Article

Detection of the Pigment Distribution of Stacked Matcha During Processing Based on Hyperspectral Imaging Technology

Qinghai He [1,2], Zhiyuan Liu [1,2], Xiaoli Li [3,*], Yong He [3] and Zhi Lin [4]

1   School of Mechanical Engineering, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China; liuzhiyuan2023@126.com (Z.L.)
2   Shandong Academy of Agricultural Machinery Science, Jinan 250100, China
3   College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China
4   Tea Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou 310008, China
*   Correspondence: xiaolili@zju.edu.cn

Abstract: Color is a key indicator for evaluating the quality of tea during processing; various processing procedures can significantly affect the content of fat-soluble pigments of tea, which in turn affects the color and quality of finished tea. Therefore, there is an urgent demand for the fast, non-destructive detection of pigments of stacked tea during processing. This paper presents the use of hyperspectral imaging technology (HSI), combined with machine learning algorithms, to detect chlorophyll a, chlorophyll b, and carotenoids in stacked matcha tea during processing. Firstly, a quantitative relationship between HSI data of tea and their pigment contents was developed based on regression analysis, and the results showed that exceptional prediction performance was achieved by the partial least squares regression (PLSR) algorithm combined with the feature band algorithm of competitive adaptive reweighting (CARS), and the R_P^2 values of detection models of chlorophyll a, chlorophyll b and carotenoids were 0.90465, 0.92068 and 0.62666, respectively. Then, these quantitative detection models were extended to each pixel in hyperspectral images, achieving point-by-point prediction of pigment components, so the distribution of pigments of stacked tea leaves during processing procedures was successfully visualized on the processing line in situ. By integrating a hyperspectral imaging system into the real-world environment, operators can monitor pigment levels in real time and thus dynamically adjust processing parameters based on real-time data. This study enhances pigment detection efficiency in tea processing, supports process optimization, and aids in quality control.

Keywords: hyperspectral imaging technology; processing procedures; in situ detection; non-destructive determination

1. Introduction

China is the world's largest producer and consumer of tea, which first originated in southwestern China more than 3000 years ago, making it the world's earliest producer of tea. In ancient times, tea was considered a plant with medicinal value, and it was popular and exported to foreign countries due to its unique flavor and refreshing aroma [1]. In recent years, more and more studies have found that green tea is beneficial to human health, such as reducing the occurrence of cardiovascular diseases and the prevention and treatment of many kinds of tumors and cancers [2,3].

Color is one of the most important attributes in the evaluation index of tea, and chlorophyll and carotenoids contained in green tea are some of the important factors affecting the color [4,5], whereas among them, chlorophyll a and chlorophyll b determine the green hue of the tea leaves, and the carotenoids determine the yellow hue, both of which are fat-soluble pigments, occupying 0.3% to 0.8% and 14.4% to 29.2% of the tea weight, respectively [6,7]. During the processing of green tea, the processing process has a greater impact on the fat-soluble pigments contained in green tea [8], in which the greening process

influences the enzyme activity therein to form more demagnetized chlorophylls, which deepen the color of tea leaves, and carotenoids, which are more significant for processing, especially after greening and drying, with decreases of up to 50% and 60%, respectively [9]. Therefore, in situ detection of pigment content during tea processing can help processors adjust processing parameters to retain or derivatize more fat-soluble pigments to ensure the "three greens" of tea color [10].

Spectroscopy is a technique that gathers information by measuring the interaction between matter and electromagnetic radiation. It operates on the principle that different wavelengths of light interact with matter in unique ways, producing distinct effects that provide valuable insights into the substance. When materials are exposed to electromagnetic radiation, they absorb light at specific wavelengths; this absorption occurs because each compound possesses a unique absorption spectrum. Consequently, spectroscopy facilitates both qualitative and quantitative analysis, allowing for the identification of substances and their concentrations. Given these distinctive properties, spectroscopy finds applications across a wide range of fields. In this context, several studies have been devoted to exploring the application of spectroscopic techniques in tea quality analysis. Yuzhen Wei et al. [11] explored the generalization performance of a tea moisture detection model under different leaf surface orientations and tea varieties. By performing fractional order differencing on the VNIR spectral data, it was found that the fractional orders of 0.4 and 0.6 significantly improved the generalization ability of the model under different varieties and leaf surface orientations. Guangxin Ren et al. [12] used a multivariate selection strategy to identify key feature wavelengths related to the quality classification of Dian Hong tea. Combining the methods of Improved Genetic Algorithm (IGA) and Particle Swarm Optimization (PSO), the Support Vector Machine (SVM) model achieved a 95.28% correct discrimination rate in tea quality prediction. Qin Ouyang et al. [13] investigated the use of VIS-NIR spectroscopy for estimating the sensory quality of color in black tea samples and found that the spectral data-based GA-BPANN model performed the best, with a correlation coefficient of 0.8935 in the prediction set, which was superior to the model based on color parameters. Ying Liu et al. [14] used VIS-NIR spectroscopy combined with an SVM model for black tea quality classification, and the correct recognition rate in the validation set using the CARS-linear kernel SVM model was 91.85%. The results showed that VIS-NIR spectroscopy is a fast, low-cost, and efficient method for tea quality prediction.

Currently, a commonly used method for tea pigment analysis is high-performance liquid chromatography (HPLC) [15]. However, HPLC requires expensive equipment, the preparation of various mobile phases, and complex post-detection data processing, making it time-consuming and labor-intensive. Additionally, it is not suitable for rapid, large-scale pigment analysis [16,17]. Moreover, existing pigmentation analysis methods cannot perform the non-destructive testing of tea [18,19]. In response to the limitations of conventional chemical testing methods, researchers have proposed the use of HSI technology to non-destructively determine pigment content [20,21]. For instance, Zijuan Zhang [22] used hyperspectral data for dimensionality reduction to identify characteristic bands, establishing a spectral index highly sensitive to anthocyanin content and developing an estimation model. Similarly, Lisong Jin [23] analyzed and compared the inversion accuracies of chlorophylls a and b using two hyperspectral methods—spectral index and pseudo-absorption coefficients—to identify optimal detection bands and developed a portable device for detecting chlorophyll a and b using a photosensitive sensor to capture leaf reflectance.

Although spectral and hyperspectral imaging techniques (HSI) have been employed to detect pigment content in tea, existing studies have yet to achieve real-time, online detection of pigments and their distribution in stacked tea leaves along the production line, remaining limited to the mere detection of spectral curves. While this approach can capture the spectral characteristics of tea samples, it falls short of providing a detailed analysis of the spatial distribution of pigments within the leaves. To address this issue, this paper integrates HSI technology with partial least squares regression (PLSR) and least squares

Support Vector Machine (LS-SVM). Additionally, the competitive adaptive reweighting algorithm (CARS) and Successive Projections Algorithm (SPA) are employed to select key feature bands from the hyperspectral data, reducing redundancy. By applying regression analysis, a quantitative relationship is established between the HSI data and the pigment content in tea. These quantitative detection models are then applied to each pixel in the hyperspectral images, enabling point-by-point prediction of pigment components and visualizing the spatial distribution of stacked teas at various stages of processing.

2. Materials and Methods

2.1. Samples Collection

The tea used in the experiment was collected from Hangzhou Jingle Tea Industry Co., Ltd. (Hangzhou, China), between April and May 2023. The variety selected for this study was matcha. A total of 180 tea samples, including 60 fresh leave samples (FR), 60 fixation leave samples (FI), and 60 dried leave samples (DR), were collected from the tea factory. In detail, the fresh leave sample (FR) represents the leaves just plucked from the tea plant, and the fixation process is to put the tea leaves into the drum, set the rotational speed of 400 r/min, and the high temperature of 200 °C to kill the tea for 10 min to utilize the high temperature to destroy the enzyme activity and to maintain the color of the green tea. The dried process is performed at a temperature of 170 °C for a period of 10–15 min, with a goal of evaporating the residual water in the tea leaves in order to maintain the quality of tea and increase the aroma. The specific collection information of samples is shown in Table 1.

Table 1. Sampling information of matcha.

Plucking Date	FR	FI	DR
4.28	5	6	5
4.30	8	5	4
5.2	6	7	7
5.4	6	5	4
5.6	7	7	7
5.8	7	7	9
5.10	4	4	4
5.12	6	6	4
5.14	4	4	5
5.16	4	4	4
5.18	3	5	7
Total	60	60	60

2.2. Hyperspectral Image Acquisition and Correction

The hyperspectral imaging unit was constructed as shown in Figure 1. It consists of a Finnish Specim FX10 hyperspectral camera (push scan) (Specim, Oulu, Finland), an LG-150 halogen lamp cold light source (Shanghai Meimei Metering Electricity Technology Co., Ltd., Shanghai, China), an electrically controlled conveyor belt, a conveyor belt speed adjustment controller, a polytetrafluoroethylene (PTFE) whiteboard (Guangzhou Jingyi Photoelectric Technology Co., Ltd., Guangzhou, China), a background blackboard, and a computer.

The hyperspectral camera was powered on and allowed to warm up for 15 min prior to data acquisition, a necessary step to prevent baseline drift in the spectral images, which could otherwise compromise the accuracy of the target spectrum. Calibration of the tea data was performed before the hyperspectral collection to ensure precise acquisition and eliminate potential interference. Under the same conditions as the tea samples, the light source was used to capture a dark environment image for reference. Following this, tea leaves were evenly spread on a conveyor belt for hyperspectral image acquisition. Each row scan contained 1024 pixels, with a spectral range spanning from 400 nm to 1000 nm, and a sampling interval of 0.6 nm, resulting in hyperspectral data collected at 951 wavelengths.

A whiteboard was used during the process to obtain the corrected hyperspectral image of the tea. The correction principle is as follows:

$$R = \frac{R_{\mathrm{raw}} - R_{dark}}{R_{\mathrm{bb}} - R_{dark}} \tag{1}$$

In the formula: R: corrected spectrum,

R_{raw}: collect the resulting raw spectra,
R_{bb}: standard spectra of the white plate,
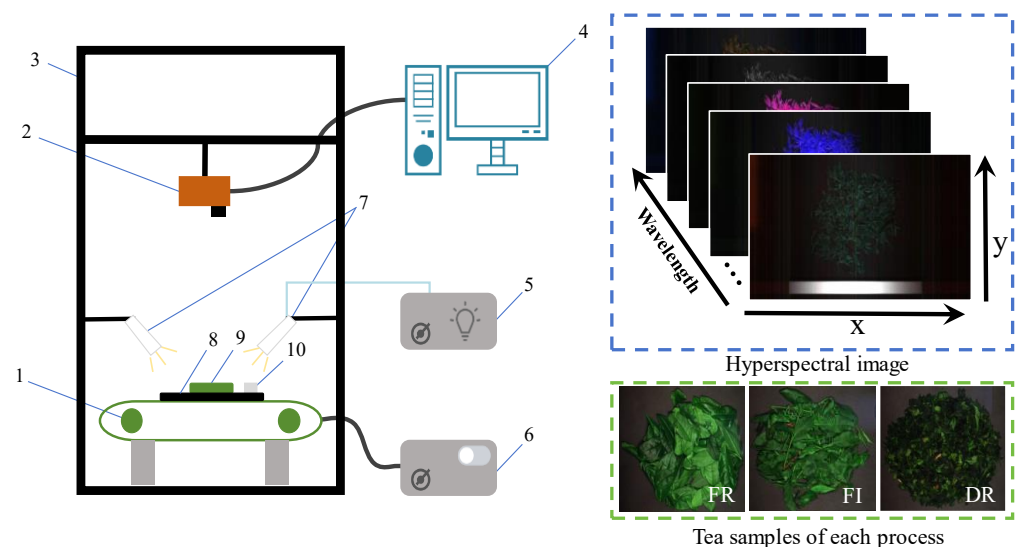R_{dark}: the dark environment spectrum collected.

Figure 1. Hyperspectral imaging device. The upper-right figure demonstrates that the acquired images contain spatial information with spectral dimensions, and the lower-right figure shows the original images of the samples for the three processing techniques. (1) Electronic controlled conveyor belt; (2) Specim FX10 hyperspectral camera; (3) bracket; (4) computer; (5) LG-150 halogen lamp cold light source; (6) conveyor belt speed adjustment controller; (7) optical fiber; (8) blackboard; (9) sample; (10) polytetrafluoroethylene (PTFE) white board.

2.3. Pigment Measured by Traditional Standard Method Destructive

Before pigment extraction, the tea samples were placed in a drying oven (STIK BAO-50A, Beijing World Trade Far East Scientific Instrument Co., Ltd., Beijing, China) for moisture removal, set at 70 °C for 72 h. After drying, the samples were transferred to an automatic rapid sample grinder (JXFSTPRP-48, Shanghai Jingxin Industrial Development Co., Ltd., Shanghai, China). The total grinding time was set to 12 min, with each interval lasting 45 s and each grinding duration set to 45 s. The grinding frequency was maintained at 65 Hz, ensuring the aperture was less than 0.3 mm, and the resulting powder passed through a 60-mesh sieve.

Approximately 0.2 g to 0.3 g of the tea powder was weighed and was weighed into a 15 mL reagent tube, to which 10 mL of an 80% acetone solution was added. The mixture was shaken thoroughly and left to stand until the powder turned gray. The reagent tube was then placed into a low-speed centrifuge (TDL-40B, Shanghai Anting Scientific Instrument Factory, Shanghai, China) and centrifuged at 1000 rpm for 10 min.

A liquid chromatograph was used to measure the pigment content. Before testing, the wavelength was set, and 3 mL of 80% acetone was added to a cleaned glass dish for calibration. The solution was then discarded into a waste tank, and the dish was rinsed with water. Next, 1 mL of the sample extract and 2 mL of acetone were added, and the absorbance was measured, with the sample number recorded.

The absorbance of the sample extraction solution was measured at 445.5 nm, 645 nm, and 663 nm, noted as I_1, I_2, and I_3, respectively. Based on these values and the sample weight (m), the concentrations of chlorophyll a, chlorophyll b, and carotenoids were calculated as follows:

$$C_A = b_1 \cdot I_3 - b_2 \cdot I_2$$
$$C_B = \frac{C_A \cdot b_3}{m} \tag{2}$$
$$C_{TC} = b_4 \cdot I_2 - b_5 \cdot I_3$$

Formula (3) is utilized to convert the concentration into the actual pigment content in dry tea as follows:

$$Chla = \frac{C_{TC} \cdot b_3}{m}$$
$$Chlb = b_6 \cdot I_1 - (C_A + C_{TC}) \cdot b_7 \tag{3}$$
$$ChlT = \frac{C_{Chlb} \cdot b_3}{m}$$

In Equation (3), *Chla*, *Chlb*, and *ChlT* represent the concentrations of chlorophyll a, chlorophyll b, and carotenoids in dry tea, expressed in mg/g. The constants b_1, b_2, b_3, b_4, b_5, b_6, and b_7 are all fixed.

2.4. Characteristic Wavelength Selection

2.4.1. Competitive Adaptive Reweighted Sampling (CARS)

Competitive Adaptive Reweighted Sampling (CARS) [24] is a spectral data analysis method for selecting characteristic wavelengths. Through adaptive reweighted sampling, CARS selects wavelengths based on the absolute regression coefficient values in the partial least squares (PLS) model, excludes wavelengths with small weights, and determines a subset of wavelengths with the lowest cross-validated root mean square error (RMSECV) through iterative validation to optimize the variable combinations. This method effectively reduces the dimensionality and computational complexity of hyperspectral data while improving model predictive performance.

2.4.2. Successive Projections Algorithm (SPA)

The Successive Projections Algorithm (SPA), proposed by Gomes et al. in 2013 [25], is a method for selecting characteristic wavelengths in spectral data analysis. By analyzing peaks and troughs in the spectral data, SPA identifies prominent spectral features, facilitating sample identification and analysis. The basic principle of SPA involves iteratively selecting one wavelength at a time and adding the wavelength that shows the maximum change in each iteration. This process extracts key feature information while reducing redundant data in the original spectral matrix.

In this paper, we propose two methods for eigenband selection: Competitive Adaptive Reweighted Sampling (CARS) and the Successive Projections Algorithm (SPA), each applied to independently select the optimal eigenbands.

2.5. Regression Model and Evaluation Index

Partial least squares regression (PLSR) [26] is a widely used statistical method for modeling and identifying the relationship between two matrices, X and Y. It works by finding the multidimensional direction in the X-space that best explains the variance in the Y-space. A least squares Support Vector Machine (LS-SVM) [27] is an enhanced version of the traditional Support Vector Machine (SVM) algorithm, unlike traditional SVM, LS-SVM transforms the optimization problem into a system of linear equations by minimizing a squared loss function. This approach simplifies and streamlines the solution process, and is well suited for pattern recognition and regression analysis. Specifically, the objective function of LS-SVM can be expressed as follows:

$$\min J(w, b) = \frac{1}{2} \|w\|^2 + \frac{1}{2} \gamma \sum_{i=1}^{N} (y_i - f(x_i))^2 \tag{4}$$

Among them, ω is the weight vector, b is the bias term, and γ is the regularization parameter. To enhance model performance, for the complex hyperspectral data in this paper, a polynomial kernel (Poly kernel) is used to process these data.

In this paper, we propose using a joint feature band selection algorithm that combines partial least squares regression (PLSR) and a least squares Support Vector Machine (LS-SVM). The model's predictive performance is evaluated using the coefficient of determination (R^2) and the root mean square error (RMSE) as key indicators. A higher R^2 value, closer to 1, indicates a better model fit, while a lower RMSE, closer to zero, signifies smaller prediction errors and improved model performance.

3. Results and Analysis

3.1. Changes in Pigment Content During Matcha Processing

The range and trend of pigment content in matcha across different processing stages are shown in Figure 2. Throughout the processing, the contents of chlorophyll a (*Chla*), chlorophyll b (*Chlb*), and total chlorophyll (*ChlT*) exhibited a gradual decrease. In Figure 2a, *Chla* had the highest content in fresh leaves (FL), with a median of 3.45 mg/g, but gradually decreased as processing continued. After fixation (FI) treatment, *Chla* content dropped to 2.41 mg/g, representing a loss of approximately 30.14%. Following drying (DR), the content further declined to 2.14 mg/g, with an overall reduction of 37.97%. The loss of *Chla* is mainly due to the demagnesium reaction that occurs during processing, where high temperatures denature proteins and chlorophyll is released from chloroplasts, while, at the same time, various types of organic acids in the cells overflow and increase in acidity, where H^+ replaces the Mg^{2+} of the chlorophyll molecule, forming a dark green or greenish-brown demagnesium chlorophyll [7,8].
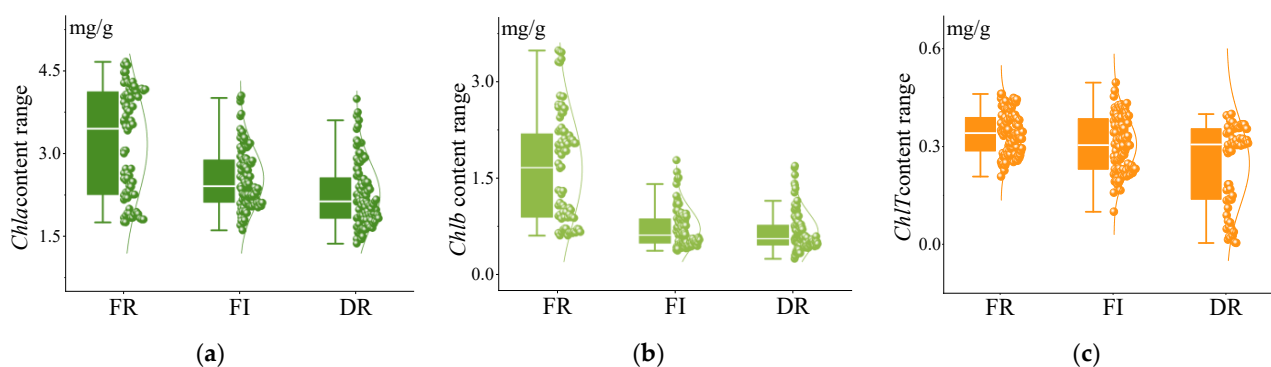

Figure 2. Variation in pigment content in different processing processes of matcha. (**a**) *Chla*, (**b**) *Chlb*, (**c**) *ChlT*.

As shown in Figure 2b, *Chlb* also decreased significantly during processing. It had a median of 1.66 mg/g in FL, which dropped to 0.62 mg/g after FI, a loss of about 62.65%, and further decreased to 0.56 mg/g after DR, with an overall reduction of 66.27%, which was more pronounced than the loss of *Chla*. This greater reduction is attributed to the presence of an additional carbonyl group in the molecular structure of *Chlb*, making it more prone to forming demagnesium chlorophyll than *Chla*.

From Figure 2c, the loss of *ChlT* was relatively smaller. Fresh leaves had the highest *ChlT* content, with a median of 0.34 mg/g, which decreased to 0.30 mg/g after FI treatment, a loss of 11.76%, and further decreased to 0.29 mg/g after DR, resulting in an overall loss of 14.71%. The reduction in *ChlT* is primarily due to their conversion into aroma compounds and other volatiles through enzymatic reactions during processing, with high temperatures accelerating these reactions [28,29].

Overall, *Chla*, *Chlb*, and *ChlT* contents showed a decreasing trend throughout processing, particularly during the drying stage. As water evaporated, increased exposure of pigments to air intensified oxidation reactions, accelerating pigment degradation. Ad-

ditionally, demagnesium reactions, along with enzymatic and non-enzymatic processes, continued during drying, further contributing to the decline in pigment content.

3.2. Spectral Response Characterization and Preprocessing

The spectral information for each sample is presented in Figure 3, showing significant differences in the spectral curves of fresh leaves (FR), fixation (FT), and drying (DR). In the 500–700 nm range, notable differences in high and low reflectance are observed across the three stages, while the spectral curve of fresh leaves diverges markedly from the other two in the 750–1000 nm range, particularly between 750 and 950 nm, where FT and DR exhibit significant variation. Chlorophyll a and b absorb light primarily in the 400–500 nm and 600–700 nm wavelengths, with weak reflectance in the 500–600 nm band due to the π electron system, conjugated double bonds, and molecular structures such as benzene and pyrrole rings. These properties allow chlorophyll to absorb light at specific wavelengths, resulting in its green color [9]. Carotenoids, with a similar conjugated double bond and isoprene structure, absorb light mainly between 400 and 550 nm, corresponding to blue and green light, while weak reflectance in the 550–700 nm range gives them a yellow to orange appearance [30,31]. In summary, the 400–1000 band covers the main spectral range of chlorophylls and carotenoids, and by analyzing the reflectance of the three processes, pigment content can be inversely predicted [32].
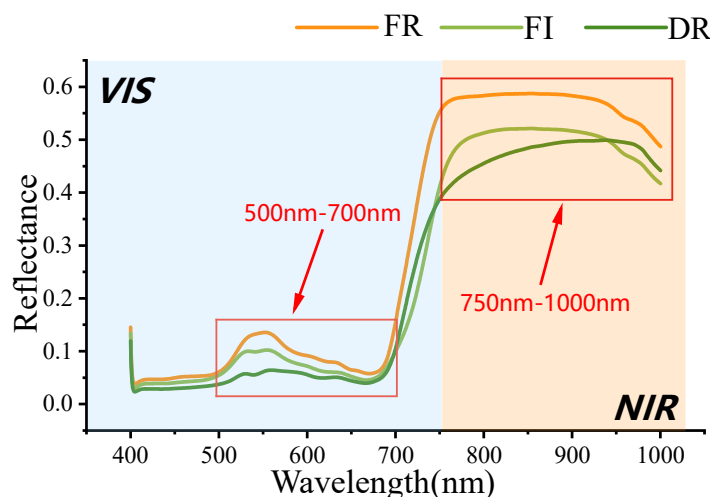

Figure 3. Average spectra of different processes. The figure shows the variability of the three processes in terms of average spectra.

In this study, hyperspectral data underwent multiple preprocessing steps to address several challenges inherent in the acquisition process.

First, the Standard Normal Variate (SNV) transformation was applied to correct for variations caused by the uneven surfaces of tea leaves and the scattering of light due to the irregular dispersion of tea across different regions. These factors can introduce significant scattered light, resulting in spectral distortions that would compromise the accuracy of the analysis.

Additionally, during HSI, the internal sensors of the instrument can generate thermal and electronic noise, further degrading the quality of the captured data [33]. To mitigate this, the Savitzky-Golay (SG) filter was employed, effectively reducing the high-frequency noise present in the spectra, thereby smoothing the data without distorting key spectral features.

Despite these preprocessing efforts, in the 400–1000 nm wavelength range, certain spikes were identified as equipment noise. Since preprocessing alone was insufficient to fully remove these artifacts, the affected spectral bands were excluded from the analysis to prevent potential biases or inaccuracies. Consequently, the final spectral curves, as shown in Figure 4d, are the product of several stages of refinement, each aimed at enhancing

data quality by addressing both physical scattering effects and sensor-based noise. This comprehensive preprocessing ensures the data are better suited for more efficient analysis and modeling.
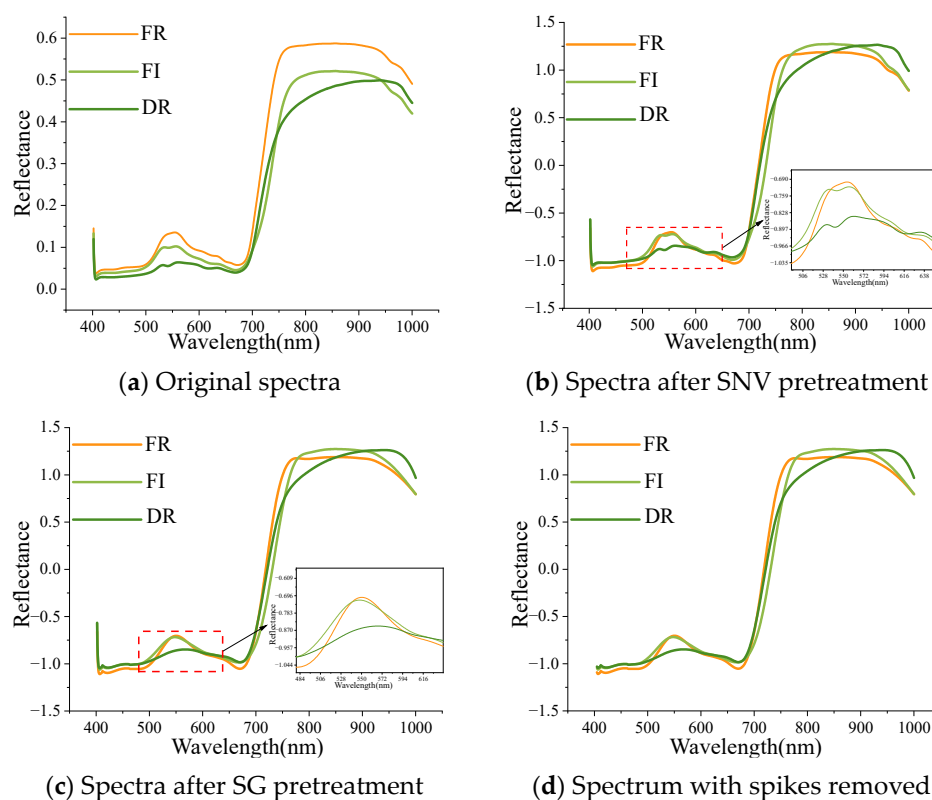

(**a**) Original spectra

(**b**) Spectra after SNV pretreatment

(**c**) Spectra after SG pretreatment

(**d**) Spectrum with spikes removed

Figure 4. Average spectra of matcha in each processing procedure after pretreatment.

3.3. Spectral Qualitative Classification Model

In order to qualitatively differentiate the spectral data during tea processing, PCA was used to visualize the individual processes. The results are shown in Figure 5, the spectral data of different processes were analyzed by PCA and the variance contributions were 88.62%, 8.46%, and 2.23%, respectively. The PCA shows distinct separation and clustering of the FL, FI, and FR groups in 3D space, indicating that the spectral data can effectively distinguish between different processing stages, which is critical for validating the spectral data and understanding the differences among the processes. In conclusion, the spectral reflectance of tea leaves shows significant variation across the three processing procedures, emphasizing the potential for using spectral reflectance to qualitatively distinguish tea samples at each stage.

3.4. Establishment and Analysis of Regression Models

Outliers in the data need to be removed before being fed into the model for training, and, in this paper, a one-class SVM is used to remove the values that are outliers in the distribution of the pigmented data. In order to visualize the changes in the data before and after screening, a detailed comparison table is provided in this paper, as shown in Table 2. This comparison table clearly demonstrates that the too-high and too-low outliers in the dataset are effectively removed after screening by the one-class SVM approach. The remaining data were divided into test and training sets at a ratio of 4:1 to prevent overfitting of the model. As shown in Table 3, the modeling results of the three pigments based on PLSR and LS-SVM without feature selection versus by feature selection with CARS and SPA are listed. Where R_c^2, RMSEC denotes the R^2 and RMSE of the training set, and where R_P^2, RMSEP denotes the R^2 and RMSE of the test set. Among the evaluation metrics listed

in Table 3, the PLSR model was found to be effective in handling the task of predicting the pigmentation content of tea leaves, and, especially when combined with the feature selection of CARS, its ability to capture the linear relationship between spectra and pigment content was optimized, with R_p^2 and RMSEP values of 0.90465 and 0.52810 for *Chla*, 0.92068 and 0.29431 for *Chlb*, and 0.62666 and 0.06855 for *ChlT*, the scatterplot is shown in Figure 6. This excellent performance is attributed to the fact that the spectral reflectance is closely related to the pigment content, especially in the visible and near-infrared regions, where specific regions in the spectral bands can directly reflect the changes in pigment content, and the PLSR model can maximize the explanatory power of these spectral features by linear combination, which makes it possible to efficiently capture the linear relationship between the pigment content and the spectra.
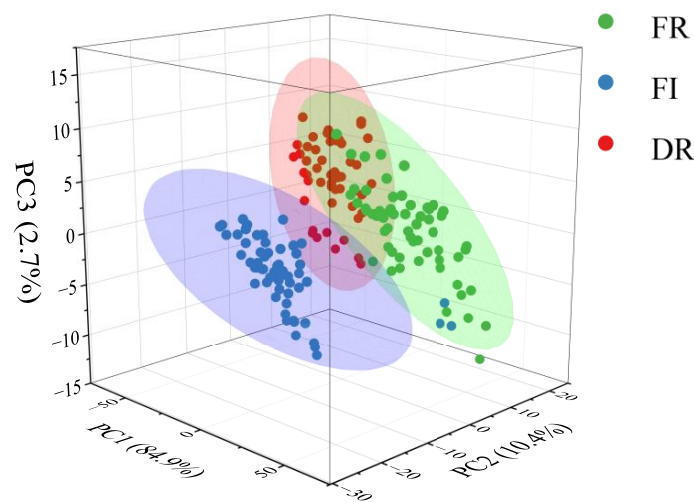

Figure 5. Distribution of tea leaves under different processing steps in three principal component spaces. The figure shows the spectra of these three processes form distinct separations and clusters in 3D space.

Table 2. Comparison of sample data before and after screening.

	Classification	Processing	Mean *	SD *	Min *	Max *	IQR *
Chla	Pre-screening	FR	2.9062	0.9384	1.7565	4.4672	1.8214
		FI	2.5349	0.5813	1.6132	3.9383	1.0337
		DR	2.2386	0.5901	1.3734	3.9866	0.8043
	Post-screening	FR	2.9077	0.8816	1.8251	4.2764	1.6626
		FI	2.5182	0.5079	1.7833	3.3932	0.9466
		DR	2.2248	0.4984	1.5481	3.4418	0.7348
Chlb	Pre-screening	FR	1.3907	0.7238	0.6057	2.7749	1.1980
		FI	0.6984	0.3077	0.3698	1.5933	0.4260
		DR	0.6235	0.2735	0.2447	1.5493	0.2916
	Post-screening	FR	1.3964	0.6835	0.6513	2.6394	1.1297
		FI	0.6853	0.2671	0.3982	1.4073	0.4096
		DR	0.6039	0.1982	0.3442	1.0889	0.2379
ChlT	Pre-screening	FR	0.3165	0.0650	0.2079	0.4611	0.0799
		FI	0.2723	0.0757	0.1581	0.4530	0.0898
		DR	0.2526	0.1305	0.0043	0.3995	0.2183
	Post-screening	FR	0.3138	0.0556	0.2420	0.4390	0.0749
		FI	0.2700	0.0687	0.1905	0.4202	0.0857
		DR	0.2583	0.1199	0.0190	0.3782	0.2092

* Mean—average value; SD—standard deviation; Min—minimum value; Max—maximum value; IQR—interquartile range.

Table 3. Comparison of modeling performance for determination of pigments in matcha samples.

Pigment	Model	Method	Modeling Result			
			R_c^2	RMSEC	R_p^2	RMSEP
Chla	PLSR	None	0.88694	0.27472	0.82198	0.26616
		CARS	**0.96808**	**0.47603**	**0.90465**	**0.52810**
		SPA	0.92860	0.42160	0.75344	0.42503
	LS-SVM	None	0.60559	0.51310	0.32543	0.51812
		CARS	0.76404	0.12558	0.63297	0.06814
		SPA	0.80212	0.03615	0.64845	0.05663
Chlb	PLSR	None	0.94344	0.14880	0.83158	0.18834
		CARS	**0.98216**	**0.26985**	**0.92068**	**0.29431**
		SPA	0.91620	0.26688	0.76371	0.30058
	LS-SVM	None	0.69362	0.34634	0.50909	0.32156
		CARS	0.56628	0.18093	0.42218	0.15377
		SPA	0.60972	0.05045	0.46779	0.03892
ChlT	PLSR	None	0.62954	0.05695	0.50115	0.05450
		CARS	**0.86013**	**0.08010**	**0.62666**	**0.06855**
		SPA	0.34428	0.08858	−0.00623	0.08988
	LS-SVM	None	0.26481	0.08023	0.15198	0.07105
		CARS	0.01485	0.09287	−0.05593	0.07929
		SPA	0.00526	0.09332	−0.04930	0.07904

(**a**) *Chla* modeling scatter plot

(**b**) *Chlb* modeling scatter plot
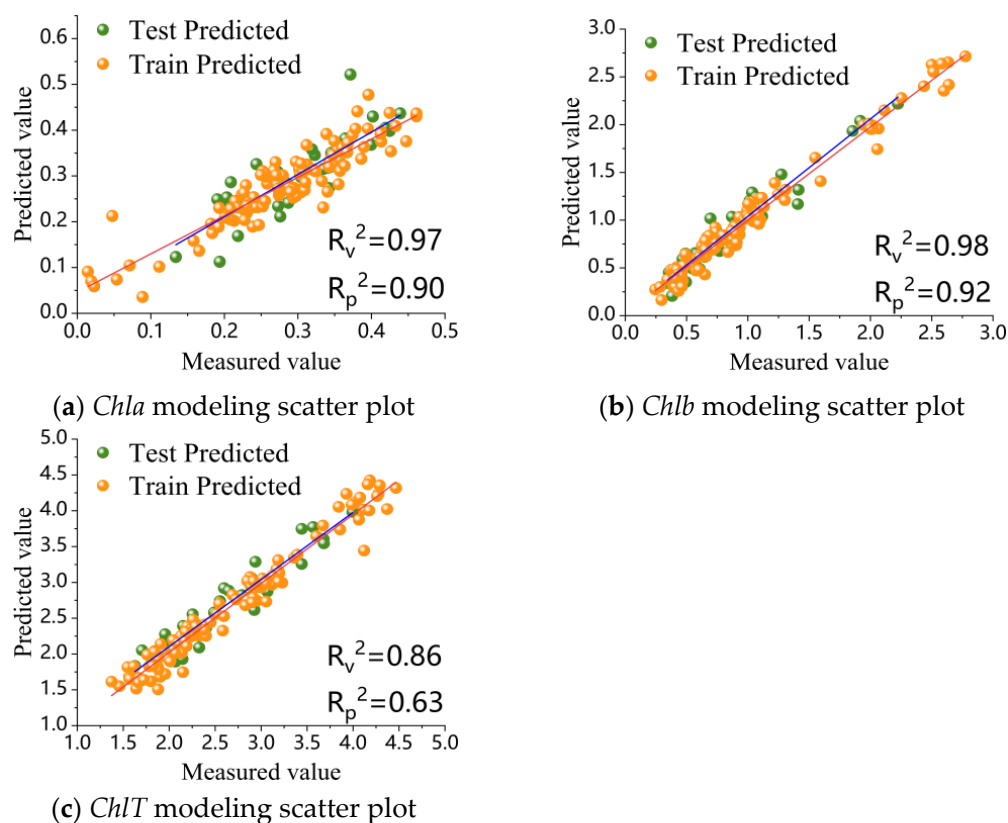

(**c**) *ChlT* modeling scatter plot

Figure 6. Scatter diagram of optimal modeling effect of three pigments.

From a macroscopic perspective, the implementation of feature selection algorithms plays a crucial role in enhancing the predictive capabilities of models, with the choice of modeling technique being a key factor in determining outcomes. Specifically, in pigment modeling, the predictive performance metrics of partial least squares regression (PLSR) models consistently surpass those of least squares Support Vector Machine (LS-SVM). This superiority stems from PLSR's ability to capture the essence of the input and output

variables, effectively reducing data complexity and addressing multicollinearity issues. Beyond prediction, PLSR offers a more comprehensive understanding of the relationships between variables, making it preferable to LS-SVM for explanatory analysis.

In contrast, the predictive performance of models using the Successive Projections Algorithm (SPA) is comparatively suboptimal. A comparative analysis of the spectral bands selected by the CARS and SPA models reveals that SPA tends to identify a more limited set of bands. Moreover, some of the bands selected by SPA lack the robust distributional characteristics seen in those chosen by CARS. This divergence in band selection may explain the lower predictive accuracy observed when SPA is combined with PLSR and LS-SVM modeling frameworks.

As shown in Figure 7, the key spectral bands of *Chla*, *Chlb*, and *ChlT* exhibit distinct patterns. Notably, the SPA selects a more limited set of characteristic wavelengths compared to the CARS algorithm. Specifically, the wavelengths identified by CARS are evenly distributed across the spectral range, providing a comprehensive representation of the pigments' spectral characteristics. In contrast, the SPA's selection is marked by an uneven distribution and a smaller number of bands, which may undermine the model's predictive accuracy. This uneven selection of bands likely contributes to the reduced predictive effectiveness observed when using the SPA to build models.
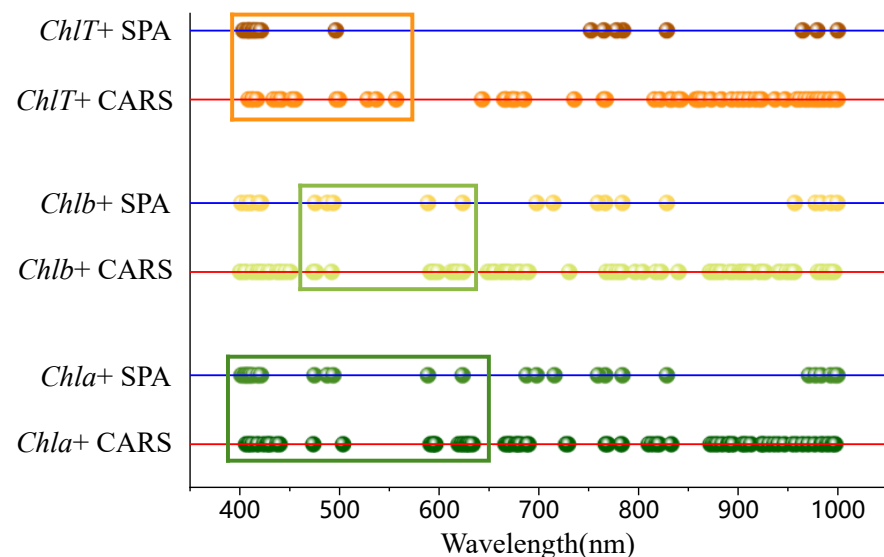

Figure 7. Distribution map of important wavelength for determination of three pigments. The figure shows the overlap of the feature-selected bands with the pigment spectral characteristic intervals.

3.5. Visualization of Tea Pigment

In this study, hyperspectral images from different batches (plucking dates) across three distinct processing procedures were selected to visualize and analyze the distribution of the three major pigments: *Chla*, *Chlb*, and *ChlT*. An optimal prediction model was employed for precise quantitative predictions at each pixel point in the original hyperspectral images. The input parameters for the model were key spectral bands filtered using the CARS feature selection technique. By leveraging these feature bands, the model generated prediction values for each pixel, enabling a comprehensive analysis and the processing of the entire image.

Figure 8 shows the visualization results of the tea samples: different colors indicate different concentrations of pigments, with bluish areas indicating lower concentrations of pigments and yellow areas indicating higher concentrations of pigments. From this, it can be observed that the contents of *Chla*, *Chlb*, and *ChlT* show a gradual increase as processing procedures. By combining HSI with mathematical modeling, the method allows accurate prediction of pigment content during tea processing and real-time monitoring of

the dynamics of pigment distribution through visualization. When integrated into real-world environments, the system allows operators to continuously monitor pigment levels and make dynamic adjustments to processing parameters, thereby enhancing detection efficiency, optimizing the production process, and improving overall quality control in tea production.
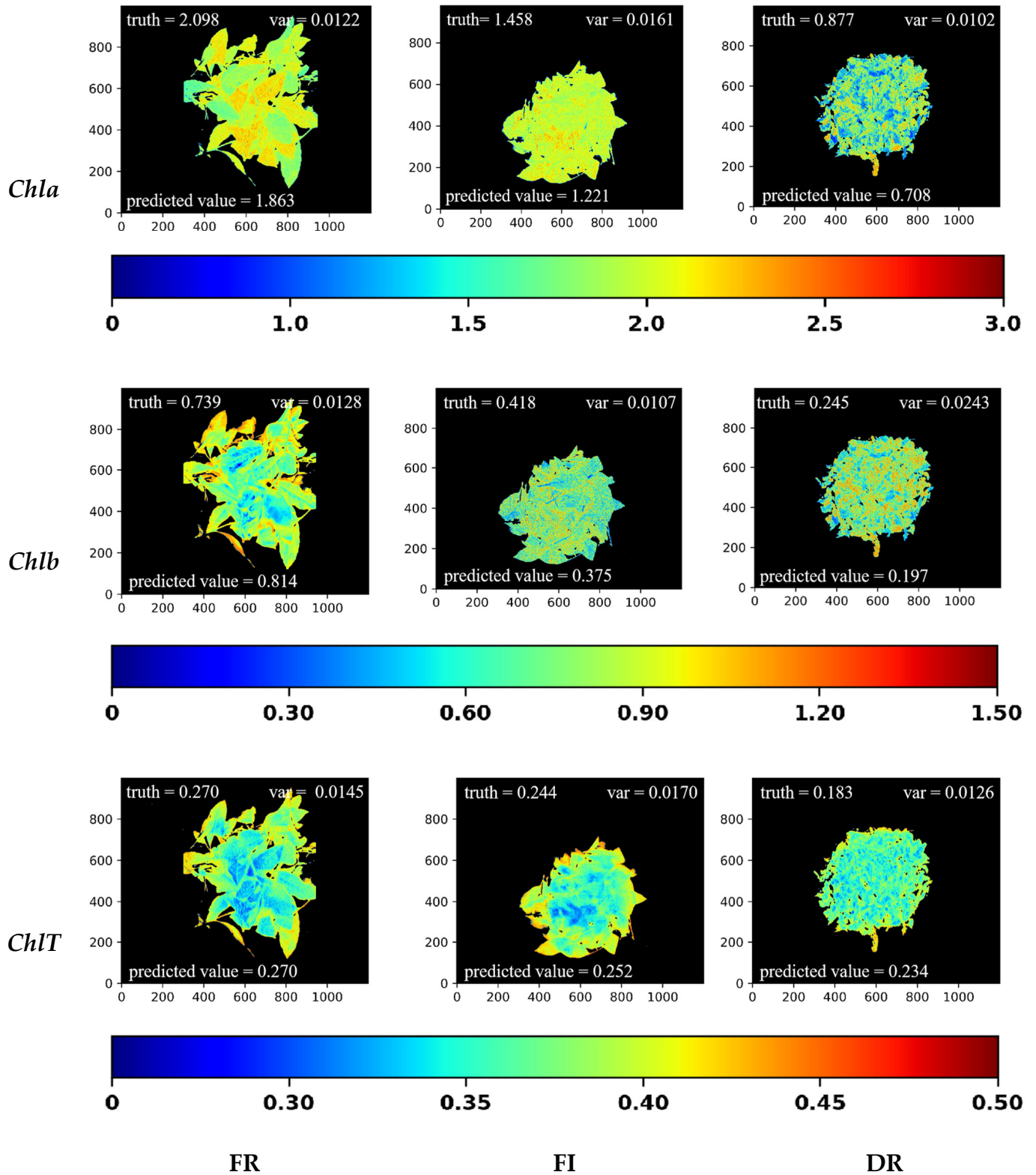

Figure 8. Visualization of pigment distribution in tea during the three processing procedures.

4. Discussion

In this study, we integrated VIS-NIR hyperspectral imaging with machine learning to create a rapid and non-destructive method for detecting *Chla*, *Chlb*, and *ChlT* content in tea. The introduction of CARS and SPA models for feature selection significantly enhanced the accuracy and stability of the prediction models for pigments. Moreover, the optimized prediction model enables precise analysis at each pixel in the hyperspectral image, facilitating real-time dynamic monitoring of pigment content during tea processing. This approach effectively addresses the limitations of traditional spectroscopic techniques, which struggle with pixel-by-pixel analysis and often fail to capture processing variations across different sample regions.

Despite the important results of this study, there are still some limitations and room for improvement.

First, although the model performed well for the prediction of *Chla* and *Chlb*, the prediction accuracy for *ChlT* was relatively low, which indicates that further improvement is still needed in the detection sensitivity of *ChlT* content. The reasons may be as follows:

The sample preprocessing and detection steps were unable to detect the true value of the sample, resulting in the spectral model not being able to match.

In the range of 400–1000 nm, the spectra of *ChlT* may overlap with the absorption spectra of other phytochromes (e.g., chlorophylls and flavonoids), leading to confusing data.

Relatively low R_p^2 values may imply that the model is not complex enough to effectively capture potentially complex features in the *ChlT* data.

To address these three issues, we propose the following:

1. In the future, more accurate testing instruments will be used for measurement to reduce measurement errors and ensure the accuracy of the true value of the samples; optimize the sample preprocessing and testing process, and plan to make multiple measurements and take the average value of each sample in order to reduce the impact of random errors in a single measurement.
2. In future studies, advanced spectral correction techniques will be used to reduce the effect of spectral overlap, and specific spectral features of *ChlT* will be extracted by feature selection methods (e.g., SPA) and spectral decomposition techniques (e.g., ICA). This will help to separate the overlapping spectral signals and extract the most representative spectral regions, thus improving the modeling results.
3. Future studies will consider more sophisticated predictive models such as deep learning frameworks (e.g., CNN or KNN), which have advantages in dealing with complex relationships, and can further improve the performance of *ChlT* content detection.

Second, the extended processing time for visualizing and generating images and reports is another challenge that needs to be addressed. Currently, it takes approximately seven seconds to generate a visualization chart on an average computer. However, in large-scale production lines, particularly in high-speed assembly environments, such delays can reduce productivity. To meet the efficiency demands of production lines, both the algorithms and hardware configurations need to be further optimized to enable data processing and visualization generation to occur within milliseconds, ensuring the practicality and efficiency of this technology in real-world production settings.

In the future, research that integrates HSI with AI technologies is expected to find widespread applications across various fields, particularly in areas requiring high-precision analysis and non-destructive testing, such as tea processing, food quality control, and production monitoring. Future efforts will prioritize reducing computational resource demands and accelerating data processing and visualization. In addition, the development of more cost-effective hardware solutions, such as simplified versions of hyperspectral inspection equipment, will also be a focus of research, and by simplifying operational processes and increasing automation, the reliance on specialized personnel can be reduced and the technology can be made more suitable for mass production environments.

5. Conclusions

In this study, we successfully integrated VIS-NIR hyperspectral imaging with machine learning algorithms to develop a rapid, non-destructive method for detecting *Chla*, *Chlb*, and *ChlT* contents in stacked tea leaves during processing. This approach also enabled spatial visualization of pigment distribution. To reduce data redundancy, feature bands were selected using the CARS and SPA models, and detection models were constructed based on PLSR and LS-SVM. The results demonstrated that PLSR, combined with CARS feature band selection, achieved the highest prediction accuracy, with R_p^2 values of 0.90465, 0.92068, and 0.62666 for *Chla, Chlb*, and *ChlT*, respectively, and RMSEP values of 0.52810, 0.29431, and 0.06855, respectively, highlighting the models' accuracy and stability.

A key innovation of this study is the extension of these quantitative detection models to each pixel of the hyperspectral image, enabling point-by-point prediction of pigment composition. This method successfully visualized the pigment distribution in stacked tea leaves during in situ processing on a production line. It offers a valuable reference for processors, allowing them to adjust processing parameters in real time based on pigment variation, thereby optimizing the color and quality of the tea. The visualization of pigment distribution not only enhances monitoring efficiency during tea processing but also provides a new technical approach to standardized quality control.

In conclusion, this study improves the efficiency of pigment content detection during tea processing, optimizes the production process, and provides strong technical support for processing decisions and standardized tea production. It lays the groundwork for the real-time adjustment of processing parameters, optimization of tea product quality, and the realization of intelligent tea processing based on the in situ physical and chemical properties of raw materials.

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

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available at https://github.com/Octoder-six-one/Spectrum%E2%80%94Pigment.git, accessed on 3 October 2024.

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

References

1. Zhang, Y.; Yan, K.; Peng, Q.; Baldermann, S.; Zhu, Y.; Dai, W.; Feng, S.; Simal-Gandara, J.; Fu, J.; Lv, H.; et al. Comprehensive Analysis of Pigment Alterations and Associated Flavor Development in Strip and Needle Green Teas. *Food Res. Int.* **2024**, *175*, 113713. [CrossRef] [PubMed]
2. Yang, C.S.; Wang, H.; Sheridan, Z.P. Studies on Prevention of Obesity, Metabolic Syndrome, Diabetes, Cardiovascular Diseases and Cancer by Tea. *J. Food Drug Anal.* **2018**, *26*, 1–13. [CrossRef] [PubMed]
3. Habtemariam, S. *The Chemical and Pharmacological Basis of Tea* (Camellia sinensis *(L.) Kuntze) as Potential Therapy for Type 2 Diabetes and Metabolic Syndrome*; Academic Press: Cambridge, MA, USA, 2019; pp. 839–906.
4. Qi, L.; Jian, O.; Liu, C.; Chen, H.; Li, J.; Xiong, L.; Liu, Z.; Huang, J. Research progress of tea quality evaluation technology. *Tea Sci.* **2022**, *42*, 316–330.

5.   Dan, Z.; Yin, X.; Gu, H.; Long, W.; Fu, H.; She, Y. Research progress of quantitative evaluation methods for tea grades. *Tea Sci.* **2023**, *43*, 733–746.

6.   Guo, L.; Lai, L.; Qu, Y.; Lin, Z.; Guo, Y. Research progress of "three green" characteristic components of green tea. *Food Ferment. Ind.* **2016**, *42*, 285–290.

7.   Na, S.; Bei, W.; Ke, O.; Qing, M.; Tong, H. Changes of major fat-soluble pigments in green tea processing and their effects on the colour and quality of tea leaves_Shu Na. *Food Ferment. Ind.* **2021**, *47*, 170–177.

8.   Suzuki, Y.; Shioi, Y. Identification of Chlorophylls and Carotenoids in Major Teas by High-Performance Liquid Chromatography with Photodiode Array Detection. *J. Agric. Food Chem.* **2003**, *51*, 5307–5314. [CrossRef]

9.   Xu, B. Analysis of Fat-Soluble Pigments in Green Tea and the Mechanism of Colour Formation of New Gaoqiao Yinfeng Shape. Master's Thesis, Hunan Agricultural University, Changsha, China, 2007.

10. Shi, J.; Wang, J.; Peng, Q. Effects of pre-harvest induction and post-harvest processing on carotenoids and their derived aroma in fresh tea leaves. *Chin. J. Food* **2023**, *23*, 365–376.

11. Wei, Y.; Li, X.; He, Y. Generalisation of Tea Moisture Content Models Based on Vnir Spectra Subjected to Fractional Differential Treatment. *Biosyst. Eng.* **2021**, *205*, 174–186. [CrossRef]

12. Ren, G.; Ning, J.; Zhang, Z. Multi-variable Selection Strategy Based on Near-infrared Spectra for the Rapid Description of Dianhong Black Tea Quality. *Mol. Biomol. Spectrosc.* **2021**, *245*, 118918. [CrossRef]

13. Ouyang, Q.; Liu, Y.; Chen, Q.; Zhang, Z.; Zhao, J.; Guo, Z.; Gu, H. Intelligent Evaluation of Color Sensory Quality of Black Tea By Visible-near Infrared Spectroscopy Technology: A Comparison of Spectra and Color Data Information. *Mol. Biomol. Spectrosc.* **2017**, *180*, 91–96. [CrossRef] [PubMed]

14. Ren, G.; Liu, Y.; Ning, J.; Zhang, Z. Assessing Black Tea Quality Based on Visible-near Infrared Spectra and Kernel-based Methods. *J. Food Compos. Anal.* **2021**, *98*, 103810. [CrossRef]

15. Li, J.; Ma, J.; Zhang, Y.; Zheng, L. Determination of 19 polyphenolic compounds in tea by ultra-high performance liquid chromatography combined with quadrupole-time of flight mass spectrometry. *Food Sci. Hum. Wellness* **2022**, *11*, 719–726. [CrossRef]

16. Zou, X.; Chen, Z.; Shi, J.; Huang, X.; Zhang, D. Rapid detection of pigment content in cucumber leaves based on near-infrared hyperspectral images. *Trans. Chin. Soc. Agric. Mach.* **2012**, *43*, 152–156.

17. Li, S.; Zhang, H.; Huai, J.; Wang, H.; Li, S.; Zhuang, L.; Zhang, J. An Online Preparative High-performance Liquid Chromatography System with Enrichment and Purification Modes for the Efficient and Systematic Separation of Panax Notoginseng Saponins. *J. Chromatogr. A* **2023**, *1709*, 464378. [CrossRef]

18. Feng, M.; Xie, J.; Mu, X. Tea leaf residue as an example to explore the extraction process of colour pigments from food waste. *Guangzhou Chem. Ind.* **2022**, *50*, 55–57, 63.

19. Wang, X.; Zhang, D.; Liu, X.; Hudie, Z.; Hong, G.; Yi, F. Optimisation of microwave-assisted extraction of black tea pigments. *Wool Text. Sci. Technol.* **2023**, *51*, 38–44.

20. Wang, F. Hyperspectral Characteristics of Chlorophyll and Total Nitrogen Content in Oil Tea and Rice in Red Soil Region. Ph.D. Thesis, Jiangxi Agricultural University, Nanchang, China, 2022.

21. Shen, Y.; Wu, P.; Huang, F.; Guo, C. Red tide algae species identification and concentration measurement based on hyperspectral imaging. *Spectrosc. Spectr. Anal.* **2023**, *43*, 3629–3636.

22. Zhang, Z. Estimation of Anthocyanin Content in Apple Leaves and Identification of Mosaic Disease Based on Hyperspectral Data. Master's Thesis, Northwest Agriculture and Forestry University, Xianyang, China, 2023.

23. Jin, L. Inversion of Chlorophyll A and B Contents in Tea Canopy Based on Hyperspectral Remote Sensing. Master's Thesis, Hangzhou Dianzi University, Hangzhou, China, 2023.

24. Li, Q.; Huang, Y.; Song, X.; Zhang, J.; Min, S. Moving Window Smoothing on the Ensemble of Competitive Adaptive Reweighted Sampling Algorithm. *Mol. Biomol. Spectrosc.* **2019**, *214*, 129–138. [CrossRef]

25. de Araújo Gomes, A.; Galvão, R.K.H.; de Araújo, M.C.U.; Véras, G.; da Silva, E.C. The Successive Projections Algorithm for Interval Selection in PLS. *Microchem. J.* **2013**, *110*, 202–208. [CrossRef]

26. Henseler, J.; Ringle, C.M.; Sarstedt, M. A New Criterion for Assessing Discriminant Validity in Variance-based Structural Equation Modeling. *J. Acad. Mark. Sci.* **2014**, *43*, 115–135. [CrossRef]

27. Suykens, J.A.; De Brabanter, J.; Lukas, L.; Vandewalle, J. Weighted Least Squares Support Vector Machines: Robustness and Sparse Approximation. *Neurocomputing* **2002**, *48*, 85–105. [CrossRef]

28. Fu, X.; Tang, J.; Yang, Z. Progress of carotenoid synthesis and metabolic regulation in tea. *Guangdong Agric. Sci.* **2021**, *48*, 18–27.

29. Liu, Y.; Liu, X.; Lin, G.; Guo, C.; Deng, L.; Yang, C. Research on new process of low-temperature vacuum and hot air combined drying of green tea. *Tea Sci.* **2013**, *33*, 345–350.

30. Lichtenthaler, H. Chlorophylls and carotenoids: Pigments of photosynthetic biomembranes. In *Methods in Enzymology*; Academic Press: Cambridge, MA, USA, 1987; pp. 350–382.

31. Goodwin, T.W. *The Biochemistry of the Carotenoids*; Springer: Dordrecht, the Netherlands, 1980; pp. 346–349.

32. Blackburn, G.A. Hyperspectral Remote Sensing of Plant Pigments. *J. Exp. Bot.* **2006**, *58*, 855–867. [CrossRef]
33. Rasti, B.; Scheunders, P.; Ghamisi, P.; Licciardi, G.; Chanussot, J. Noise Reduction in Hyperspectral Imagery: Overview and Application. *Remote Sens.* **2018**, *10*, 482. [CrossRef]