agriculture

MDPI

Article

Nondestructive Testing Model of Tea Polyphenols Based on Hyperspectral Technology Combined with Chemometric Methods

Xiong Luo, Lijia Xu, Peng Huang, Yuchao Wang, Jiang Liu, Yan Hu, Peng Wang and Zhiliang Kang *

College of Mechanical and Electrical Engineering, Sichuan Agricultural University, Ya'an 625000, China; 2019217012@stu.sicau.edu.cn (X.L.); 10887@sicau.edu.cn (L.X.); 14130@sicau.edu.cn (P.H.); 14225@sicau.edu.cn (Y.W.); 2019317021@stu.sicau.edu.cn (J.L.); 2020317020@stu.sicau.edu.cn (Y.H.); 2019317016@stu.sicau.edu.cn (P.W.)

* Correspondence: 12200@sicau.edu.cn; Tel.: +86-186-0835-1703

Abstract: Nondestructive detection of tea's internal quality is of great significance for the processing and storage of tea. In this study, hyperspectral imaging technology is adopted to quantitatively detect the content of tea polyphenols in Tibetan teas by analyzing the features of the tea spectrum in the wavelength ranging from 420 to 1010 nm. The samples are divided with joint x-y distances (SPXY) and Kennard-Stone (KS) algorithms, while six algorithms are used to preprocess the spectral data. Six other algorithms, Random Forest (RF), Gradient Boosting (GB), Adaptive boost (AdaBoost), Categorical Boosting (CatBoost), LightGBM, and XGBoost, are used to carry out feature extractions. Then based on a stacking combination strategy, a new two-layer combination prediction model is constructed, which is used to compare with the four individual regressor prediction models: RF Regressor (RFR), CatBoost Regressor (CatBoostR), LightGBM Regressor (LightGBMR) and XGBoost Regressor (XGBoostR). The experimental results show that the newly-built Stacking model predicts more accurately than the individual regressor prediction models. The coefficients of determination R_c^2 and R_p^2 for the prediction of Tibetan tea polyphenols are 0.9709 and 0.9625, and the root mean square error RMSEC and RMSEP are 0.2766 and 0.3852 for the new model, respectively, which shows that the content of Tibetan tea polyphenols can be determined with precision.

Keywords: hyperspectral; tea polyphenols; sample division; feature selection; regression model; nondestructive detection

Citation: Luo, X.; Xu, L.; Huang, P.; Wang, Y.; Liu, J.; Hu, Y.; Wang, P.; Kang, Z. Nondestructive Testing Model of Tea Polyphenols Based on Hyperspectral Technology Combined with Chemometric Methods. *Agriculture* **2021**, *11*, 673. https://doi.org/10.3390/agriculture11070673

Academic Editor: Fabio Sciubba

Received: 4 July 2021
Accepted: 13 July 2021
Published: 16 July 2021

1. Introduction

Tea is one of the three most popular non-alcoholic beverages in the world. Tea polyphenols are an important part of tea and a vital source of bioactive chemicals, with the ability of anti-oxidation, anti-cancer, anti-bacterial, anti-inflammation and anti-arteriosclerosis [1–3], and they play an important role in the medical and food industries. In addition, there is also a certain correlation between the content of tea polyphenols and the quality of tea [4]. Not only beneficial to human health, high-quality tea is also sold at a much higher price in the market. The traditional detection methods of tea polyphenols are mainly either physical or chemical [5–7], which are not only costly and complicated but also time-consuming and vulnerable to subjective factors [8]. Therefore, it is of great significance to develop a fast and nondestructive online detection technology to test tea polyphenols. Hyperspectral imaging technology is based on a lot of narrow-band image data technology. It combines imaging technology with spectral technology to detect the two-dimensional geometric space and one-dimensional spectral information of the target and obtain continuous and narrow-band data with high spectral resolution. Hyperspectral imaging is a new generation of photoelectric detection technology and can be adopted in this field for its low cost, fast speed, reliability and its ability to leave the samples intact in the test.

Near-infrared spectroscopy technology (NIR) is an optical detection method known for its fast speed and no direct touch of the samples [9]. It has been used in quality inspections of many agricultural products [10–12]. Wang et al. [13] established a pear juiciness detection model at 650–1100 nm, with an external verification determination coefficient of 0.93 and root mean square error of 0.97%. Pennisi et al. [14] established freshness models of different species of fish based on near-infrared spectroscopy technology. Jens et al. [15] designed a potato dry matter content detection model based on NIR technology. Previous studies have shown that the use of spectroscopy technology to detect the quality of agricultural products is feasible.

With the development of spectral technology, image analysis is added to spectroscopy [16], and hyperspectral imaging technology emerges with time. Compared with multispectral images, hyperspectral images have a richer image and spectral information [17]. At present, the use of hyperspectral technology to detect agricultural product quality is still in its infancy. However, as a fast and nondestructive detection technology, hyperspectral imaging has great application prospects. There has been only a small amount of research on agricultural-product quality detection based on hyperspectral technology [18,19].

The modeling results established based on hyperspectral technology are affected by many factors. The method of feature data preprocessing is a major factor affecting the results. Common spectral data preprocessing methods include orthogonal signal correction (OSC) [20], first derivative (FD) [21,22], second derivative (SD) [22], multivariate scattering correction (MSC) [21–23], standard normal variable transformation (SNVT) [21–23], Savitzky–Golay filter (SG) [21,24]. It was shown that these methods could reduce the influence of external factors and improve detection accuracy to some extent.

The selection of spectral characteristic bands is another important factor affecting the model results. Effective selection of characteristic bands can save computing resources [25] and improve model performance. In recent years, researchers have proposed many characteristic band selection methods, such as interval partial least squares (iPLS) [26,27], synergy interval partial least square (siPLS) [28,29], backward interval partial least square (biPLS) [30–32]. These feature-selection algorithms divide all features into several intervals and then select a small part of the interval with good effect as the characteristic band by iteration. However, the spectral features selected by this "bundling" method are likely to miss some important features.

To avoid the presence of bias introduced by manual data splitting, there are a number of computational methods that can be used for sample selection, such as random selection (RS), Kennard-Stone (KS) [33,34], or sample set partitioning based on joint x-y distances (SPXY) [35–37] algorithm.

The purpose of this research is to explore the feasibility of fast and nondestructive online detection of Tibetan tea polyphenol content based on hyperspectral image technology. Different data preprocessing methods are used to process the acquired hyperspectral data of Tibetan tea. This paper selects the best preprocessing method by establishing the model and analyzing the modeling results.

2. Materials and Methods

2.1. Samples

A total of three grades of Ya'an Tibetan tea were selected for the test, including 32 samples for the first grade, 33 for the second grade and 37 for the third grade. Each group of samples was individually packaged in a sealed plastic bag and stored in a 5 °C thermostat for the subsequent determination of spectral data and tea polyphenol physicochemical data. The measurement process of tea polyphenol content is as follows.

1. Reagent preparation.
    (a) Mother liquor: The milled tea (0.6 g) and 5 mL 70% methanol solution were placed in a 10 mL centrifuge tube and shaken. After bathing at 70 °C for 10 min, the tube was removed, allowed to cool, and then centrifuged for

10 min at 3500 r/min, and the supernatant was collected. The precipitation was extracted according to the above extraction procedure. The collected supernatants from the above extraction were mixed, then diluted to 10 mL with 70% aqueous methanol, and filtered through a 0.45 μm filter.

(b) Test solution: 1 mL mother solution (a) was added into a 100 mL volumetric flask, and distilled water was added to dilute to 100 mL and shaken well.

(c) Gallic acid working solution: 1.0, 2.0, 3.0, 4.0 and 5.0 mL of gallic acid standard solution (1000 μg /mL) was added into five 100 mL volumetric flasks, diluted with distilled water to 100 mL, and shaken well. Finally, five groups of working fluid were obtained. The concentrations were 10, 20, 30, 40 and 50 μg/mL.

2. Determination of the content of tea polyphenols.

A total of 1.0 mL each of gallic acid working solution (c), distilled water and test solution (b) were added into the scale tube. A total of 5.0 mL of Folinol reagent (concentration 10%) was added to each test tube. After 4 min, 4.0 mL 7.5 % sodium carbonate (Na_2CO_3) solution was added, then we added water to a constant volume scale. The mixture was then stored at room temperature for 60 min. The absorbance (A, A_0) was measured by a spectrophotometer at the wavelength of 765 nm with a 10 mm colorimetric vessel. The standard curve was prepared according to the absorbance of the gallic acid working solution and the concentration of gallic acid in each working solution. By comparing the absorbance of the sample and the standard working solution, the content of tea polyphenols was calculated as follows:

$$c = \frac{(A - A_0) \times V \times d \times 100}{S_{std} \times \omega \times 10^6 \times m} \quad (1)$$

c (%) is the content of tea polyphenol (percentage of tea polyphenols in dry matter of tea), A represents the absorbance of the sample test solution, A_0 is the absorbance of the blank reagent solution, V (mL) is the volume of sample extract, d is the dilution factor (take 100 here), S_{std} represents the slope of the gallic acid standard curve, ω is the dry matter content of the sample (percentage of tea sample quality before and after drying) and m (g) is the mass of the sample.

The measurement results and the sample results based on the SPXY algorithm (see Section 3.1) are shown in Table 1. The Tibetan tea polyphenol data from the test is used as standard data for future use.

Table 1. Tea polyphenol content statistics and sample partition results based on the SPXY algorithm.

Sample	Number of Samples	Tea Polyphenol Content (%)			
		Minimum	Maximum	Mean	Standard Deviation
Total	102	4.0590	9.4098	7.1225	1.6827
Calibration set	76	4.0590	9.4098	7.2576	1.6010
Prediction set	26	4.5110	9.4079	6.7276	1.8792

Note: "%" is the percentage of tea polyphenols in the dry matter of tea, the same as line 120.

2.2. Hyperspectral Image Acquisition

The hyperspectral data of the Tibetan tea test is acquired using a GaiaSorter hyperspectral sorter made by Beijing Zolix Company, which provides an effective spectral band of 387–1035 nm, a spectral resolution of 2.8 nm and 256 spectral channels. We spread the tea leaves evenly into a square in a container (about 65 cm × 65 cm). The hyperspectral acquisition system is shown in Figure 1. Due to the influence of dark currents at the beginning and end of the spectral band, only the 420–1010 nm band is retained as raw spectral data. The sample platform is set to move at a speed of 4.0 mm · s^{-1}, the distance of the imaging object is 170 mm, and the camera exposure time was set to 16 ms. We placed the tea to be tested on the stage. Under the illumination of a uniform light source, the

platform is moved horizontally at a set speed, and the hyperspectral camera can obtain continuous hyperspectral images of the samples on the platform. The acquired images are then calibrated using Equation (2):

$$I = \frac{I_{raw} - I_b}{I_w - I_b} \quad (2)$$

where I is the corrected image, I_{raw} represents the raw image, I_b is the standard black image and I_w represents the standard white image.

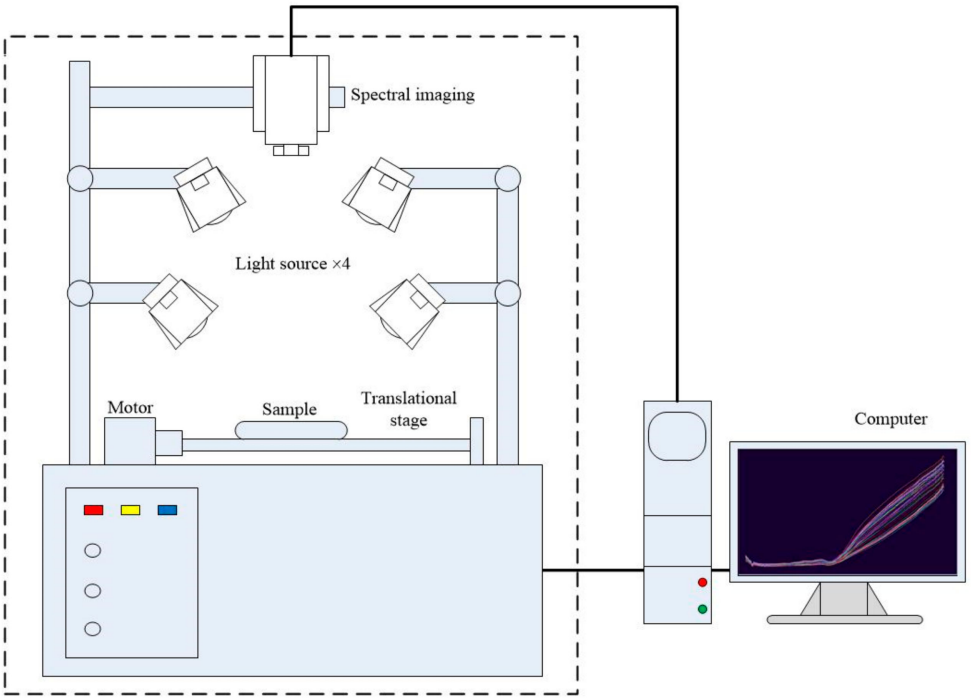

Figure 1. Schematic diagram of the hyperspectral imaging system.

ENVI5.1 software is used to calculate the average spectral value of the region of interest (151 × 151 pixels) in the hyperspectral image.

2.3. Hyperspectral Data Preprocessing

Random noise is often generated during the acquisition of spectra by the external environment, instrument response and other factors unrelated to the nature of the sample to be measured, and disorderly fluctuations in the spectral data appear. Therefore, this article uses six preprocessing algorithms, including SG, MSC, SNVT, FD, SD and Z-score standardization (ZSS), to eliminate the noise in the raw spectrum (RAW) data. Python 3.8 (Python Software Foundation) is adopted in all data processing and modeling.

2.4. Sample Partitioning

2.4.1. Kennard-Stone (KS)

The KS algorithm [33] regards all samples as candidate samples of a training set and selects the two samples with the farthest Euclidean distance into the training set. Then, by calculating the Euclidean distance between the remaining samples and the known samples in the training set, the two samples nearest to the selected samples are selected and put

into the training set, and the above steps are repeated until the number of samples reaches the set value. The formula for calculating Euclidean distance is:

$$d_x(p,q) = \sqrt{\sum_{i=1}^{n}[x_p(i) - x_q(i)]^2}; p,q \in [1,n] \quad (3)$$

where x_p and x_q represent two different samples and represent the number of spectral bands.

2.4.2. Sample Set Partitioning Based on Joint X-Y Distances (SPXY)

The SPXY algorithm is developed on the basis of the KS algorithm. When SPXY calculates the sample distance, the sample label (Y) and the sample feature (X) are taken into account at the same time. The specific calculation is as follows [36]:

$$d_y(p,q) = \sqrt{(y_p - y_q)^2} = |y_p - y_q|; p,q \in [1,n] \quad (4)$$

$$d_{x,y}(p,q) = \frac{d_x(p,q)}{max_{p,q}d_x(p,q)} + \frac{d_y(p,q)}{max_{p,q}d_y(p,q)}; p,q \in [1,n] \quad (5)$$

where $d_x(p,q)$ represents the spectral distance and $d_y(p,q)$ represents the chemical measurement value distance.

2.5. Feature Selection and Modeling

The acquired hyperspectral data often contains a lot of redundant information, which will have a certain impact on the accuracy and efficiency of the final modeling. Six methods [38–41], Gradient Boosting (GB), Adaptive Boosting (AdaBoost), Random Forest (RF), Categorical Boosting (CatBoost), LightGBM and XGBoost, are used to select hyperspectral feature bands. Random forest regression (RFR), categorical boosting regression (CatBoostR), LightGBM regression (LightGBMR), XGBoost regression (XGBoostR) and model integration strategy stacking are used in the model. Stacking is a combined model that trains the base learner from the initial data set and then uses the predicted value of the base-learner as a new feature to train the meta-learner.

2.6. Model Reliability

Model evaluation takes the coefficient of determination (R^2) [42] and root mean square error (RMSE) [43] as evaluation criteria, and the calculation method is shown in Equations (6) and (7).

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

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

where y_i and $\hat{y}_i$ are the measured value and predicted value of the sample, respectively, $\overline{y}$ represents the average value of the sample and n is the number of samples. When the predicted value ($\hat{y}_i$) of the model is closer to the true value (y_i), the better the effect, in other words, a good model should have small RMSE values (the closer the value of RMSE is to 0, the better the effect of the model). Furthermore, the models with high R^2 values are better than the models with low R^2 values (the closer the value of R^2 is to 1, the better the effect of the model). At the same time, the smaller the difference in the determination coefficient between the calibration set and the independent test set of the model, the better. If the gap is too large, it indicates that the model is under-fitting or over-fitting.

3. Results

3.1. Spectral Preprocessing and Sample Division

In the process of collecting hyperspectral data, due to the influence of environmental factors, the acquired spectral data has certain noises, which will adversely affect the performance of the model. Therefore, the spectral data is preprocessed before modeling. Six methods, including SG, MSC, SNVT, ZSS, FD and SD, are used to preprocess the spectral data of the tea samples. In order to make the established model representative, the division of the data set is also very important. This paper uses the KS and SPXY sample division algorithm to divide the 102 groups of samples into the calibration set and the prediction set at a ratio of 3:1.

Gradient Boosting regression (GBR) is used to model and predict the raw data and preprocessed spectral data. The modeling results based on different preprocessing algorithms and different sample partitioning algorithms are shown in Figure 2.

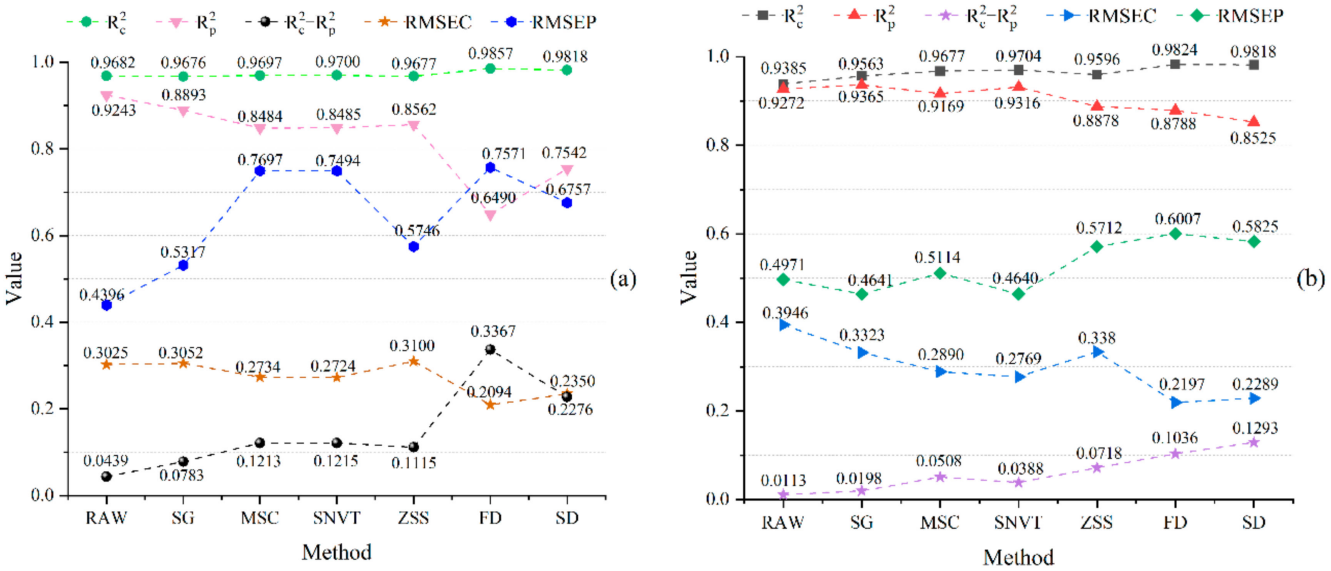

Figure 2. Prediction results of GBR model with different inputs. (**a**) Modeling results based on KS partition data set; (**b**) Modeling results based on SPXY partition data set.

Figure 2a demonstrates the modeling results of the data set divided by the KS algorithm. The coefficients of determination of the calibration set (R_c^2) are all above 0.96. The RAW-KS-GBR model has the best effect, with coefficients of determination of the calibration set at 0.9682, the calibration set root mean square error (RMSEC) at 0.3025, the prediction set determination coefficient (R_p^2) at 0.9243 and the prediction set root mean square error (RMSEP) at 0.4396. In the FD-KS-GBR model R_c^2 is the largest, being 0.9857, but R_p^2 is the smallest, only 0.6490, indicating that the FD-KS-GBR model has a serious overfitting problem. Figure 2b is the modeling result of the data set divided by the SPXY algorithm. The determination coefficient R_c^2 of the model calibration set established by FD and SD preprocessing spectral data is above 0.98, but the values ofR_p^2 do not exceed 0.88. The value of SNVT-SPXY-GBR model R_c^2 is 0.9704, RMSEC is 0.2769, R_p^2 is 0.9316 and RMSEP is 0.4640. The value of SG-SPXY-GBR model R_c^2 is 0.9563, RMSEC is 0.3323, R_p^2 is 0.9365 and RMSEP is 0.4641.

The data set divided by the KS algorithm is more prone to overfitting than the model established by the SPXY algorithm, so the SPXY-GBR model is generally better than the KS-GBR model. Based on Figure 2, comparing two different data set partitioning methods, and six different preprocessing algorithm modeling results, the models with the better effects are RAW-KS-GBR, SG-SPXY-GBR and SNVT-SPXY-GBR. The SG-SPXY-GBR model has the highest R_p^2 value of 0.9365, and its R_c^2 value also reaches 0.9563, with a small discrepancy

between them. This manifests that the model established with SG as the preprocessing algorithm and SPXY as the sample division method not only provides high accuracy but also has better robustness. In summary, the SG algorithm is finally selected to preprocess the original hyperspectral data of Tibetan tea. The original spectral characteristic curve RAW and the spectral characteristic curve after SG preprocessing are shown in Figure 3.

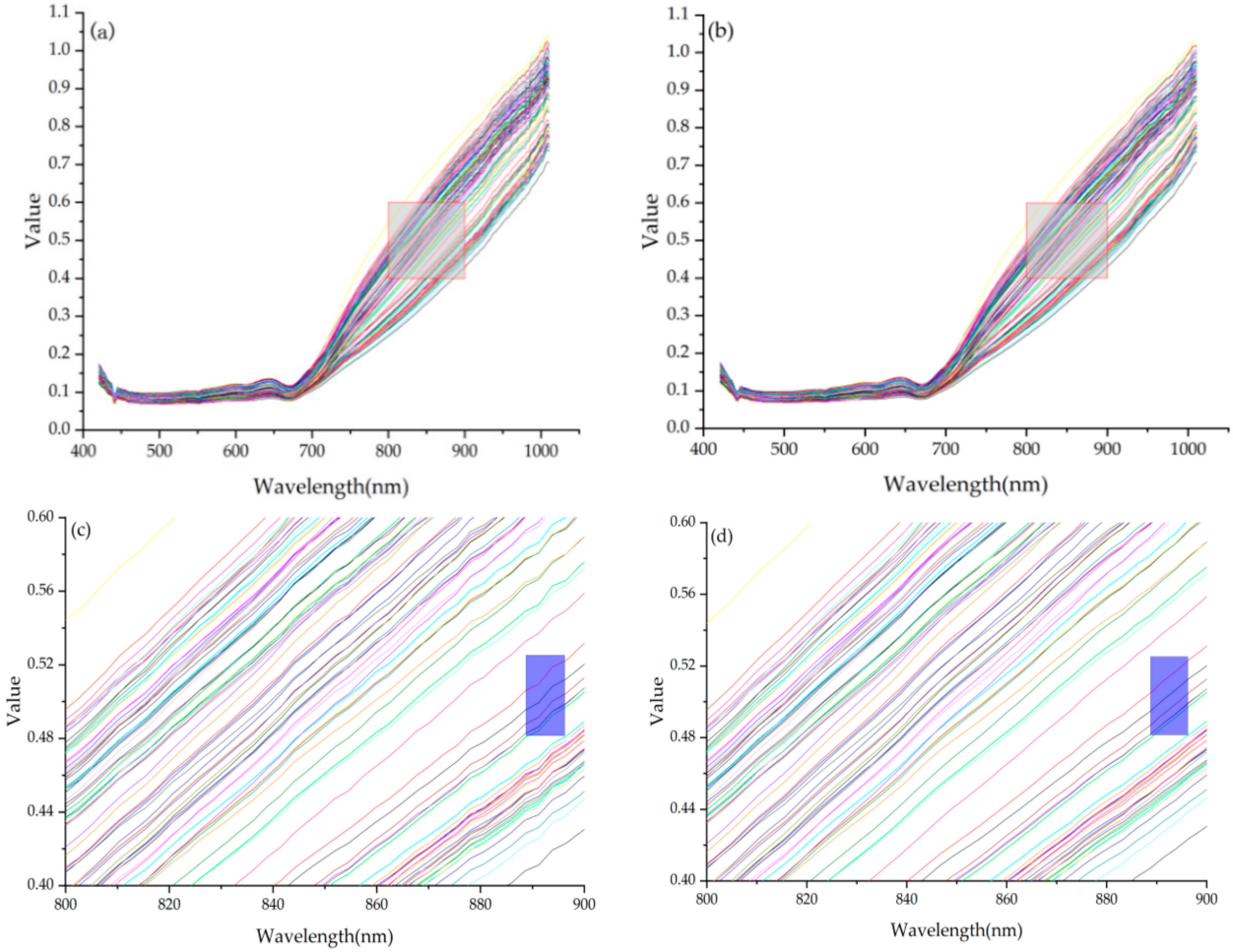

Figure 3. Tibetan tea spectrum curve. (**a**) Raw data; (**b**) Data preprocessed by SG algorithm; (**c**) Enlarged view of the red frame in Figure (**a**); (**d**) Enlarged view of the red frame in Figure (**b**).

The spectral curve Figure 3b, after SG preprocessing, is smoother than the raw spectral data in Figure 3a. Figure 3c,d show partial enlarged views corresponding to the red boxes in Figure 3a,b. The blue shaded part clearly shows this point of view, indicating that the algorithm can effectively filter out noise. The SPXY algorithm is selected to divide the calibration set and the test set. After the division, the statistical results of Tibetan tea polyphenol content are shown in Table 1. Figure 4 shows the prediction results of Tibetan tea polyphenols content by GBR model after SG algorithm preprocessing and SPXY algorithm partitioning of the data set. The horizontal axis represents the actual measured value, and the vertical axis represents the predicted value of the established model.

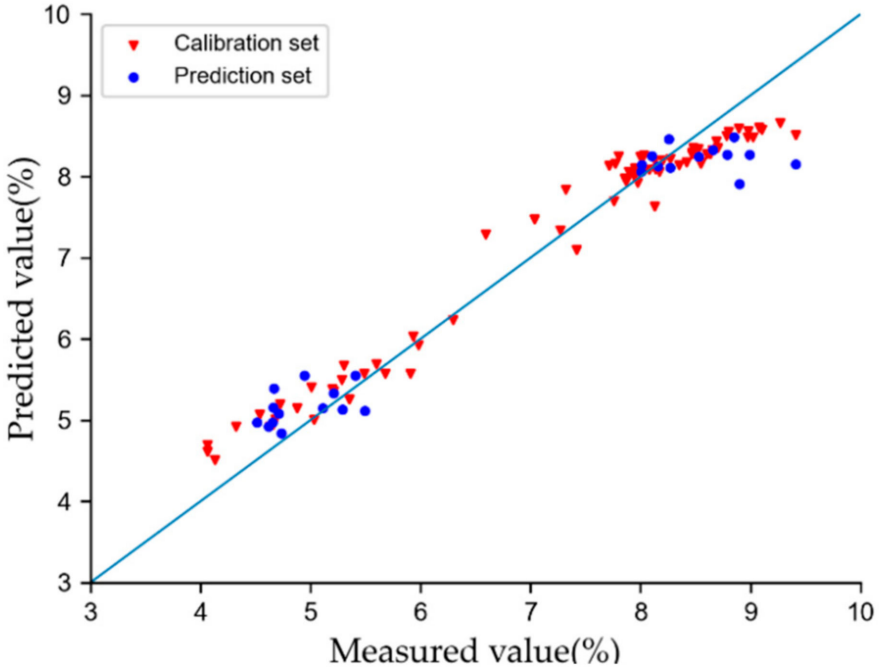

Figure 4. Prediction results of tea polyphenols after pretreatment of raw spectra by SG algorithm, and "%" indicates the percentage of tea polyphenols in the dry matter of the tea sample.

3.2. Selection of Characteristic Bands of Tibetan Tea Hyperspectral Data

The data noise after SG algorithm preprocessing has been improved to a certain extent, but there is still a lot of information unrelated to the prediction of tea polyphenol content in the data. If the spectrum number is not further extracted, the high-dimensional spectrum data will undoubtedly affect the accuracy and robustness of the model. In this study, six algorithms, including GB, AdaBoost, RF, CatBoost, LightGBM and XGBoost, have been used to select the top 30 Tibetan tea spectral characteristic bands. The final characteristic bands obtained are shown in Figure 5.

The feature selection algorithms RF and CatBoost take the wavelength of 522.66 nm as the second most important feature, while XGBoost takes the band of 564.55 nm as the first feature, which only ranks fifth in GB algorithm, fourth in AdaBoost algorithm and seventh in RF algorithm. The characteristic wavelengths extracted by different algorithms are mostly distributed between 420 and 700 nm. The experimental results show that the characteristic wavelengths extracted by different algorithms are different but also share some qualities. The features extracted by the above six feature extraction algorithms will be used as the input of the subsequent regression prediction algorithm.

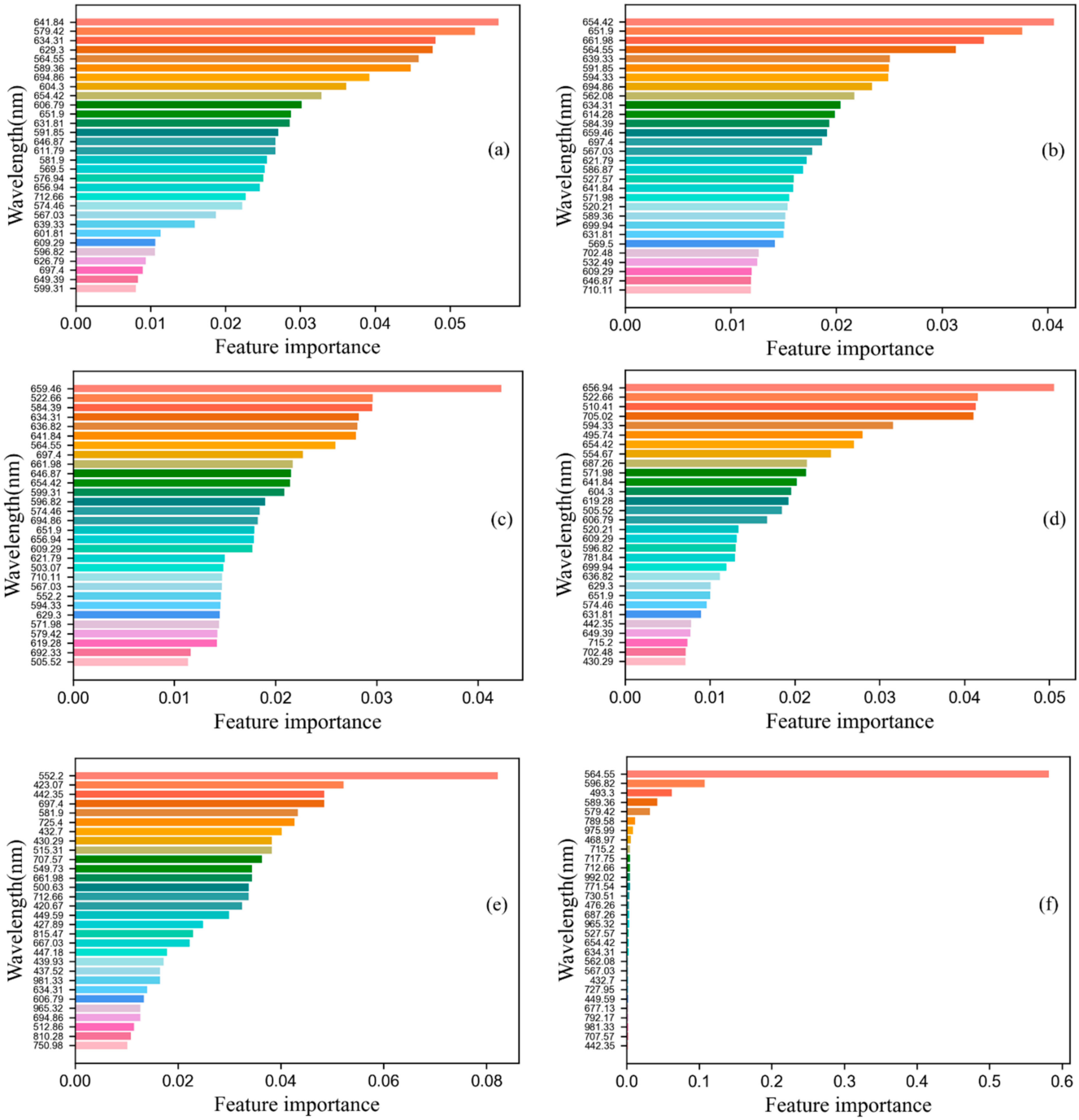

Figure 5. Feature bands selected by different algorithms. (**a**) GB; (**b**) AdaBoost; (**c**) RF; (**d**) CatBoost; (**e**) LightGBM and (**f**) XGBoost.

3.3. Results of Models

3.3.1. Full-Band Modeling Results

The SG algorithm is used to preprocess the original spectral data, and the processed data is used for modeling and prediction. Table 2 shows the prediction results of different individual models. Among them, the CatBoostR model is the most accurate, with its R_c^2 and R_p^2 at 0.9578 and 0.9493, respectively. The model of RFR prediction effect is poor, and the coefficient of determination of the calibration set is only 0.9040.

Table 2. Detection results based on different models of full spectrum.

Model	R_c^2	RMSEC	R_p^2	RMSEP
RFR	0.9040	0.4929	0.9470	0.4244
CatBoostR	0.9578	0.3266	0.9493	0.4149
LightGBMR	0.9419	0.3833	0.9259	0.5015
XGBoostR	0.9574	0.3283	0.9463	0.4271

3.3.2. Modeling Results of Characteristic Bands

Six groups of Tibetan tea spectral features are selected using different feature extraction algorithms and used as inputs to the RFR, CatBoostR, LightGBMR and XGBoostR models. At the same time, based on the stacking combination strategy, RFR, LightGBM and XGBoostR are used as three base-learners, and CatBoostR is used as a meta-learner to build a new predictive model (Stacking model). The built Stacking model is shown in Figure 6.

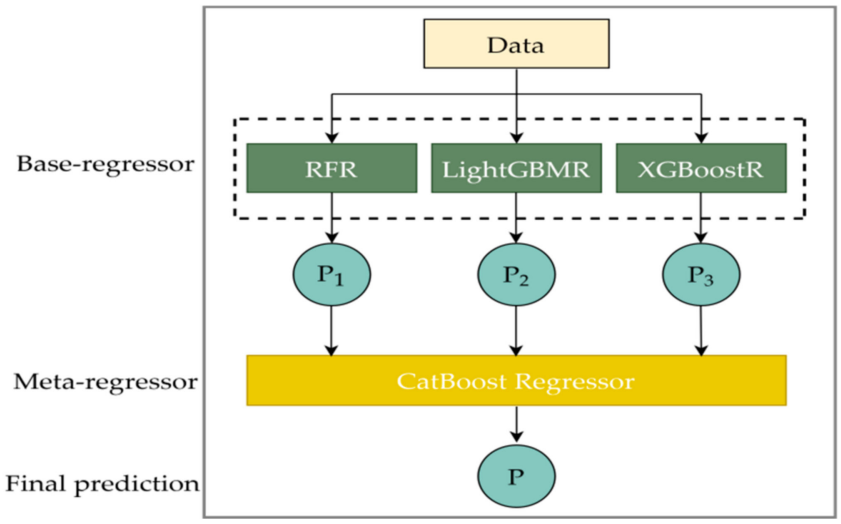

Figure 6. Flow chart depicting the stacking regressor model used for tea polyphenols prediction.

Table 3 shows the prediction results of different models. Compared with the full-band modeling results, even if the feature dimension is reduced, the model performance is not reduced accordingly. The modeling accuracy has been improved to a certain extent, and the robust performance has also been further improved. The prediction accuracy of the CatBoostR model is generally acceptable, with RMSEC lower than 0.35 and RMSEP lower than 0.45. The R_c^2 and R_p^2 of the XGBoost + CatBoostR model are 0.9744 and 0.9509, respectively, and the RMSEC and RMSEP are 0.2546 and 0.4084, respectively. The R_c^2 and R_p^2 of the LightGBM + CatBoostR model are 0.9753 and 0.9520, respectively, and the RMSEC and RMSEP are 0.2499 and 0.4035, respectively. The R_c^2 and R_p^2 of the CatBoost + CatBoostR model are 0.9697 and 0.9563, respectively, and the RMSEC and RMSEP are 0.2766 and 0.3852, respectively. The RMSEC and RMSEP values of the CatBoost + CatBoostR model are closer. Therefore, this model is considered the best among the four individual models.

Table 3. Predictive performance of models using the characteristic wavelengths extracted by different algorithms.

Model	Method	R_c^2	RMSEC	R_p^2	RMSEP
RFR	GB	0.9413	0.3852	0.9411	0.4471
	AdaBoost	0.9416	0.3844	0.9398	0.4521
	RF	0.9496	0.3570	0.9457	0.4295
	CatBoost	0.9540	0.3411	0.9523	0.4023
	LightGBM	0.9530	0.3449	0.9516	0.4052
	XGBoost	0.9586	0.3236	0.9478	0.4211
CatBoostR	GB	0.9539	0.3413	0.9467	0.4255
	AdaBoost	0.9556	0.3350	0.9406	0.4487
	RF	0.9597	0.3193	0.9494	0.4146
	CatBoost	0.9697	0.2766	0.9563	0.3852
	LightGBM	0.9753	0.2499	0.9520	0.4035
	XGBoost	0.9744	0.2546	0.9509	0.4084
LightGBMR	GB	0.9492	0.3588	0.9203	0.5201
	AdaBoost	0.9492	0.3585	0.9406	0.4491
	RF	0.9505	0.3539	0.9468	0.4249
	CatBoost	0.9617	0.3114	0.9418	0.4446
	LightGBM	0.9575	0.3278	0.9365	0.4643
	XGBoost	0.9510	0.3521	0.9352	0.4692
XGBTR	GB	0.9524	0.3468	0.9315	0.4822
	AdaBoost	0.9669	0.2893	0.9237	0.5088
	RF	0.9591	0.3218	0.9322	0.4798
	CatBoost	0.9686	0.2819	0.9457	0.4296
	LightGBM	0.9557	0.3346	0.9457	0.4295
	XGBoost	0.9578	0.3266	0.9524	0.4020
Stacking	GB	0.9600	0.3181	0.9357	0.4674
	AdaBoost	0.9608	0.3148	0.9538	0.3962
	RF	0.9579	0.3262	0.9452	0.4313
	CatBoost	0.9709	0.2711	0.9625	0.3569
	LightGBM	0.9653	0.2964	0.9536	0.3968
	XGBoost	0.9702	0.2746	0.9526	0.4013

In the Stacking model built in this article, the model with the characteristic band extracted by the CatBoost algorithm as the input is the most effective. The value of R_c^2 is 0.9709, RMSEC is 0.2711, R_p^2 is 0.9625 and RMSEP is 0.3568. The prediction accuracy is higher than that of other individual regressors, and as a result, it is the most accurate. Figure 7a is the prediction result of the CatBoost + Stacking model on the content of tea polyphenols in Tibetan tea. The horizontal axis represents the actual measured value, and the vertical axis represents the predicted value of the established model. Due to the small number of samples with a tea polyphenol content of about 7%, the data partition algorithm SPXY did not allocate the test set near this value. Therefore, in the data set divided by SPXY, the sample corresponding to the tea polyphenol content of the calibration set of 7.2671% is selected as one of the test set samples, and the sample corresponding to the tea polyphenol content of the test set of 8.7892% is selected as one of the calibration set samples. If the replaced data is input into the CatBoost + Stacking model, Figure 7b shows the prediction results. The value of R_c^2 is 0.9686, RMSEC is 0.2833, R_p^2 is 0.9577 and RMSEP is 0.3703.

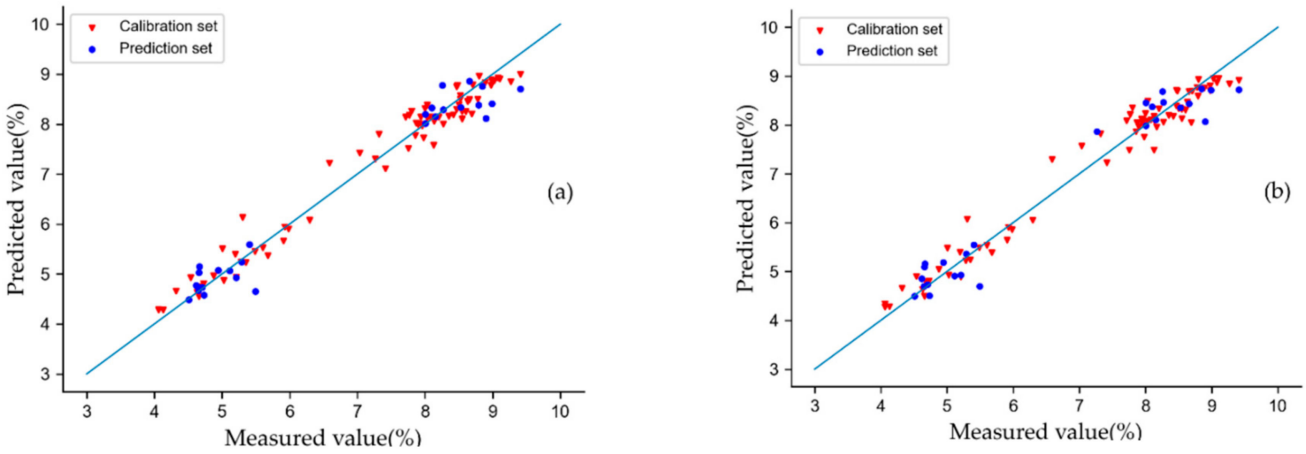

Figure 7. Prediction results of tea polyphenols based on CatBoost + Stacking model, "%" indicates the percentage of tea polyphenols in the dry matter of tea sample. (**a**) The modeling results before replacing the samples. (**b**) The modeling results after replacing the samples.

4. Discussions

A detection model of Tibetan tea polyphenols is established based on hyperspectral technology. The test results show that the spectral data preprocessing algorithm SG can effectively eliminate noise. Band selection can improve the prediction accuracy and robustness of the model.

The final modeling results show that the characteristic band selection method used in this study is effective. The 233 feature variables are reduced to 30, but the accuracy of the model does not decrease as a result. Generally speaking, the effect of the CatBoostR individual model is better than other individual models. The calibration set of LightGBM + CatBoostR and XGBoost + CatBoostR models has performed well. However, the prediction set does not perform well, and the difference between the RMSEC and RMSEP of the model is large, and the robustness of the model is low. Among all the models, the CatBoost + Stacking model built in this paper is the most effective. The determination coefficients R_c^2 and R_p^2 are 0.9709 and 0.9625, respectively, and the RMSEC and RMSEP are 0.2711 and 0.3569, respectively. The data divided by the SPXY algorithm has no test sample with a tea polyphenol content of about 7%. In order to improve the credibility of the model, one sample is selected for replacement in the calibration set and the test set (see Section 3.3.2). The final result is slightly lower than the result before replacing the sample. The reason for this phenomenon may be due to the fact that there are fewer samples with a tea polyphenol content of about 7%, and the model has not been trained perfectly. Trained models and examples can be found here: https://github.com/luo-rochon/example_for_stacking_model (accessed on 25 June 2021).

Traditional detection methods for total tea polyphenols include ferrous tartrate colorimetry [44–46], potassium permanganate titration [45,47,48], folin phenol colorimetry [45,49], electrochemical method [50], among which folin phenol colorimetry is the most widely used. The principle of ferrous tartrate colorimetry [44] is to use polyphenols to react with ferrous tartrate under a certain pH value to form a blue-violet complex, which is quantified by spectrophotometry. This method has good reproducibility but requires a large sample size and a long measurement cycle. In addition, the detection result is slightly higher than the true value.

The potassium permanganate oxidation titration method [48,51] uses potassium permanganate to oxidize tea polyphenols to fade the potassium permanganate solution. The decrease in absorbance is measured at the maximum absorption wavelength, which can indirectly determine the content of tea polyphenols. The method is convenient and does not require the use of expensive equipment. However, in addition to being able to oxidize

some non-polyphenolic substances, the titration endpoint is difficult to grasp, resulting in large measurement errors.

Folin phenol colorimetry [49,51] is a common method for the determination of plant phenols in the world. Polyphenol compounds have the -OH group in tea polyphenols that are easily oxidized and appear blue. The absorbance is measured at a wavelength of 765 nm. This method has the most accurate measurement results, but the measurement process is more complicated, and the detection cost is relatively high.

According to the electrochemical properties of substances in the solution and its change rule, an electrochemical analysis method [50] was established based on the existence of certain electrical parameters such as potential, conductivity, current and electricity and the concentration of the measured substance. This method has the advantages of intuitive sensitivity, simple and rapid, wide determination range and is not susceptible to color, precipitation and other non-polyphenol organic compounds. However, the preparation process of the electrode and the surface treatment of the electrode needs to be further studied, and a chemical buffer is also needed to increase the cost of a single measurement.

In summary, most of the traditional tea polyphenol detection methods are more or less the use of certain chemical reagents, resulting in increased measurement costs and environmental pollution, in addition to the sample damage. As a new detection technology, the biggest advantage of hyperspectral technology is that it can quickly, nondestructively and in real-time detect agricultural products. The deficiency is that the test sensitivity is low, and the quantitative analysis of unknown samples must be realized by establishing a correction model. The establishment process of the correction model is relatively complex and requires a large number of training samples. Finally, it has to be mentioned that hyperspectral equipment is more expensive, but its reusability can make up for this defect.

At present, there is still a lack of tools and methods for the rapid and nondestructive determination of tea polyphenol content in the tea production process. In this study, only the total polyphenol content was predicted. In future research, we will explore the feasibility of tea polyphenol monomer detection based on this technology. The combination of hyperspectral technology and an integrated algorithm can be used for the online determination of Tibetan tea polyphenol content. At the same time, it provides a reference for the internal quality testing of other agricultural products.

Author Contributions: Conceptualization, X.L.; methodology, Z.K.; software, X.L.; validation, L.X., P.H. and Y.W.; formal analysis, P.W.; investigation, X.L.; resources, P.H.; data curation, Y.W.; writing—original draft preparation, X.L.; writing—review and editing, X.L. and Y.H.; visualization, J.L.; supervision, Z.K.; project administration, L.X.; funding acquisition, Z.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the subject double support program of Sichuan Agricultural University (Grant NO. 035-1921993093).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This data can be found here: https://github.com/luo-rochon/Experimental-data (accessed on 13 May 2021); https://www.kaggle.com/lxrochon/tibetan-tea-experimental-data (accessed on 13 May 2021).

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

References

1. Frei, B.; Higdon, J.V. Antioxidant activity of tea polyphenols in vivo: Evidence from animal studies. *J. Nutr.* **2003**, *133*, 3275S–3284S. [CrossRef]
2. Lin, J.-K.; Liang, Y.-C.; Lin-Shiau, S.-Y. Cancer chemoprevention by tea polyphenols through mitotic signal transduction blockade. *Biochem. Pharmacol.* **1999**, *58*, 911–915. [CrossRef]
3. Khan, N.; Mukhtar, H. Tea polyphenols for health promotion. *Life Sci.* **2007**, *81*, 519–533. [CrossRef] [PubMed]
4. Guo, Y.; Chen, Q.; Huang, J.R.; Xue-Yuan, W.U.; Qiong, W.U. The tea flavor quality and its ingredients. *J. Tea Commun.* **2015**, *3*, 312–318.

5. Liu, Y.; Zhou, S.; Han, W.; Li, C.; Liu, W.; Qiu, Z.; Chen, H. Detection of Adulteration in Infant Formula Based on Ensemble Convolutional Neural Network and Near-Infrared Spectroscopy. *Foods* **2021**, *10*, 785. [CrossRef]
6. Kang, M.Y.; Yue, L.I.; Wei, W.J.; Jing, B.Y. Determination of Tea Polyphenols by HPLC. *Food Res. Dev.* **2014**, *35*, 14–15.
7. Zhao, F.; Lin, H.-T.; Zhang, S.; Lin, Y.-F.; Yang, J.-F.; Ye, N.-X. Simultaneous determination of caffeine and some selected polyphenols in Wuyi Rock tea by high-performance liquid chromatography. *J. Agric. Food Chem.* **2014**, *62*, 2772–2781. [CrossRef]
8. Wang, L.; Wang, P.; Wu, L.; Xu, L.; Huang, P.; Kang, Z. Computer Vision Based Automatic Recognition of Pointer Instruments: Data Set Optimization and Reading. *Entropy* **2021**, *23*, 272. [CrossRef]
9. Arndt, M.; Drees, A.; Ahlers, C.; Fischer, M. Determination of the Geographical Origin of Walnuts (*Juglans regia* L.) Using Near-Infrared Spectroscopy and Chemometrics. *Foods* **2020**, *9*, 1860. [CrossRef]
10. McVey, C.; Gordon, U.; Haughey, S.A.; Elliott, C.T. Assessment of the Analytical Performance of Three Near-Infrared Spectroscopy Instruments (Benchtop, Handheld and Portable) through the Investigation of Coriander Seed Authenticity. *Foods* **2021**, *10*, 956. [CrossRef]
11. Imanian, K.; Pourdarbani, R.; Sabzi, S.; García-Mateos, G.; Arribas, J.I.; Molina-Martínez, J.M. Identification of Internal Defects in Potato Using Spectroscopy and Computational Intelligence Based on Majority Voting Techniques. *Foods* **2021**, *10*, 982. [CrossRef]
12. Basile, T.; Marsico, A.D.; Perniola, R. NIR analysis of intact grape berries: Chemical and physical properties prediction using multivariate analysis. *Foods* **2021**, *10*, 113. [CrossRef]
13. Wang, F.; Zhao, C.; Yang, G. Development of a Non-Destructive Method for Detection of the Juiciness of Pear via VIS/NIR Spectroscopy Combined with Chemometric Methods. *Foods* **2020**, *9*, 1778. [CrossRef]
14. Pennisi, F.; Giraudo, A.; Cavallini, N.; Esposito, G.; Merlo, G.; Geobaldo, F.; Acutis, P.L.; Pezzolato, M.; Savorani, F.; Bozzetta, E. Differentiation between Fresh and Thawed Cephalopods Using NIR Spectroscopy and Multivariate Data Analysis. *Foods* **2021**, *10*, 528. [CrossRef]
15. Wold, J.P.; O'Farrell, M.; Andersen, P.V.; Tschudi, J. Optimization of Instrument Design for In-Line Monitoring of Dry Matter Content in Single Potatoes by NIR Interaction Spectroscopy. *Foods* **2021**, *10*, 828. [CrossRef] [PubMed]
16. Nogales-Bueno, J.; Rodríguez-Pulido, F.J.; Baca-Bocanegra, B.; Pérez-Marin, D.; Heredia, F.J.; Garrido-Varo, A.; Hernández-Hierro, J.M. Reduction of the Number of Samples for Cost-Effective Hyperspectral Grape Quality Predictive Models. *Foods* **2021**, *10*, 233. [CrossRef]
17. Nicolaï, B.M.; Defraeye, T.; De Ketelaere, B.; Herremans, E.; Hertog, M.L.; Saeys, W.; Torricelli, A.; Vandendriessche, T.; Verboven, P. Nondestructive measurement of fruit and vegetable quality. *Annu. Rev. Food Sci. Technol.* **2014**, *5*, 285–312. [CrossRef] [PubMed]
18. Wei, X.; He, J.; Zheng, S.; Ye, D. Modeling for SSC and firmness detection of persimmon based on NIR hyperspectral imaging by sample partitioning and variables selection. *Infrared Phys. Technol.* **2020**, *105*, 103099. [CrossRef]
19. Yang, C.; Zhao, Y.; An, T.; Liu, Z.; Jiang, Y.; Li, Y.; Dong, C. Quantitative prediction and visualization of key physical and chemical components in black tea fermentation using hyperspectral imaging. *LWT* **2021**, *141*, 110975. [CrossRef]
20. Hu, G.; He, D.-J.; Kenneth, A. Soil phosphorus and potassium estimation using visible-near infrared reflectance spectroscopy with direct orthogonal signal correction. *Trans. Chin. Soc. Agric. Mach* **2015**, *46*, 139–145.
21. Wang, H.; Wang, K.; Zhu, X.; Zhang, P.; Yang, J.; Tan, M. Integration of Partial Least Squares Regression and Hyperspectral Data Processing for the Nondestructive Detection of the Scaling Rate of Carp (*Cyprinus carpio*). *Foods* **2020**, *9*, 500. [CrossRef]
22. Rindang, A.; Ayu, P. Prediction of water content in Lintong green bean coffee using FT-NIRS and PLS method. In Proceedings of the IOP Conference Series: Earth and Environmental Science, Medan, Indonesia, 10 October 2020; p. 012047.
23. Xu, L.; Zhou, Y.-P.; Tang, L.-J.; Wu, H.-L.; Jiang, J.-H.; Shen, G.-L.; Yu, R.-Q. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. *Anal. Chim. Acta* **2008**, *616*, 138–143. [CrossRef]
24. Schoot, M.; Kapper, C.; van Kollenburg, G.H.; Postma, G.J.; van Kessel, G.; Buydens, L.M.; Jansen, J.J. Investigating the need for preprocessing of near-infrared spectroscopic data as a function of sample size. *Chemom. Intell. Lab. Syst.* **2020**, *204*, 104105. [CrossRef]
25. Wang, P.; Liu, J.; Xu, L.; Huang, P.; Luo, X.; Hu, Y.; Kang, Z. Classification of Amanita Species Based on Bilinear Networks with Attention Mechanism. *Agriculture* **2021**, *11*, 393. [CrossRef]
26. Borin, A.; Poppi, R.J. Application of mid infrared spectroscopy and iPLS for the quantification of contaminants in lubricating oil. *Vib. Spectrosc.* **2005**, *37*, 27–32. [CrossRef]
27. Pereira, A.F.C.; Pontes, M.J.C.; Neto, F.F.G.; Santos, S.R.B.; Galvao, R.K.H.; Araujo, M.C.U. NIR spectrometric determination of quality parameters in vegetable oils using iPLS and variable selection. *Food Res. Int.* **2008**, *41*, 341–348. [CrossRef]
28. Li, Y.; Guo, M.; Shi, X.; Wu, Z.; Li, J.; Ma, Q.; Qiao, Y. Online near-infrared analysis coupled with MWPLS and SiPLS models for the multi-ingredient and multi-phase extraction of licorice (Gancao). *Chin. Med.* **2015**, *10*, 1–10. [CrossRef] [PubMed]
29. Wang, X.; Bao, Y.; Liu, G.; Li, G.; Lin, L. Study on the best analysis spectral section of NIR to detect alcohol concentration based on SiPLS. *Procedia Eng.* **2012**, *29*, 2285–2290. [CrossRef]
30. Zou, X.; Zhao, J.; Li, Y. Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of 'Fuji'apple based on BiPLS and FiPLS models. *Vib. Spectrosc.* **2007**, *44*, 220–227. [CrossRef]
31. Li, P.-F.; Wang, J.-H.; Cao, N.-N.; Han, D.-H. Selection of variables for MLR in Vis/NIR spectroscopy based on BiPLS combined with GA. *Spectrosc. Spectr. Anal.* **2009**, *29*, 2637–2641.
32. Shi, J.-Y.; Zou, X.-B.; Zhao, J.-W.; Mao, H.-P. Selection of wavelength for strawberry nir spectroscopy based on bipls combined with saa. *J. Infrared Millim. Waves* **2011**, *30*, 458–462. [CrossRef]

33. Kennard, R.W.; Stone, L.A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137–148. [CrossRef]
34. Morais, C.L.; Santos, M.C.; Lima, K.M.; Martin, F.L. Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation Kennard-Stone algorithm approach. *Bioinformatics* **2019**, *35*, 5257–5263. [CrossRef]
35. Zhan, X.-R.; Zhu, X.-R.; Shi, X.-Y.; Zhang, Z.-Y.; Qiao, Y.-J. Determination of hesperidin in tangerine leaf by near-infrared spectroscopy with SPXY algorithm for sample subset partitioning and Monte Carlo cross validation. *Spectrosc. Spectr. Anal.* **2009**, *29*, 964–968.
36. Galvao, R.K.H.; Araujo, M.C.U.; José, G.E.; Pontes, M.J.C.; Silva, E.C.; Saldanha, T.C.B. A method for calibration and validation subset partitioning. *Talanta* **2005**, *67*, 736–740. [CrossRef] [PubMed]
37. Huichun, Y.U.; Lou, N.; Yin, Y.; Liu, Y.H. Predictive Model for Detection of Maize Toxins with Sample Set Partitioning Based on Joint x-y Distance (SPXY) Algorithm and Successive Projections Algorithm (SPA) Based on Hyperspectral Imaging Technology. *Food Sci.* **2018**, *39*, 328–335.
38. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *arXiv* **2017**, arXiv:1706.09516.
39. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
40. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
41. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
42. Renaud, O.; Victoria-Feser, M.-P. A robust coefficient of determination for regression. *J. Stat. Plan. Inference* **2010**, *140*, 1852–1862. [CrossRef]
43. Nørgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J.P.; Munck, L.; Engelsen, S.B. Interval partial least-squares regression (i PLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc.* **2000**, *54*, 413–419. [CrossRef]
44. Liu, B.-B.; Chen, S.-H.; Zou, J.-Y.; Li, Q.-W.; Chen, H.-J. Determination of tea polyphenol in Huangli buds by ferrous tartrate method. *Cent. South Pharm.* **2011**, *10*, 740–741.
45. Tan, H.P.; Zou, Y.; Shan-Rong, Y.E.; Chen, L. Review of tea polyphenols analyses for tea. *China Meas. Test. Technol.* **2008**, *4*, 4–11.
46. Wang, L.; Wu, L.; Yao, Y.; Xia, J. Spectrophotometric Determination of Tea-Polyphenol with Ferrous Tartrate. *Chin. J. Spectrosc. Lab.* **1997**, *3*, 52–54.
47. Yuan, Y.; Zhang, S.F.; Shuang, S.M.; Dong, C. Determination of Tea Polyphenols in Tea by 1.10-phenanthroline-iron(II) Indicator. *Food Sci.* **2008**, *29*, 403–405.
48. Wen, X.; Changqing, T.U. Discoloring Spectrophotometric Determination of Tea Polyphenols by Potassium Permanganate. *Food Ind.* **2019**, *40*, 278–281.
49. Wang, Y.B.; Yan, X.U.; Wei, Y.C.; Liu, X.Q.; Liu, X.H. Determination of total polyphenols in the zinc complex of polyphenol by Folin-Ciocalteu colorimetry. *Chem. Res.* **2011**, *22*, 76–78.
50. Kou, L.; Liang, R.; Qin, W.; Liang, R. Potentiometric determination of total polyphenols in green tea based on complexation-reaction-induced response. *Int. J. Electrochem. Sci.* **2014**, *9*, 3190–3198.
51. Wang, L.L.; Chen, J.; Song, Z.S.; Yang, J.G.; Chen, L. Advances in Study on Test Method of Tea Polyphenols in Tea. *Tea Sci. Technol.* **2013**, *4*, 6–12.