MDPI

Establishment and Accuracy Evaluation of Cotton Leaf Chlorophyll Content Prediction Model Combined with Hyperspectral Image and Feature Variable Selection

Siyao Yu [1,2,3], Haoran Bu [1,2] (ID), Xue Hu [1,2,*], Wancheng Dong [1,2,3] and Lixin Zhang [1,4]

1 College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832000, China; yusiyao960526@163.com (S.Y.); bhr9763@163.com (H.B.); dongwancheng1982@126.com (W.D.); zhlx2001329@163.com (L.Z.)
2 Engineering Research Center for Production Mechanization of Oasis Characteristic Cash Crop, Ministry of Education, Shihezi 832000, China
3 Collaborative Innovation Center of Province-Ministry Co-Construction for Cotton Modernization Production Technology, Shihezi 832000, China
4 Bingtuan Energy Development Institute, Shihezi University, Shihezi 832000, China
* Correspondence: huxue@foxmail.com; Tel.: +86-188-9953-3580

Abstract: In order to explore the feasibility of rapid non-destructive detection of cotton leaf chlorophyll content during the growth stage, this study utilized hyperspectral technology combined with a feature variable selection method to conduct quantitative detection research. Through correlation spectroscopy (COS), a total of 882 representative samples from the seedling stage, bud stage, and flowering and boll stage were used for feature wavelength screening, resulting in 213 selected feature wavelengths. Based on all wavelengths and selected feature wavelengths, a backpropagation neural network (BPNN), a backpropagation neural network optimized by genetic algorithm (GA-BPNN), a backpropagation neural network optimized by particle swarm optimization (PSO-BPNN), and a backpropagation neural network optimized by sparrow search algorithm (SSA-BPNN) prediction models were established for cotton leaf chlorophyll content, and model performance comparisons were conducted. The research results indicate that the GA-BPNN, PSO-BPNN, and SSA-BPNN models established based on all wavelengths and selected feature wavelengths outperform the BPNN model in terms of performance. Among them, the SSA-BPNN model (referred to as COS-SSA-BPNN model) established using 213 feature wavelengths extracted through correlation analysis showed the best performance. Its determination coefficient and root-mean-square error for the prediction set were 0.920 and 3.26% respectively, with a relative analysis error of 3.524. In addition, the innovative introduction of orthogonal experiments validated the performance of the model, and the results indicated that the optimal solution for achieving the best model performance was the SSA-BPNN model built with 213 feature wavelengths extracted using the COS method. These findings indicate that the combination of hyperspectral data with the COS-SSA-BPNN model can effectively achieve quantitative detection of cotton leaf chlorophyll content. The results of this study provide technical support and reference for the development of low-cost cotton leaf chlorophyll content detection systems.

Keywords: chlorophyll; cotton; hyperspectral; improved neural networks; non-destructive testing; orthogonal experiment

1. Introduction

Chlorophyll is a fundamental component in plant organs, and its content is an important physicochemical parameter that reflects crop growth. Accurate and efficient quantitative estimation of cotton leaf chlorophyll content (*CLCC*) is of great significance for yield prediction and field management decision making [1–3]. Traditional chlorophyll

content detection usually involves field sampling and indoor testing, which is not only time-consuming and laborious, but also destructive and lagging [4]. The use of hyperspectral technology for the determination of plant physicochemical parameters, such as chlorophyll, has gradually become an important tool for evaluating crop physicochemical parameters, due to its advantages of low consumption and rapid and non-damaging detection, among others [5–8].

Over time, neural networks have gained significant popularity in spectral qualitative analysis and quantitative prediction due to their advantages in learning, fault tolerance, real-time processing, and fitting non-linear problems [9–14]. A backpropagation neural network (BPNN), as one of the representative algorithms in machine learning, is a multilayer forward neural network that utilizes a backpropagation learning algorithm and has shown good performance in non-linear pattern recognition and classification [15–17]. Combining the color values (R, G, B, H, S, and I) of grape skin, using BPNN to predict grape ripeness has proven to be a great method for predicting grape ripeness [18]. Some scholars have constructed winter wheat chlorophyll retrieval models based on BPNN and regression analysis and compared actual measured values with model estimated values. The results showed that the inversion model based on BPNN demonstrated significantly higher accuracy than the regression analysis model [19]. Some scholars have also used partial least squares regression, principal component regression, and BPNN to establish models for estimating chlorophyll content in corn leaves, and the results also showed that the BPNN network model had the best prediction effect [20]. Although the BPNN model can achieve good detection performance, it still has some limitations, such as slow convergence speed, susceptibility to local optima, and overfitting problems, as mentioned in previous studies [21–23]. To address the limitations of the BPNN model, Li et al. [24] optimized the BPNN model using a genetic algorithm to establish an ecosystem health assessment model for 16 regions in Yunnan Province. Furthermore, it has been demonstrated that the optimized model based on high-spectral data for predicting the gelatinization characteristics of millet using the backpropagation neural network optimized by particle swarm optimization (PSO-BPNN) approach exhibits higher expressive capacity than the BPNN model [25].

While there has been a substantial amount of research on using machine learning methods for crop nutrient and related parameter detection using hyperspectral technology, the studies have primarily focused on rice [26,27], maize [20,28,29], and wheat [5,6,22]. There is relatively less literature available on cotton as the subject of study.

In order to investigate the impact of spectral band selection and modeling methods on the quantitative prediction of chlorophyll content, this study utilized hyperspectral data preprocessing using the Savitzky–Golay five-point quadratic smoothing method. Feature wavelength selection was performed using correlation spectroscopy (COS). DH10 cotton plants at the seedling, bud, and flowering stages were quantitatively assessed for mixed-leaf chlorophyll content using BPNN, backpropagation neural network optimized by genetic algorithm (GA-BPNN), PSO-BPNN, and backpropagation neural network optimized by sparrow search algorithm (SSA-BPNN). By collecting hyperspectral imaging information of cotton leaves at different growth stages using hyperspectral instruments under laboratory conditions, representative spectral data are obtained to establish a quantitative relationship between spectra and chlorophyll content. Comparing the performance of BPNN, GA-BPNN, PSO-BPNN, and SSA-BPNN models, the optimal detection model for cotton leaf chlorophyll content is selected, and its feasibility is further explored using orthogonal experiments. The rapid detection model for chlorophyll content in DH10 cotton established in this study provides a reference for the detection of chlorophyll content in other cotton varieties. This study innovatively applied orthogonal experiments in quantitative detection research, providing a new perspective on enhancing model reliability and improving modeling efficiency through the combination selection of models using orthogonal experiments. At the same time, it offers the corresponding technical support and theoretical basis for the development of low-cost cotton chlorophyll content rapid detection systems.

2. Materials and Methods

2.1. Sampling Site

In this paper, samples were collected from the second company experimental base of Shihezi University in Shihezi, Xinjiang Uygur Autonomous Region (86.08° E, 44.31° N), which is located in a temperate continental climate with large temperature differences and sufficient sunshine hours (annual sunshine hours reach 2500–3500 h), and the sampling area is shown in Figure 1.
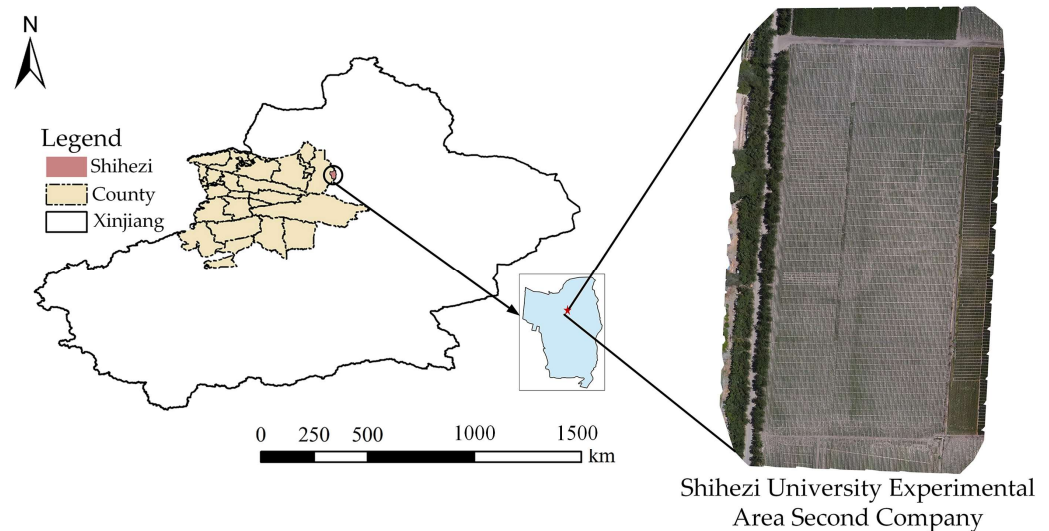

Figure 1. Schematic diagram of the geographical location of the sampling site.

2.2. Data—Acquisition and Pre-Processing

2.2.1. Field Sample Collection

The study was conducted on DH10-type cotton with a planting area of 18.84 m × 40 m (Figure 1), and cotton leaf samples were collected in the field at three time points: June 13 (seedling stage), July 10 (bud stage), and August 5 (flowering and boll stage).

Based on field surveys and relevant literature, a combination of "five-point sampling" and "random sampling" methods was utilized for selecting field cotton plants. Cotton plants were randomly selected and labeled at sampling points. Starting from the top leaves of each cotton plant, the third main leaf of the third branch was plucked. This position typically exhibits good development and represents the sample well. After labeling, the leaves were sealed in bags and stored in a portable refrigeration unit to preserve the samples. After excluding samples that were damaged due to improper storage, a total of 882 samples were obtained, with 259, 308, and 315 samples collected during the seedling stage, bud stage, and flowering and boll stage, respectively.

2.2.2. Hyperspectral Image Acquisition

Hyperspectral images of cotton leaves were acquired in the laboratory using a hyperspectral imaging system (ISUZU OPTICS Co., Ltd., Suzhou, China). The hyperspectral imaging system (Figure 2) mainly consists of an imaging spectrometer, a 150 W light source providing parallel light, a precision delivery unit set (Zhuo Li Hanguang, SC300-1A, Beijing, China), a 14-bit thermoelectrically cooled electron-multiplying charge-coupled device (EM-CCD), and a camera (Andor Luca EMCCD DL-604M, Andor Technology plc., N. Ireland). The spectral range and the respective rates were 400–1000 nm and 2.8 nm. The number of wavelengths was 846. The system uses line scan to acquire hyperspectral information of the sample. To eliminate baseline drift, the light source and camera are turned on and preheated for 30 min before hyperspectral image data acquisition. The parameters of the hyperspectral image acquisition system were set as follows: the angle between the light source and the vertical plane was 45°, the exposure time $T = 0.016$ s, the distance between

the sample and the lens was 28 cm, and the image acquisition speed $V = 1.35$ mm/s. During the test, the blade was placed at the center of the carrier table.
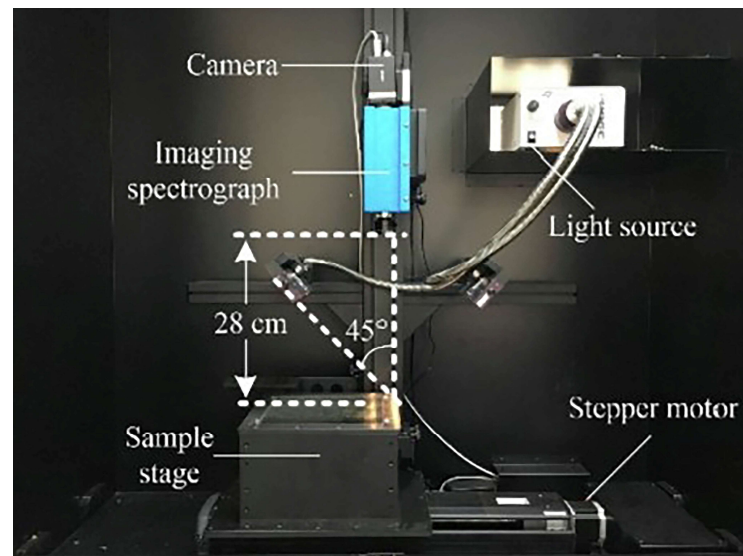

Figure 2. Schematic diagram of hyperspectral image data acquisition system composition.

2.2.3. Hyperspectral Image Correction

Due to the spatial light intensity conversion in the halogen lamp and the dark current in the CCD camera that may affect spectra with low reflectance, black–white correction of the instrument and black–white calibration of the hyperspectral image are required before collecting hyperspectral data [30–32]. Under the same system conditions as the sample collection, a white calibration image W was obtained by scanning a white calibration board with a diffuse reflection efficiency of 99%, and a black calibration image B was obtained by closing the camera shutter. This completes the calibration of the hyperspectral image. The collected absolute image I is transformed into a relative image E using the following formula:

$$E = \frac{I - B}{W - B} \tag{1}$$

2.2.4. Hyperspectral Information Extraction

The corrected sample images were analyzed using image segmentation techniques to select the regions of interest (ROIs) for each sample and extract representative spectral information from them [33]. The representative sample's original image is shown in Figure 3a. Due to the distinct color contrast between the leaf portion and the main stem portion in the hyperspectral image of the leaf, a support vector machine (SVM) was utilized to select the RGB values of pixels as features for image segmentation. The segmented sample image is shown in Figure 3b, which serves as the sample region of interest (ROI) (Figure 3c). Original pixel dimensions for the spectral image are 83,776 × 80,304, but the ROI is the extracted complete leaf area after image segmentation using SVM for each leaf. The size of each leaf's pixels is not fixed. The average spectrum of all pixels within the ROI is extracted as the representative spectral information of the sample (Figure 3d).

From Figure 3d, it can be observed that within the visible light wavelength range, the 500–600 nm region represents a high reflectance area, with a peak appearing around 550 nm. The 400–500 nm and 600–700 nm regions show low reflectance. Reflectance shows a steep increasing trend from 700 to 760 nm. In the near-infrared wavelength range, 760–1000 nm represents a strong reflectance region, and the curve appears almost horizontal.
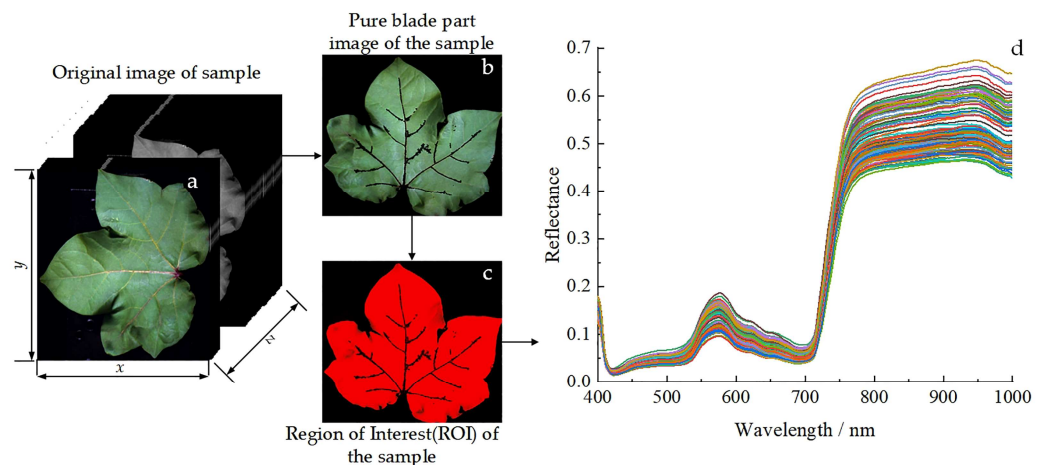
Figure 3. Hyperspectral images of cotton leaves and average spectra of ROI in three stages. (**a**). Original image sample; (**b**). Pure blade part image of the sample; (**c**). Region of Interest (ROI) of the sample; (**d**). The average spectrum of all pixels: Different colored lines represent different sample spectral curves.

2.2.5. Hyperspectral Data Processing

Due to the limitations of the instrument itself, it may introduce some unfavorable factors, such as noise and dark current. Additionally, it is also influenced to some extent by its own non-quality-related information. For example, phenomena such as baseline drift in the spectral curve, multicollinearity, and noise issues contribute to the presence of redundant information in this data. Redundant information not only affects the response time of the model, but can also potentially impact its performance. Therefore, in order to maintain the integrity of the image and avoid the influence of these unfavorable factors on the acquired sample spectral curves, it is necessary to process the raw spectral information. This study utilized the Savitzky–Golay five-point quadratic smoothing method for preprocessing the spectral data. Building upon this method, the correlation between spectral parameters and cotton leaf chlorophyll content was investigated, leading to the selection of characteristic wavelength bands.

2.2.6. CLCC Determination

After completing the hyperspectral data collection for all samples, *CLCC* was measured using the spectrophotometric method. A total of 0.5 g of cotton leaves was taken, the veins were removed, and the leaves were crushed and placed in a mortar. Quartz sand and calcium carbonate powder were added with an 80% acetone solution with a volume fraction of 2–3 mL, and ground until the tissue turned white. Then, 10 mL of acetone solution was added and ground into a uniform pulp and left to stand in the dark at room temperature (25 °C) for 10 min. After filtration, the mortar and pestle were repeatedly rinsed to ensure all leaf pigments entered the volumetric flask. Finally, the solution was made up to 50 mL using a 95% ethanol solution, and the total mass concentration of chlorophyll in the extract (mg/L) was measured using the TPX04 nutrient detector at an absorbance of 652 nm [34], which is calculated as follows:

$$C_{a+b} = \frac{A_{652} \times 1000}{34.5} \tag{2}$$

where 34.5 is the absorbance coefficient of chlorophyll *a* and *b* at a wavelength of 652 nm. In turn, *CLCC* (mg/g) was measured as

$$CLCC = \frac{C_{a+b} \times V}{M \times 1000} \tag{3}$$

In Equation (2), C_{a+b} represents the total mass concentration of chlorophyll (mg/L); V denotes the total volume of the extraction solution (mL); M refers to the fresh mass of the leaf sample (g). The statistical data of chlorophyll content for a total of 882 cotton leaf samples at the seedling stage, bud stage, and boll stage of DH10 cotton are presented in Table 1.

Table 1. Chlorophyll leaf content of cotton samples during the growth period.

$CLCC/(\mathrm{mg \cdot g^{-1}})$	DH10 Cotton Leaf Samples during the Growth Period
Maximum	1.74
Median	1.28
Minimum	1.03
Average	1.36

The distribution of cotton chlorophyll content is illustrated in Figure 4. The average chlorophyll content gradually converges towards the median, indicating a normal distribution of the content. These data distribution characteristics of chlorophyll content may be beneficial for training the *CLCC* prediction model.
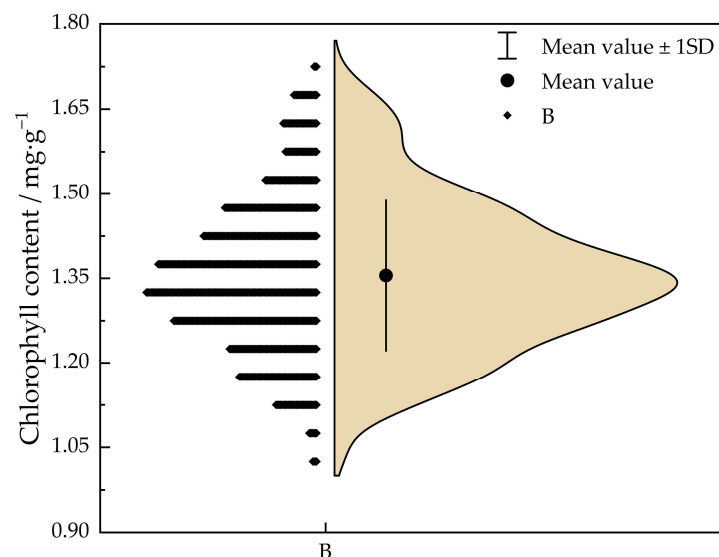

Figure 4. Schematic diagram of chlorophyll content distribution.

2.3. Model Construction and Accuracy Evaluation Standards

2.3.1. Model Construction

In this study, BPNN, GA-BPNN, PSO-BPNN, and SSA-BPNN were used to construct cotton leaf chlorophyll content prediction models. A total of 882 samples from the three periods were mixed, and based on a random selection principle, 93% of the samples were used as training samples for model building, while the remaining 7% were used for prediction.

The BPNN algorithm is a forward-propagation, backpropagation algorithm. During the forward-propagation process, input samples pass through the input layer, hidden layer, and finally reach the output layer. When there is a significant error between the output and the actual results, the backward propagation process is initiated. During the backpropagation process, the error signal is propagated back along the original path of connection, and the weights and thresholds of neurons in each layer are modified to reduce the error [9–17]. The above forward and backward propagations are repeated until the requirements are met, completing the training of the network model. In this study, when modeling based on the BPNN algorithm, the main parameter settings were as follows: the

number of nodes in the input and output layers was 1; the number of nodes in the hidden layer was 9; and the iteration count, learning rate, and target were set as 200, 0.01, and 10^{-6}, respectively.

The GA is a parallel, random search optimization method that was proposed in 1962 by Professor Holland from the University of Michigan, USA [35]. It is derived from simulating the genetic mechanisms of the natural world and the theory of biological evolution. It incorporates the principles of natural selection and survival of the fittest into the encoded population formed for parameter optimization. It selects, crosses over, and mutates individuals based on the chosen fitness function, ensuring that individuals with higher fitness values are preserved, while those with lower fitness values are eliminated. The new population inherits information from the previous generation, while also being superior to it. This process continues in a repetitive cycle until the conditions are met. The basic elements of a genetic algorithm include chromosome encoding methods, fitness function, genetic operations, and running parameters. It possesses characteristics such as high-level heuristic search and parallel computing. When using the GA algorithm to optimize the BPNN model, the main parameter settings are as follows: 50 iterations, a population size of 10, a crossover probability of 0.4, and a mutation probability of 0.2.

The PSO algorithm is a population-based intelligent optimization algorithm. It is inspired by the collective behavior of biological populations and applied to solve optimization problems. Each particle in the algorithm represents a potential solution to the problem, and each particle corresponds to a fitness value determined by the fitness function. The velocity of a particle determines the direction and distance of its movement. The velocity is dynamically adjusted based on the particle's own movement experience and that of other particles, allowing individuals to search for optimization within the feasible solution space. When using the PSO algorithm to optimize the BPNN model, the main parameter settings are as follows: the acceleration coefficient is set to 1.494, the population size is 20, each particle has a dimension of 2, the population is updated 100 times, and the velocity of the particles is set between −1.0 and 1.0.

The SSA is a novel swarm intelligence optimization algorithm introduced in 2020 [36]. It is primarily inspired by the foraging behavior and anti-predator behavior of sparrows: individuals in the population monitor the behavior of other individuals in the group. Attackers within the population compete with high-intake companions for food resources to enhance their predation rate. Additionally, when the sparrow population becomes aware of danger, they exhibit anti-predator behavior. When using the SSA algorithm to optimize the BPNN model, the main parameter settings are as follows: the safety value is set to 0.6, the proportion of discoverers in the population is 0.7, and the rest are joiners. The proportion of sparrows sensing danger is 0.2. The initial population size is 30, each particle has a dimension of 2, and the population is updated 50 times.

2.3.2. Model Accuracy Evaluation Criteria

The accuracy evaluation parameters of the *CLCC* prediction model established in this article include the root-mean-square error of calibration (*RMSEC*), the root-mean-square error of prediction (*RMSEP*), the coefficient of determination on the training set (R_c^2), the coefficient of determination on the prediction set (R^2), and the relative prediction deviation (*RPD*). When the R^2 value is higher and the *RMSEP* value is lower, the regression effect of the model is better. The formula for calculating *RPD* is as follows:

$$RPD = \sqrt{\frac{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2}{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}} \tag{4}$$

where y_i is the actual value of the *i*th sample, $\hat{y}_i$ is the predicted value of the *i*th sample, $\overline{y}$ is the actual mean value, and *n* is the number of samples. When *RPD* > 2.0, the model is

considered to be good at prediction. When $1.4 < RPD \leq 2.0$, then the model can make a rough prediction of chlorophyll content, but the prediction accuracy needs to be improved. When $RPD \leq 1.4$, the model is considered to have poor accuracy and does not have prediction ability.

3. Results

3.1. Feature Wavelength Screening

The correlation coefficient curve between smooth spectral reflectance and cotton chlorophyll content within the wavelength range of 400–1000 nm is shown in Figure 5. The results indicate a negative correlation in the wavelength range of 412–424 nm, with a prominent dip in the correlation coefficient curve occurring around 416 nm. The correlation coefficient at the bottom of the dip is −0.25. There is a positive correlation in the wavelength ranges of 400–411 nm and 422–1000 nm. The correlation coefficient reaches its maximum at 900 nm, with a value of 0.44. Using a significance test with a threshold of $p = 0.01$, a total of 213 feature wavelengths were found to exhibit highly significant positive correlations in the ranges of 404–406 nm, 522–667 nm, and 697–1000 nm.
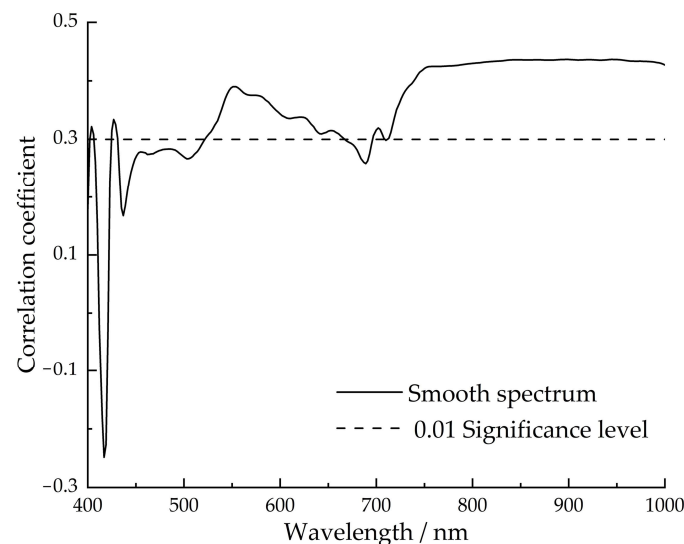

Figure 5. The correlation coefficient between chlorophyll content and smooth spectral wavelength.

3.2. Evaluation of CLCC Prediction Model during the Growth Period

3.2.1. Results and Analysis of BPNN Model

A BPNN model was established using 213 selected feature wavelengths obtained through Savitzky–Golay quadratic smoothing method applied to all wavelengths and filtered based on correlation analysis. The model was used to predict *CLCC*. Based on the model's performance, an optimal model suitable for chlorophyll content detection in cotton was derived. The model evaluation results are shown in Table 2. From Table 2, it can be observed that compared to using all wavelengths, the BPNN model established using feature wavelengths has fewer input variables and shows improved performance. Among them, the number of feature wavelengths selected based on correlation analysis is 213, accounting for 25.18% of the total number of wavelengths. The R_c^2 of the training set and the R^2 of the prediction set for the constructed BPNN model both increased by 9.40% and 6.60%, respectively. Additionally, the *RPD* increased from 1.285 to 1.443, indicating an improvement in the predictive performance of the model. It indicates that utilizing the hyperspectral data in conjunction with the COS-BPNN model can effectively achieve quantitative detection of cotton leaf chlorophyll content.

Table 2. BPNN model evaluation results of cotton leaf samples with different quantities and wavelengths.

Feature Extraction Methods	Number of Wavelengths	Calibration Set		Validation Set		RPD
		RMSEC/%	R_c^2	RMSEP/%	R^2	
Original Spectrum	846	2.91	0.611	2.84	0.655	1.285
COS	213	3.20	0.705	3.05	0.721	1.443

3.2.2. Results and Analysis of the GA-BPNN Model

The evaluation results of the GA-BPNN model for *CLCC*, established using all wavelengths and feature wavelengths, are shown in Table 3. The *RPD* values of the GA-BPNN models for *CLCC*, based on different numbers of wavelengths, are all greater than 1.4 and close to 2.0. This indicates that the model's predictive performance is improved compared to the BPNN model. Among them, the model built using the 213 feature wavelengths selected by the COS method has superior performance. The calibration set's R_c^2 and *RMSEC* are 0.790 and 2.60%, respectively, while the prediction set's R^2 and *RMSEP* are 0.814 and 2.58%, respectively. The GA-BPNN model built showed an improvement of 5.10% and 6.70% in the calibration set's R_c^2 and the prediction set's R^2, respectively. The *RPD* increased from 1.798 to 2.188, indicating an enhancement of the model's predictive performance. This indicates that using the hyperspectral combined COS-GA-BPNN model can effectively achieve quantitative detection of *CLCC*.

Table 3. GA-BPNN model evaluation results of cotton leaf samples with different quantities and wavelengths.

Feature Extraction Methods	Number of Wavelengths	Calibration Set		Validation Set		RPD	Prediction Time/s
		RMSEC/%	R_c^2	RMSEP/%	R^2		
Original Spectrum	846	2.67	0.739	2.64	0.747	1.798	297.63
COS	213	2.60	0.790	2.58	0.814	2.188	169.52

To validate the efficiency of the model, the prediction time of the GA-BPNN model was also statistically analyzed (Table 3). As the number of feature wavelengths decreases, the model's prediction time shortens. The running time of the GA-BPNN model built using the feature wavelengths selected by the COS method is 56.96% of the model built using all wavelengths, indicating a significant improvement in prediction model efficiency.

3.2.3. Results and Analysis of the PSO-BPNN Model

The evaluation results of the PSO-BPNN model for cotton chlorophyll content, established using all wavelengths and feature wavelengths, are shown in Table 4. The *RPD* values of the PSO-BPNN models for cotton leaf chlorophyll content, constructed using different numbers of wavelengths, are all greater than 2.0, indicating the excellent predictive performance of the models. Among them, the PSO-BPNN model established based on the full spectral range has calibration set R_c^2 and *RMSEC* values of 0.804 and 2.25%, respectively. The prediction set R^2 and *RMSEP* values are 0.820 and 2.13%, respectively, with an *RPD* of 2.432. The PSO-BPNN model established based on 213 feature wavelengths selected by the COS method has calibration set R_c^2 and RMSEC values of 0.882 and 2.28%, respectively. The prediction set R^2 and *RMSEP* values are 0.885 and 2.58%, respectively, with an RPD of 2.784. Compared to the PSO-BPNN model, the COS-PSO-BPNN model shows significant improvement in predictive performance. The calibration set R_c^2 and prediction set R^2 have increased by 7.80% and 6.50% respectively. Additionally, the *RPD* value has increased from 2.432 to 2.784, indicating a substantial enhancement of the model predictive ability. As the number of feature wavelengths decreases, the prediction time of the model is reduced. The runtime of the PSO-BPNN model established using feature wavelengths selected by the COS method is 48.09% of the model built using all wavelengths. Compared to the

model built using all wavelengths, there is a significant improvement in the efficiency of the predictive model.

Table 4. PSO-BPNN model evaluation results of cotton leaf samples with different quantities and wavelengths.

Feature Extraction Methods	Number of Wavelengths	Calibration Set		Validation Set		RPD	Prediction Time/s
		RMSEC/%	R_c^2	*RMSEP*/%	R^2		
Original Spectrum	**846**	**2.25**	0.804	2.13	0.820	2.432	226.51
COS	213	2.28	0.882	2.58	0.885	2.784	108.92

3.2.4. Results and Analysis of the SSA-BPNN Model

The evaluation results of the SSA-BPNN model, using all wavelengths and selected feature wavelengths, are presented in Table 5. All models exhibited *RPD* values greater than 3.0 and determination coefficients greater than 0.9. Based on different numbers of wavelengths, the performance of the cotton chlorophyll content SSA-BPNN model was significantly superior to the BPNN model. Among them, the SSA-BPNN model established based on the full wavelength range showed a calibration set R_c^2 value of 0.914 and an *RMSEC* value of 4.08%, while the prediction set had an R^2 value of 0.909 and an *RMSEP* value of 3.62%. The *RPD* value was 3.233. The SSA-BPNN model established using the COS method to select 213 feature wavelengths had a calibration set R_c^2 value of 0.930 and an *RMSEC* value of 3.18%, while the prediction set had an R^2 value of 0.920 and an *RMSEP* value of 3.26%. The *RPD* value increased to 3.524. Compared to the COS-PSO-BPNN and PSO-BPNN models, the calibration set R_c^2 and prediction set R^2 increased by 1.60% and 1.10%, respectively, and the *RPD* value improved from 3.233 to 3.524. The model's predictive performance has been enhanced to some extent. The runtime of the SSA-BPNN model, built using the COS method to select feature wavelengths, was 34.27% of the model built using all wavelengths. Compared to the model built using all wavelengths, the prediction model efficiency was significantly improved.

Table 5. SSA-BPNN model evaluation results of cotton leaf samples with different quantities and wavelengths.

Feature Extraction Methods	Number of Wavelengths	Calibration Set		Validation Set		RPD	Prediction Time/s
		RMSEC	R_c^2	*RMSEP*	R^2		
Original Spectrum	846	4.08	0.914	3.62	0.909	3.233	189.79
COS	213	3.18	0.930	3.26	0.920	3.524	65.04

3.2.5. Model Comparison

Comparing the results from Tables 2–5, the predictive model performance for cotton leaf chlorophyll content, based on feature wavelengths selected using the COS method, was superior to the model established using the full spectrum of wavelengths. The BPNN model established using all wavelengths had R_c^2 and RPD values of 0.611 and 1.285, respectively, for the calibration set. The GA-BPNN model established using all wavelengths had R_c^2 and *RPD* values of 0.739 and 1.798, respectively, for the calibration set. The PSO-BPNN model established using all wavelengths had R_c^2 and *RPD* values of 0.804 and 2.432, respectively, for the calibration set. The SSA-BPNN model established using all wavelengths had R_c^2 and *RPD* values of 0.914 and 3.233, respectively, for the calibration set. Among them, the SSA-BPNN showed the most significant improvement, with R_c^2 and *RPD* values for the calibration set increasing by 0.303 and 1.948, respectively, compared to the BPNN model.

The results indicate that, when modeling based on all wavelengths, the SSA-BPNN model outperforms the BPNN, GA-BPNN, and PSO-BPNN models in terms of performance. The BPNN model established based on feature wavelengths selected using the COS method

had R^2, *RMSEP*, and *RPD* values of 0.721, 3.05%, and 1.443, respectively, for the prediction set. The GA-BPNN model established using feature wavelengths selected with the COS method had R^2, *RMSEP*, and *RPD* values of 0.814, 2.58%, and 2.188, respectively, for the prediction set. The PSO-BPNN model established using feature wavelengths selected with the COS method had R^2, *RMSEP*, and *RPD* values of 0.885, 2.58%, and 2.784, respectively, for the prediction set. The SSA-BPNN model established using feature wavelengths selected with the COS method had R^2, *RMSEP*, and *RPD* values of 0.920, 3.26%, and 3.524, respectively, for the prediction set. The results indicate that, after selecting feature wavelengths using the COS method, the regression performance of the SSA-BPNN model is significantly better than the BPNN model. The RPD of the GA-BPNN and PSO-BPNN models, optimized using the GA and PSO algorithms, respectively, increased from 1.443 to 2.188 and 2.784, respectively. This indicates that compared to the BPNN model, both the GA-BPNN and PSO-BPNN models exhibit improved regression performance, as well.

The schematic diagram of the fitted prediction models is shown in Figure 6. The coefficient of determination (R_f^2) and the residual sum of squares (*RSS*) express the degree of model fit. A higher coefficient of determination and a lower value of residual sum of squares indicate a better fit of the model. As shown in Figure 6, the COS-SSA-BPNN model has the highest R_f^2 value of 0.911 and the lowest *RSS* value of 0.066, indicating a good fit of this model. The COS-SSA-BPNN model also has a narrower 95% confidence interval for prediction errors and a more concentrated distribution of data points, suggesting stronger overall data consistency, stability, and representativeness. This implies higher reliability of sample parameters and stronger predictive ability for the COS-SSA-BPNN model.
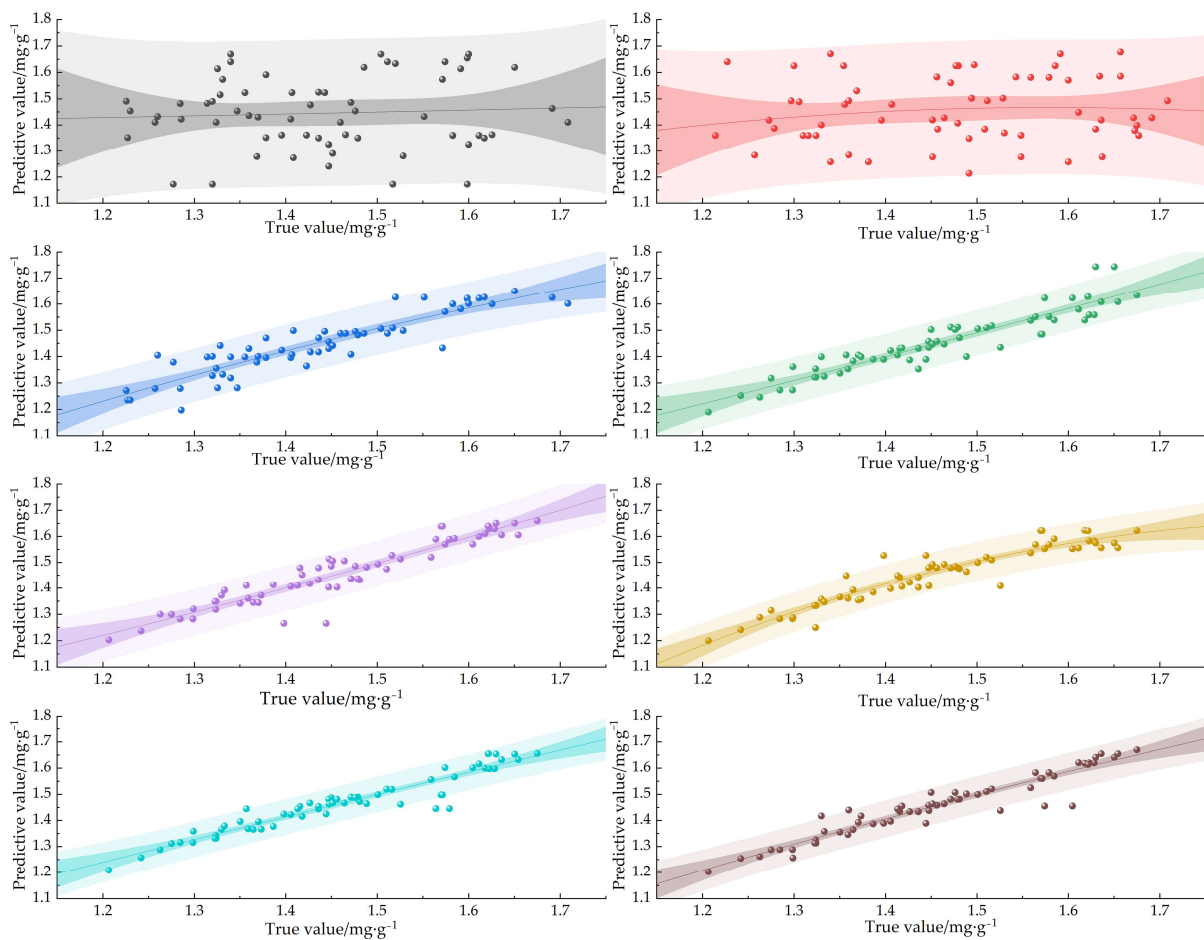

Figure 6. Prediction model for chlorophyll content in cotton leaves. Note: The dark stripe area is a 95% confidence band; Light-colored bands are 95% predicted bands.

By analyzing the model evaluation criteria (R_c^2, R^2, *RMSEC*, *RMSEP*, and *RPD*) and polynomial fitting situation for each model, it can be concluded that the SSA-BPNN model, built using feature wavelengths selected with the COS method, performs the best. Furthermore, this model has the shortest prediction time and highest efficiency. Therefore, it can be concluded that the combination of hyperspectral data and the COS-SSA-BPNN model is effective for quantitative detection of chlorophyll content in cotton leaves.

3.3. Orthogonal Experiment Verification

3.3.1. Orthogonal Experiment Design Plan

An orthogonal experiment with two factors, modeling wavelength quantity and modeling method, was designed using a two-factor four-level orthogonal design. The experiment followed an $L_8(4^2)$ orthogonal table design, with each experimental group repeated seven times and averaged, resulting in a total of eight experimental groups. Verify the optimal results of the *CLCC* prediction model, as described in Section 3.2. The experimental factors and levels are shown in Table 6, and the orthogonal experimental plan is presented in Table 7.

Table 6. Orthogonal experiment factor level table for chlorophyll content prediction.

Level	1	2	3	4
Data processing method	Original spectrum/A1	COS/A2	—	—
Modeling methods	BPNN/B1	GA-BPNN/B2	PSO-BPNN /B3	SSA-BPNN/B4

Table 7. Experimental scheme.

Serial Number	Data Processing Method	Modeling Methods
1	A1	B1
2	A1	B2
3	A1	B3
4	A1	B4
5	A2	B1
6	A2	B2
7	A2	B3
8	A2	B4

3.3.2. Experimental Results and Analysis

The orthogonal experiment results of DH10 *CLCC* are shown in Table 8. In this study, the prediction set R^2, *RMSEP*, and *RPD* were selected as reference indicators to describe the prediction effectiveness of the model for *CLCC*.

Table 8. Experimental results.

Indicator Name / Group Name	A	B	R^2	*RMSEP*	*RPD*
1	1	1	0.655	2.84	1.285
2	1	2	0.747	2.64	1.798
3	1	3	0.820	2.13	2.432
4	1	4	0.909	3.62	3.233
5	2	1	0.721	3.05	1.443
6	2	2	0.814	2.58	2.188
7	2	3	0.885	2.58	2.784
8	2	4	0.920	3.26	3.524

The optimal solution is the combination of preferable levels for each factor within the tested range. Higher values of R^2 and *RPD* indicate better performance, while a lower value of *RMSEP* is preferred. Table 9 presents the analysis of the experimental results.

Table 9. Analysis of test results.

Index		A	B
R^2	K_1	3.131	1.376
	K_2	3.340	1.561
	K_3	—	1.705
	K_4	—	1.829
	R	0.209	0.453
	Optimization	A2B4	
$RMSEP$	K_1	11.23	5.89
	K_2	11.47	5.22
	K_3	—	4.71
	K_4	—	6.88
	R	0.24	2.17
	Optimization	A1B3	
RPD	K_1	8.748	2.728
	K_2	9.939	3.986
	K_3	—	5.216
	K_4	—	6.757
	R	1.191	4.029
	Optimization	A2B4	

In this study, a visual analysis was conducted on each individual indicator to determine the optimal level combination for each indicator. Then, considering the practical application requirements, a comprehensive comparison analysis was performed using a comprehensive balance method to evaluate and determine the optimal solution. K_i represents the sum of the corresponding experimental results when the level number of any column (A, B) is i (i = 1, 2, 3, 4). R represents the range, which is calculated as R = max $\{K_1, K_2, K_3, K_4\}$–min $\{K_1, K_2, K_3, K_4\}$ for any given column. Analyzing R^2, the maximum values of K_i for factors A and B occur at K_2 = 3.340 and K_4 = 1.829, respectively. The optimal combination for this indicator is A2B4. Analyzing $RMSEP$, the minimum values of K_i for factors A and B occur at K_1 = 11.23 and K_3 = 4.71, respectively. The optimal combination for this indicator is A1B3. Analyzing RPD, the maximum values of K_i for factors A and B occur at K_2 = 9.939 and K_4 = 6.757, respectively. The optimal combination for this indicator is A2B4. The $RMSEP$ values for all eight experiments are less than 5%, and the measured chlorophyll content is 1.36 mg/g. This corresponds to 0.068 mg/g for the eight experimental groups, which has a relatively small impact on practical applications. Therefore, priority should be given to considering the R^2 and RPD values of the model. Taking practical considerations into account, the optimal solution is determined to be A2B4, which is consistent with the analysis results in Section 3.2. The results indicate that the COS-SSA-BPNN model is effective at detecting chlorophyll content in cotton leaves.

4. Discussion

In this study, the spectral information based on the visible and near-infrared (VNIR) wavelength range (400–1000 nm) was combined with machine learning techniques (BPNN, GA-BPNN, PSO-BPNN, SSA-BPNN). This successful integration allowed for the accurate determination of chlorophyll content in different growth stages of cotton. The model established using SSA-BPNN demonstrated the best predictive performance for cotton chlorophyll content.

Generally speaking, the reflectance spectrum of green plants is primarily influenced by leaf pigments within the visible light range, resulting in strong absorption and low reflectance. The negative correlation between *CLCC* and the spectrum within the visible light range indicates that higher chlorophyll content leads to lower spectral reflectance and stronger absorption. However, the samples used in this study come from three different growth stages, which introduces certain differences in the relationship between chlorophyll content and spectral information. Furthermore, the reflectance spectrum beyond visible

light is mainly influenced by cell structure and leaf water content. Although there are specific wavelength bands where *CLCC* and the spectrum demonstrate a highly significant correlation, it cannot be excluded that other factors may influence the relationship, presenting as numerical correlations.

In this study, due to the large amount of data, the detection performance of the BPNN model was relatively poorer, potentially indicating better performance for relatively smaller datasets, as confirmed by the research of Wei and Sun [18,19]. However, models optimized using algorithms demonstrated more significant predictive advantages, and similar phenomena can be found in the literature [21–25].

In Tables 2–5, there is an observed phenomenon where the *RMSEC* and *RMSEP* values increase, despite an increase in R_c^2 and R^2. However, the variation in *RPD* values follows the expected pattern. This may be attributed to the wide time span between the data points, which corresponds to the seedling stage (13 June), bud stage (10 July), and flowering stage (5 August), leading to variations in chlorophyll content. It is yet to be further investigated whether this phenomenon is a result of the automatic selection of sample data for the calibration and prediction sets during the modeling process, causing differences in the data.

Studies by researchers have shown that feature band selection can contribute to the improvement of predictive model performance [7,20,24,36]. Based on the preprocessing in this study, feature band selection was conducted using correlation analysis [18], and the selected spectra with a strong correlation were found to be more beneficial for predicting chlorophyll content [37,38]. The results of orthogonal experiments were consistent with the results obtained through individual comparative analysis of model predictions.

Section 3.3 innovatively applies orthogonal experiments to quantitative detection research, which is consistent with the results obtained in Section 3.2 through comparative methods. By validating the model performance, it also confirms the feasibility of this approach. In this study, eight models need to be established, and when preprocessing methods, modeling algorithms, or research targets increase, more models will be required, consuming a significant amount of time for model optimization. Taking the example of a three-factor, three-level design, we would need to establish 27 sets of models. However, using this approach, only nine sets of models need to be built based on the modeling scheme. This may be a novel research approach that can reduce modeling options and improve work efficiency. But, further investigation is needed to determine whether it can be applied to other detection studies.

The SSA-BPNN model established in this study can be used for quantitative estimation of chlorophyll content during the growth stages of DH10 cotton. However, the structure and parameters of the SSA-BPNN model were designed based on a specific cotton variety from Xinjiang. Further research is needed to determine whether the model can be successfully applied to the estimation of chlorophyll content in different varieties of cotton.

5. Conclusions

In this study, on the basis of high-spectral technology, machine learning techniques (BPNN, GA-BPNN, PSO-BPNN, and SSA-BPNN) were successfully employed in conjunction with the VNIR spectral range (400–1000 nm) to determine the chlorophyll content at different growth stages of cotton. Additionally, orthogonal experiments were introduced to validate the performance of the models, providing a new approach for studying quantitative detection models under the influence of multiple factors. The main conclusions are as follows:

(1) Spectral information of samples of cotton leaf chlorophyll content was obtained based on visible near-infrared hyperspectral imaging technology. The spectral data were preprocessed using the Savitzky–Golay quadratic smoothing method. The model performance of cotton leaf chlorophyll content prediction was compared between the model built with all wavelengths and the one built with feature wavelengths selected through correlation analysis. It was determined that the model built with the selected feature wavelengths exhibited better performance.

(2) The performance of the SSA-BPNN, GA-BPNN, and PSO-BPNN models built with all 846 wavelengths and 213 feature wavelengths extracted using COS were superior to the BPNN model. Among them, the SSA-BPNN model built with the 213 feature wavelengths extracted using the COS method exhibited the best performance and highest efficiency. Its *RPD* was 3.524, and the determination coefficients for the calibration set and prediction set were 0.930 and 0.920, respectively. The root-mean-square errors were 3.18% and 3.26% for the calibration set and prediction set, respectively.

(3) An orthogonal experiment was conducted to validate the optimal results, and the results indicated that the optimal solution was A2B4, which corresponded to the SSA-BPNN model built with the 213 feature wavelengths extracted using the COS method. This finding was consistent with the optimal results obtained in this study.

This study demonstrates that the combination of hyperspectral imaging and the COS-SSA-BPNN model can effectively achieve quantitative detection of cotton leaf chlorophyll content. The rapid detection model for chlorophyll content in DH10 cotton established in this study provides a reference for the detection of chlorophyll content in other cotton varieties. At the same time, it offers the corresponding technical support and theoretical basis for the development of low-cost cotton leaf chlorophyll content rapid detection systems.

References

1. Nijs, I.; Behaeghe, T.; Impens, I. Leaf nitrogen content as a predicto of photosynthetic capacity in ambient and global change conditions. *J. Biogeogr.* **1995**, *22*, 177–183.

2. Daughtry, C.S.T.; Walthall, C.L.; Kim, M.S.; De Colstoun, E.B.; McMurtrey, J.E., III. Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance. *Remote Sens. Environ.* **2000**, *74*, 229–239.

3. Li, S.T.; Liu, X.Y.; He, P. Analyses on nutrient requirements in current agriculture production in China. *J. Plant Nutr. Fertil.* **2017**, *33*, 1416–1432.

4. Wei, Q.; Zhang, B.Z.; Wei, Z. Estimation of canopy chlorophyll contention winter wheat by UAV multispectral remote sensing. *J. Triticeae Crops* **2020**, *40*, 8.

5. Shah, S.H.; Angel, Y.; Houborg, R.; Ali, S.; McCabe, M.F. A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat. *Remote Sens.* **2019**, *11*, 920.

6. Deng, S.Q.; Zhao, Y.; Bai, X.Y.; Li, X.; Sun, Z.D.; Liang, J.; Cheng, S. Inversion of chlorophyll and leaf area index for winter wheat based on UAV image segmentation. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 136–145.

7. Xiao, Q.; Tang, W.; Zhang, C.; Zhou, L.; Feng, L.; Shen, J.; Yan, T.; Gao, P.; He, Y.; Wu, N. Spectral Preprocessing Combined with Deep Transfer Learning to Evaluate Chlorophyll Content in Cotton Leaves. *Plant Phenomics* **2022**, *2022*, 9813841.

8. Yang, P.Q.; Tol, C.V.D.; Campbell, P.K.E.; Middleton, E.M. Fluorescence correction vegetation index (FCVI): A physically based reflectance index to separate physiological and non-physiological information in far-red sun-induced chlorophyll fluorescence-Science Direct. *Remote Sens. Environ.* **2020**, *240*, 111676.

9. Li, Y.; Ma, B.; Li, C.; Yu, G. Accurate prediction of soluble solid content in dried Hami jujube using SWIR hyperspectral imaging with comparative analysis of models. *Comput. Electron. Agric.* **2022**, *193*, 106655.
10. Berger, K.; Verrelst, J.; Féret, J.-B.; Hank, T.; Wocher, M.; Mauser, W.; Camps-Valls, G. Retrieval of aboveground crop nitrogen content with a hybrid machine learning method. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *92*, 102174.
11. Petteri, N.; Nathaniel, N.; Tarmo, L. Crop yield prediction with deep convolutional neural networks. *Comput. Electron. Agric.* **2019**, *163*, 104859.
12. Zhou, Y.; Zheng, J. Inversion Model Design of Chlorophyll a Based on BP Neural Network and Remote Sensing Image. In *Innovative Computing: IC 2020*; Springer: Singapore, 2020; pp. 531–538.
13. An, T.; Yu, S.; Huang, W.; Li, G.; Tian, X.; Fan, S.; Dong, C.; Zhao, C. Robustness and accuracy evaluation of moisture prediction model for black tea withering process using hyperspectral imaging. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2021**, *269*, 120791.
14. Danilo, T.O.; Rouverson, S.P.D.; Walter, M.; Cristiano, Z. Convolutional neural networks in predicting cotton yield from images of commercial fields. *Comput. Electron. Agric.* **2020**, *171*, 105307.
15. Yeh, T.-S. Bifurcation curves of positive steady-state solutions for a reaction–diffusion problem of lake eutrophication. *J. Math. Anal. Appl.* **2016**, *449*, 1708–1724.
16. Zhang, Q.; Smith, D.W.; Baxter, C.W. Artificial neural networks: A tool with significant potential in environmental engineering and science-introduction. *J. Environ. Eng. Sci.* **2004**, *3*, III–IV.
17. Zhang, Z. A gentle introduction to artificial neural networks. *Ann. Transl. Med.* **2016**, *4*, 370.
18. Wei, X.; Wu, L.; Ge, D.; Yao, M.; Bai, Y. Prediction of the Maturity of Greenhouse Grapes Based on Imaging Technology. *Plant Phenomics* **2022**, *2022*, 9753427. [PubMed]
19. Sun, B.Y.; Chang, Q.R.; Liu, M.Y. Inversion chlorophyll mass fraction in winter wheat canopy by hyperspectral reflectance. *Atca Agric. Boreali-Occident. Sin.* **2017**, *26*, 552–559.
20. Li, Y.Y.; Chang, Q.R.; Liu, X.Y.; Yan, L.; Luo, D.; Wang, S. Estimation of maize leaf SPAD value based on hyperspectrum and BP neural network. *Trans. Chin. Soc. Agric. Eng.* **2016**, *32*, 135–142.
21. Zhou, Y.; Lu, A.J.; Liu, X. Prediction of chlorophyll a content in water body based on BP neural network with improved Genetic Algorithm. *Electron. Test* **2022**, 37–42.
22. YYang, T.C.; Lu, J.S.; Liao, F.; Qi, H.; Yao, X.; Cheng, T.; Zhu, Y.; Cao, W.X.; Tian, Y.C. Retrieving potassium levels in wheat blades using normalised spectra. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102412. [CrossRef]
23. Mohamed, S.A.; Metwaly, M.M.; Metwalli, M.R.; AbdelRahman, M.A.E.; Badreldin, N. Integrating Active and Passive Remote Sensing Data for Mapping Soil Salinity Using Machine Learning and Feature Selection Approaches in Arid Regions. *Remote Sens.* **2023**, *15*, 1751.
24. Li, Y.; Wu, Y.; Liu, X. Regional ecosystem health assessment using the GA-BPANN model: A case study of Yunnan Province, China. *Ecosyst. Health Sustain.* **2022**, *8*, 2084458.
25. Wang, G.L.; Wang, W.J.; Cheng, K.; Liu, X.; Zhao, J.G.; Li, H.; Guo, E.H.; Li, Z.W. Hyperspectral imaging combined with back propagation neural network optimized by sparrow search algorithm for predicting gelatinization properties of millet flour. *Food Sci.* **2022**, *43*, 65–70.
26. Conrad, A.O.; Li, W.; Lee, D.-Y.; Wang, G.-L.; Rodriguez-Saona, L.; Bonello, P. Machine Learning-Based Presymptomatic Detection of Rice Sheath Blight Using Spectral Profiles. *Plant Phenomics* **2020**, *2020*, 8954085. [PubMed]
27. Palacios-Cabrera, H.; Jimenes-Vargas, K.; González, M.; Flor-Unda, O.; Almeida, B. Determination of Moisture in Rice Grains Based on Visible Spectrum Analysis. *Agronomy* **2022**, *12*, 3021.
28. Colovic, M.; Yu, K.; Todorovic, M.; Cantore, V.; Hamze, M.; Albrizio, R.; Stellacci, A.M. Hyperspectral Vegetation Indices to Assess Water and Nitrogen Status of Sweet Maize Crop. *Agronomy* **2022**, *12*, 2181.
29. Simic, M.A.; Matthew, R.; Patrick, R.; Tharindu, A.; Anuruddha, M. The importance of leaf area index in mapping chlorophyll content of corn under different agricultural treatments using uav images. *Int. J. Remote Sens.* **2018**, *39*, 5415–5431.
30. Li, J.; Chen, L. Comparative analysis of models for robust and accurate evaluation of soluble solids content in 'Pinggu' peaches by hyperspectral imaging. *Comput. Electron. Agric.* **2017**, *142*, 524–535.
31. Yao, H.B. *Hyperspectral Imaging for Food Quality Analysis and Control*; Elsevier: Amsterdam, The Netherlands, 2010; pp. 45–78.
32. Zhang, B.H.; Li, J.B.; Fan, S.X.; Huang, W.Q.; Zhang, C. Principles and applications of hyperspectral imaging technique in quality and safety inspection of fruits and vegetables. *Spectrosc. Spectr. Anal.* **2014**, *34*, 2743–2751.
33. Anna, S.; Piotr, B.; Monika, Z.; Wojciech, M.; Bozena, S. Detection of fungal infections in strawberry fruit by vnir/swir hyperspectral imaging. *Postharvest Biol. Technol.* **2018**, *139*, 115–126.
34. Gao, J.F. *Plant Physiology Experimental Techniques*; Higher Education Press: Beijing, China, 2006; pp. 74–76.
35. Nagasubramanian, K.; Jones, S.; Sarkar, S.; Singh, A.K.; Singh, A.; Ganapathysubramanian, B. Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems. *Plant Methods* **2018**, *14*, 86. [PubMed]
36. Ouyang, C.T.; Zhu, D.L.; Wang, F.Q. A Learning Sparrow Search Algorithm. *Comput. Intell. Neurosci.* **2021**, *2021*, 3946958.

37. Bai, Z.X.; Zhu, R.G.; Wang, S.C.; Zheng, M.C.; Gu, J.F.; Cui, X.M.; Zhang, Y.X. Quantitative detection of fox meat adulteration in mutton by hyper spectral imaging combined with characteristic variables screening. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 276–284.
38. Liu, J.; Chang, Q.R.; Liu, M.; Yin, Z.; Ma, W.J. Chlorophyll Content Inversion with Hyperspectral Technology for Apple Leaves Based on Support Vector Regression Algorithm. *J. Agric. Mach.* **2016**, *47*, 260–265+272.