in the figure is the automatically generated frequency threshold after the completion of the genetic algorithm, and the green horizontal line shows the cutoff value for the model with the minimum RMSECV. Figure 3b shows the cross-validation (CV)-explained variance as a function of the number of included variables. The CV-explained variance is a metric used to assess the predictive performance of a model and indicates the model's ability to explain the observed data. When the number of variables increases to 29, the red asterisks represent the number of variables corresponding to the F-test results that are significant in the model. When the number of variables increases to 124, the CV-explained variance increases to a maximum value of 96.67% and remains relatively constant or decreases, and the green asterisks indicate the global maximum, i.e., the model with the largest CV-explained variance among all

the numbers of variables, a value that corresponds to the optimal number of variables that provides the best model prediction performance. Figure 3c shows the variation in the cross-validation's root mean square error (RMSECV) with the number of selected feature wavelengths. The RMSECV decreases with an increase in the number of selected feature wavelengths. The red and green asterisks' horizontal coordinate corresponds to Figure 3, where the red asterisk's position has an RMSECV of 0.1663 and the green asterisk's position has an RMSECV of 0.161, which remains constant thereafter, indicating that the number of the selected feature wavelength is more appropriate.

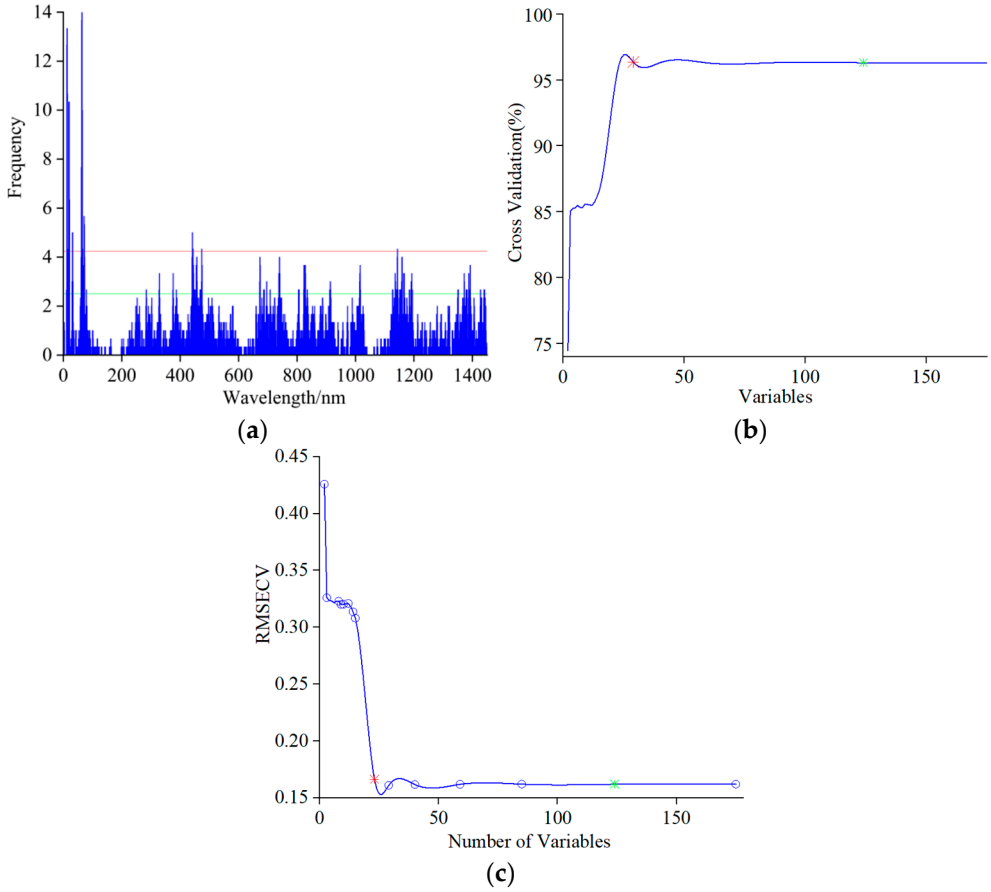

Figure 3. Results of feature wavelength extraction by GA. (**a**) Frequency plot of selected variables (**b**) CV change curve with the number of variables. (**c**) RMSECV change curve with the number of variables. The red line is the frequency threshold automatically generated after the completion of the genetic algorithm, and the green horizontal line indicates the cutoff value for the model with the smallest RMSECV. Red asterisks represent the number of variables corresponding to significant F-test results in the model. The green asterisk indicates the global maximum, i.e., the model with the largest variance explained by the CV among all the number of variables.

3.3.3. LASSO

Figure 4 shows the LASSO regression path diagram, indicating how the different coefficients vary with the regularization parameter λ. In the LASSO regression path diagram, the horizontal axis is log(λ) and the vertical axis is the absolute value of the LASSO regression coefficients. MSE with error bars is a commonly used method for visualizing model performance, which can help us understand more intuitively the predictive performance and stability of different models. The method represents the MSE value of each model and its confidence interval by plotting the error bars so that it can show both the mean error and the error range of the model. The MSE value of each model and its confidence interval are calculated first, using 10-fold cross-validation to calculate 10 MSE values for each model, and then their mean and standard error are calculated as the MSE value and

confidence interval of the model. As shown in the figure, the error bars consist of a central point and two line segments, with the central point representing the average MSE value of the model and the vertical line segments representing the confidence intervals, which extend up and down from the central point to represent the upper and lower limits of the confidence intervals.

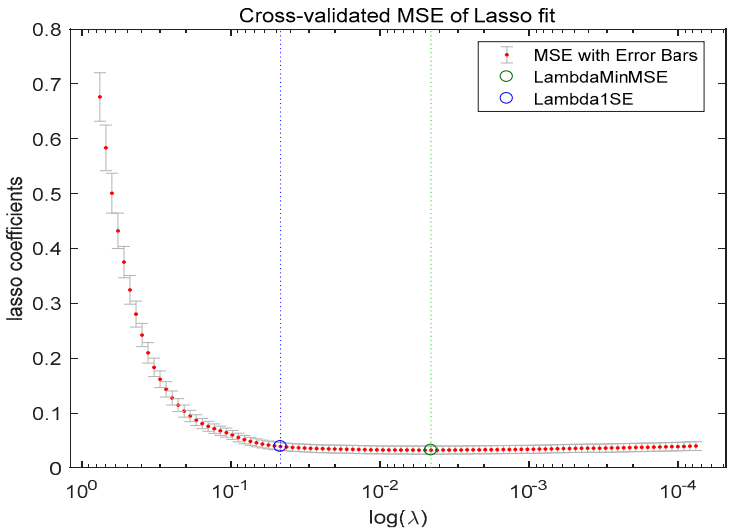

Figure 4. LASSO regression pathway plot.

In this study, both Lambda1 and LambdaMinMSE methods are used to more comprehensively assess the performance and stability of the LASSO model. In this figure, the dashed lines of different colors correspond to different features. The blue dashed line in the figure shows the optimal regularization parameter for the LambdaMinMSE method by calculating the MSE values at different λ values, corresponding to the λ = 0.0046 that minimizes the MSE value. The green dashed line in the figure shows the MSE value and the standard error of the MSE value by calculating the MSE value at different λ values by the Lambda1SE method and finding the λ = 0.0467 that makes the MSE value less than the minimum MSE value plus the standard error as the optimal regularization parameter. The cross-validation error curve is used to determine the optimal λ value, and the optimal λ value is the path diagram that minimizes the cross-validation error for the LambdaMinMSE method, corresponding to λ = 0.0046. The features with non-zero coefficients in matrix B at a value of λ = 0.0046 are selected as the final set of features. The final LASSO regression screening yielded the following feature wavelengths: 350, 368, 392, 412, 413, 522, etc., for a total of 32 feature wavelengths.

3.3.4. Result and Analysis of Different Models

The choice of modeling method has a great impact on the performance and stability of the model. As shown in Table 3, it can be seen that the predictive performance and stability of the CARS–PLS, GA-PLS, and LASSO-PLS models are better than those of the raw-data PLS model. This indicates that the appropriate modeling method can improve the predictive ability and stability of the models. However, these characteristic wavelength selection methods only optimize the models by means of mathematical analysis. Neglecting the characteristic absorption peaks of different groups of compounds in pear soluble solids for different wavelength spectra may result in unfavorable model enhancement. A total of 33 characteristic wavelengths were selected by the LASSO–PLS method, which had the best predictive performance and stability compared with CARS–PLS and GA–PLS, with an r^2_{pre} of 0.9754 and RMSEP of 0.135. Therefore, in order to further improve the modeling effect, the most effective mathematical analysis method (LASSO) was chosen to select the characteristic wavelengths, which were weighted and fused with the characteristic wavelengths corresponding to the response spectra of the chemical groups in the pear

SSC selected through the correlation analysis of the pear SSC with the spectra to build the LASSO–GWF–PLS model.

Table 3. Results obtained from different models with different wavelength number.

Models	Wavelength Number	Calibration Set		Prediction Set	
		r^2_{cal}	*RMSEC*	r^2_{pre}	*RMSEP*
PLS	1451	0.9491	0.184	0.9321	0.154
CARS–PLS	29	0.9751	0.173	0.9736	0.131
GA–PLS	19	0.9662	0.162	0.9637	0.151
LASSO–PLS	33	0.9762	0.178	0.9754	0.135

3.4. Correlation Analysis between Near-Infrared Spectral Characteristics and SSC of Pears

The method of picking the characteristic wavelengths of pear SSC chemical group response spectra uses spectroscopic techniques to analyze the chemical composition, molecular structure, and other information in pear samples to determine the characteristic wavelength related to the SSC and improve the accuracy and stability of the prediction model. Firstly, the information was checked to determine all groups of compounds in pear soluble solids and the absorption peaks of these groups and, secondly, the correlation coefficients of each characteristic wavelength with pear soluble solids were calculated, and the characteristic wavelengths were selected by the peaks and valleys of the correlation coefficient plot. Figure 5 shows the correlation coefficient plot between pear soluble solids and spectral data, which reflects the correlation between the spectral response at different wavelengths and the content of pear soluble solids. The spectral responses at 465 nm and 522 nm correspond to the π-π leap absorption peaks of aromatic compounds. The spectral response at 980 nm corresponds to the vibrational absorption peak of the hydroxyl functional group in the pear, because the hydroxyl functional group can vibrate in this wavelength range to produce an absorption peak. The spectral responses at 657 nm and 695 nm correspond to the absorption peaks of carotenoids and chlorophylls in pears because carotenoids and chlorophylls absorb light energy in this wavelength range, resulting in absorption peaks. The spectral responses at 1230 nm and 1260 nm correspond to C-H stretching vibration absorption peaks in pears, because the C-H bonds in fatty acids and lipids undergo stretching vibrations in this wavelength range, resulting in absorption peaks. The spectral response at 1460 nm corresponds to the C-H bending vibration absorption peak in pears, because the C-H bonds in proteins and amino acids undergo bending vibrations in this wavelength range, resulting in absorption peaks. The spectral response at 1729 nm corresponds to the C=O stretching vibration absorption peak in pears, because the C=O bonds in esters and ketones undergo stretching vibrations in this wavelength range, resulting in absorption peaks.

From the response relationship between pear quality attributes and spectra, it can be determined that there is a high correlation between SSC and some group spectra. LASSO selected characteristic wavelength set S1, pear SSC chemical group response spectra corresponding to wavelength set S2, and LASSO-GWF characteristic wavelength set S, as shown in Table 4.

Table 4. Different feature wavelength sets.

Wavelength Sets	Wavelength Number	Feature Wavelength
S1	33	350, 368, 392, 412, 413, 522, 523, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 678, 682, 735, 736, 743, 744, 1460, 1461, 1473, 1474, 1475, 1482, 1729, 1740
S2	9	465, 522, 657, 695, 980, 1230, 1260, 1460, 1729
S	38	350, 368, 392, 412, 413, 465, 522, 523, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 678, 682, 695, 735, 736, 743, 744, 980, 1230, 1260, 1460, 1461, 1473, 1474, 1475, 1482, 1729, 1740

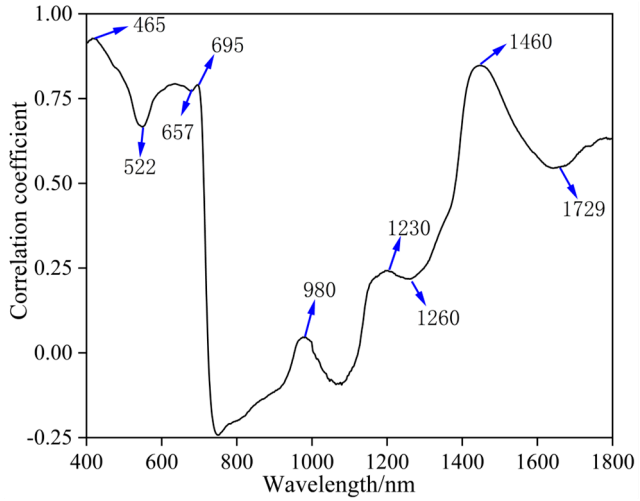

Figure 5. Correlation coefficient between pear soluble solids and spectral data.

3.5. Analysis of the Results of the Weighted Model

In this study, nine response spectra with high correlation with pear soluble solids were selected, and four of these wavelengths, namely 522 nm, 695 nm, 1460 nm, and 1729 nm, were consistent with the wavelengths selected by the LASSO algorithm, so the remaining five highly correlated wavelengths were fused with the thirty-three wavelengths selected by the LASSO algorithm to construct a wavelength set (containing thirty-eight wavelengths in total). Subsequently, each wavelength was considered an independent characterization factor of pear soluble solids, and the results were analyzed by the weighted model as shown in Figure 6 (Note: weighting value is zero at a wavelength of 648 nm, so only 37 wavelengths are shown in Figure 6). A total of 37 characterization factors were correlated with pear soluble solids, and their weight values w_j were in the range of 0.181–0.948.

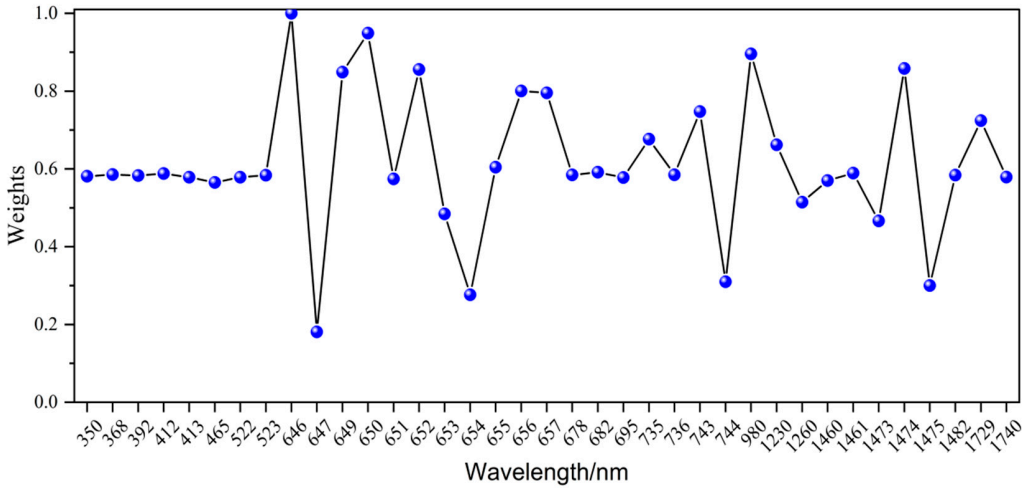

Figure 6. Weighting values of different wavelength characterization factors.

3.6. LASSO–GWF–PLS Model Analysis

The LASSO algorithm selects the feature wavelengths with predictive power from a large number of wavelengths to obtain a feature subset S1. The feature wavelengths corresponding to the response spectra of pear soluble groups are obtained by correlation analysis between pear soluble solids and spectra to obtain a feature subset S2. The obtained S1 and S2 are combined to obtain a feature set S that contains all feature wavelengths. For each feature wavelength λ_i, its weight w_i is calculated, and the weight w_i of each feature

wavelength is used as its weight in the model for the weighted fusion of features. The set of feature wavelengths after weighted fusion is input into the PLS model to establish the prediction model between the feature variables and the target variables. In this study, the most relevant feature wavelengths were selected by LASSO regression, combined with the soluble solids group response spectra of pears, and PLS was used to establish the relationship model between the response variable (i.e., soluble solids content of pears) and the predictor variable (i.e., the feature wavelengths selected by LASSO regression and the wavelengths corresponding to group response spectra). The most relevant features selected by LASSO regression were used, and the effect of the wavelength corresponding to the group response spectra was also taken into account to establish a more accurate prediction model.

The results of the LASSO–GWF–PLS model are shown in Figure 7. The quantitative analysis model has a good prediction effect, and the coefficients of determination of the calibration set (a) and the prediction set (b) are 0.992 and 0.998, respectively. The model is robust and adaptable for detecting SSC in pears. It can be also found that the R^2 of the prediction set improved from 0.975 to 0.998 compared with the LASSO–PLS prediction model, indicating that the LASSO–GWF–PLS method has better predictive ability.

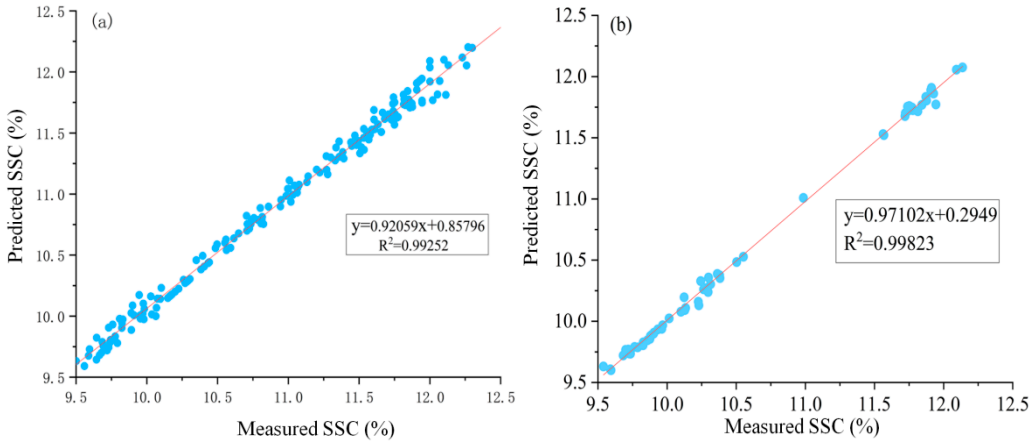

Figure 7. Measured vs. predicted values for SSC by LASSO–GWF–PLS model. (**a**) Correction set (**b**) Prediction set.

4. Conclusions

In this study, the LASSO–PLS method with the best prediction effect was developed to realize the fast and accurate assessment of SSC in pears. Compared with a PLS model with the original data, r^2_{cal} was improved from 0.949 to 0.976, and r^2_{pre} was improved from 0.932 to 0.975. In order to further improve the model's effect, the characteristic wavelength groups selected by the LASSO algorithm were weighted and fused with the pear soluble solids group response spectra and, finally, the LASSO–GWF–PLS model was proposed. Compared with the LASSO–PLS method employing mathematical analysis only, the r^2_{cal} of the LASSO–GWF–PLS model was improved from 0.949 to 0.992, and r^2_{pre} was improved from 0.932 to 0.998. This indicated that the fusion of the spectra of the group responses of the compounds in pear soluble solids had an enhancing effect on the model. The performance of the prediction model was optimized by selecting an appropriate mathematical analysis of the characteristic wavelength selection method fused with the group response spectra. The method can be applied for the other fruits' quality analysis, such as sugar content, acidity, etc., which provides a new idea for the quality analysis of fruits and other food products.

Author Contributions: B.Z.: methodology; P.L.: software; P.L. and M.L.: original draft preparation; W.L. and H.Z.: review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the National Natural Science Foundation of China (62265007 and 32260622) and supported by the Jiangxi Provincial Natural Science Foundation (20224BAB212007).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Xia, Y.; Fan, S.; Tian, X.; Huang, W.; Li, J. Multi-factor fusion models for soluble solid content detection in pear (*Pyrus bretschneideri* 'ya') using Vis/NIR online half-transmittance technique. *Infrared Phys. Technol.* **2020**, *110*, 103443. [CrossRef]
2. Deng, J.; Jiang, H.; Chen, Q. Characteristic wavelengths optimization improved the predictive performance of near-infrared spectroscopy models for determination of aflatoxin B-1 in maize. *J. Cereal Sci.* **2022**, *105*, 103474. [CrossRef]
3. Jiang, H.; Wang, J.; Chen, Q. Comparison of wavelength selected methods for improving of prediction performance of PLS model to determine aflatoxin B1 (AFB1) in wheat samples during storage. *Microchem. J.* **2021**, *170*, 106642. [CrossRef]
4. Wang, T.; Li, G.; Dai, C. Soluble Solids Content prediction for Korla fragrant pears using hyperspectral imaging and GsMIA. *Infrared Phys. Technol.* **2022**, *123*, 104119. [CrossRef]
5. Xin, Z.H.; Ju, S.C.; Zhang, D.Y.; Zhou, X.G.; Guo, S.; Pan, Z.G.; Wang, L.S.; Cheng, T. Construction of spectral detection models to evaluate soluble solids content and acidity in Dangshan pear using two different sensors. *Infrared Phys. Technol.* **2023**, *131*, 104632. [CrossRef]
6. Martins, J.A.; Rodrigues, D.; Cavaco, A.M.; Antunes, M.D.; Guerra, R. Estimation of soluble solids content and fruit temperature in 'Rocha' pear using Vis-NIR spectroscopy and the SpectraNet-32 deep learning architecture. *Postharvest Biol. Technol.* **2023**, *199*, 112281. [CrossRef]
7. Chen, S.B.; Yang, H.; Luo, R.; Hu, Z. Rapid Quantitative Model and Optimization of Potato Soluble Solids by Near Infrared Spectroscopy. *Anhui Agric. Sci.* **2021**, *49*, 205–209.
8. Guo, Y.; Guo, J.; Shi, Y.; Li, X.; Liu, Y.; Huang, H.; Li, Z. Prediction of soluble solids in Hami melon by CARS-SVM. *Food Mach.* **2021**, *37*, 81–85.
9. Liu, Y.; Chen, X.; Ouyang, A. Non-Destructive Measurement of Soluble Solid Content in Gannan Navel Oranges by Visible/Near-Infrared Spectroscopy. *Acta Opt. Sin.* **2008**, *28*, 478–481.
10. Zheng, K.; Li, Q.; Wang, J. Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra. *Chemom. Intell. Lab. Syst.* **2012**, *112*, 48–54. [CrossRef]
11. Yang, Y.; Zhao, C.; Huang, W.; Tian, X.; Fan, S.; Wang, Q. Optimization and compensation of models on tomato soluble solids content assessment with online Vis/NIRS diffuse transmission system. *Infrared Phys. Technol.* **2022**, *121*, 104050. [CrossRef]
12. Zheng, K.; Feng, T.; Zhang, W. Variable selection by double competitive adaptive reweighted sampling for calibration transfer of near infrared spectra. *Chemom. Intell. Lab. Syst.* **2019**, *191*, 109–117. [CrossRef]
13. Liu, J.; Zeng, C.; Wang, N.; Shi, J.; Sun, Y. Rapid biochemical methane potential evaluation of anaerobic co-digestion feedstocks based on near infrared spectroscopy and chemometrics. *Energies* **2021**, *14*, 1460. [CrossRef]
14. Li, W.; Suhayb, M.K.; Thangavelu, L. Implementation of AdaBoost and genetic algorithm machine learning models in prediction of adsorption capacity of nanocomposite materials. *J. Mol. Liq.* **2022**, *350*, 118527. [CrossRef]
15. Yao, J.; Wu, Z.; Liu, Y. Predicting membrane fouling in a high solid AnMBR treating OFMSW leachate through a genetic algorithm and the optimization of a BP neural network model. *J. Environ. Manag.* **2022**, *307*, 114585. [CrossRef] [PubMed]
16. Hong, Y.Y.; Chan, Y.H.; Cheng, Y.H. Week-ahead daily peak load forecasting using genetic algorithm-based hybrid convolutional neural network. *IET Gener. Transm. Distrib.* **2022**, *12*, 2416–2424. [CrossRef]
17. Yoon, D.; Kim, K.; Cha, D.-H. Development of model output statistics based on the least absolute shrinkage and selection operator regression for forecasting next-day maximum temperature in South Korea. *Q. J. R. Meteorol. Soc.* **2022**, *148*, 1929–1944. [CrossRef]
18. Hu, X.; Shen, F.; Zhao, Z.; Qu, X.; Ye, J. An individualized gait pattern prediction model based on the least absolute shrinkage and selection operator regression. *J. Biomech.* **2020**, *112*, 110052. [CrossRef]
19. Narala, S.; Li, S.; Klimas, N.K.; Patel, A.B. Application of least absolute shrinkage and selection operator logistic regression for the histopathological comparison of chondrodermatitis nodularis helicis and hyperplastic actinic keratosis. *J. Cutan. Pathol.* **2021**, *48*, 739–744. [CrossRef]
20. Chu, X.L. *Chemometric Methods in Modern Spectral Analysis*; Chemical Industry Press: Beijing, China, 2022.
21. Yu, Y.; Zhang, Q.; Huang, J. Nondestructive determination of SSC in Korla Fragrant Pear using a portable near-infrared spectroscopy system. *Infrared Phys. Technol.* **2021**, *116*, 103785. [CrossRef]
22. Cruz, S.; Guerra, R.; Brazio, A. Nondestructive simultaneous prediction of internal browning disorder and quality attributes in 'Rocha' pear (*Pyrus communis* L.) using VIS-NIR spectroscopy. *Postharvest Biol. Technol.* **2021**, *179*, 111562. [CrossRef]
23. Zaveri, S.T. Hyperspectral endmember extraction using Pearson's correlation coefficient. *Int. J. Comput. Sci. Eng.* **2021**, *24*, 89–97.
24. Lv, Y.; Yang, H. A multi-model modeling approach based on weighted kernel Fisher criterion feature extraction. *Chin. J. Chem. Eng.* **2014**, *22*, 22–28.

25. Asri, M.N.M.; Verma, R.; Mahat, N.A.; Nor, N.A.M.; Desa, W.N.S.M.; Ismail, D. Raman spectroscopy with self-organizing feature maps and partial least squares discriminant analysis for discrimination and source correspondence of red gel ink pens. *Microchem. J.* **2022**, *175*, 107170. [CrossRef]
26. Wang, H.; Chu, X.; Chen, P.; Li, J.; Liu, D.; Xu, Y. Partial least squares regression residual extreme learning machine (PLSRR-ELM) calibration algorithm applied in fast determination of gasoline octane number with near-infrared spectroscopy. *Fuel* **2022**, *309*, 122224. [CrossRef]
27. Xie, Z.; Feng, X.; Chen, X. Subsampling for partial least-squares regression via an influence function. *Knowl.-Based Syst.* **2022**, *245*, 108661. [CrossRef]
28. Li, Z.; Pang, W.; Liang, H.; Chen, G.; Duan, H.; Jiang, C. Fast Quantitative Modelling Method for Infrared Spectrum Gas Logging Based on Adaptive Step Sliding Partial Least Squares. *Energies* **2022**, *15*, 1325. [CrossRef]
29. Deng, L.; Ma, L.; Cheng, K.K.; Xu, X.; Raftery, D.; Dong, J. Sparse PLS-Based Method for Overlapping Metabolite Set Enrichment Analysis. *J. Proteome Res.* **2021**, *20*, 3204–3213. [CrossRef]
30. Li, J.; Tian, X.; Huang, W.; Zhang, B.; Fan, S. Application of Long-Wave Near Infrared Hyperspectral Imaging for Measurement of Soluble Solid Content (SSC) in Pear. *Food Anal. Methods* **2016**, *9*, 3087–3098. [CrossRef]
31. Wang, S.; Han, P.; Cui, G. The NIR Detection Research of Soluble Solid Content in Watermelon Based on SPXY Algorithm. *Spectrosc. Spectr. Anal.* **2022**, *39*, 738–742.