eISSN 1981-1829

Ciência e Agrotecnologia
Science and Agrotechnology

Prediction of chlorophyll relative content in tea plant canopy using optimize GRNN algorithm and RPA multispectral images

Predição do conteúdo relativo de clorofila na copa da planta de chá usando algoritmo GRNN otimizado e imagens multiespectrais RPA

Qingyan Zhou[1*], Jincheng Zhang[1], Tangwei Wei[1], Wen Xing[2], Jing Wang[1], Youhua Zhang[1]

ABSTRACT

To quickly and accurately assess tea plant growth, this study aims to find a new way to predict the chlorophyll content in tea plant canopies using machine learning. Using remotely piloted aircraft equipped with multispectral cameras, images of tea plantation areas are captured and reflectance from four spectral bands is extracted, leading to the calculation of vegetation indices. Simultaneously, chlorophyll relative content in the tea plant canopies was collected on the ground using a detector. Four models, namely Random Forest (RF), Backpropagation neural network (BPNN), Radial basis function network (RBFN), and General Regression Neural Network (GRNN), were constructed to predict the chlorophyll relative content in tea plant canopies. Subsequently, important remote sensing variables were identified through RF filtering, followed by a comparison of the predictive performance of machine learning models under different input conditions. Lastly, by integrating the Sparrow Search Algorithm (SSA) to optimize the smoothing factor in the GRNN, the study explores the impact of optimization algorithms on the predictive performance of the GRNN model. Experiments indicate that within the established machine learning models, the GRNN demonstrates the highest predictive accuracy. By ranking the importance of remote sensing variables through RF, 18 significant remote sensing variables were selected, which enhanced the predictive accuracy of the machine learning models. The optimization of the GRNN smoothing factor through the SSA algorithm can significantly enhance the predictive accuracy of the GRNN model. Based on a series of experiments, the established RFSSA-GRNN prediction model demonstrates good predictive performance, with an reaching 0.84.

Index terms: Leaf physiological parameters; vegetation index; feature screening; intelligent optimization algorithm.

RESUMO

Para avaliar de forma rápida e precisa o crescimento de plantas de chá, este estudo visa encontrar uma nova maneira de prever o conteúdo de clorofila em dosséis de plantas de chá usando aprendizado de máquina. Utilizando aeronaves pilotadas remotamente equipadas com câmeras multiespectrais, imagens de áreas de plantações de chá são capturadas e a reflectância de quatro bandas espectrais é extraída, levando ao cálculo de índices de vegetação. Simultaneamente, o conteúdo relativo de clorofila nos dosséis das plantas de chá foi coletado no terreno usando um detector. Quatro modelos, nomeadamente Floresta Aleatória (RF), rede neural de retropropagação (BPNN), rede de função de base radial (RBFN) e Rede Neural de Regressão Geral (GRNN), foram construídos para prever o conteúdo relativo de clorofila nos dosséis das plantas de chá. Subsequentemente, variáveis importantes de sensoriamento remoto foram identificadas através de filtragem RF, seguidas por uma comparação do desempenho preditivo dos modelos de aprendizado de máquina sob diferentes condições de entrada. Por último, ao integrar o Algoritmo de Busca de Pardal (SSA) para otimizar o fator de suavização no GRNN, o estudo explora o impacto dos algoritmos de otimização no desempenho preditivo do modelo GRNN. Experimentos indicam que, dentro dos modelos de aprendizado de máquina estabelecidos, o GRNN demonstra a maior precisão preditiva. Ao classificar a importância das variáveis de sensoriamento remoto através de RF, 18 variáveis significativas de sensoriamento remoto foram selecionadas, o que melhorou a precisão preditiva dos modelos de aprendizado de máquina. A otimização do fator de suavização do GRNN através do algoritmo SSA pode melhorar significativamente a precisão preditiva do modelo GRNN. Com base em uma série de experimentos, o modelo de predição estabelecido RFSSA-GRNN demonstra um bom desempenho preditivo, com um alcançando 0,84.

Termos para indexação: Parâmetros fisiológicos da folha; índice de vegetação; triagem de recursos; algoritmo de otimização inteligente.

Agricultural Sciences

Ciênc. Agrotec., 48:e016123, 2024
http://dx.doi.org/10.1590/1413-7054202448016123

Editor: Renato Paiva

[1]School of Information and Computer, Anhui Agricultural University, Hefei, Anhui, China
[2]School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu, Sichuan, China

[*]Corresponding author: swallowchou@126.com

Received in October 13, 2023 and approved in February 08, 2024

Introduction

Tea originated in the southwestern part of China around 600,000 to 700,000 years ago and has since become a crucial component of the lifestyles of approximately 300 million people globally (Pan et al., 2022). In 2021, China's annual tea production reached approximately 3 million tons, accounting for around 50% of the world's total, making it the leading tea-producing country worldwide (Miao et al., 2022). Chlorophyll content in plant leaves has been widely utilized as a primary indicator material and method for estimating photosynthetic

capacity, health status, and resistance to various diseases (Krause, & Weis 1991). Moreover, the chlorophyll content can serve as an indicator of both the stress levels experienced by plants and the levels of nitrogen content (Wang et al., 2019). Wang et al. (2004) demonstrated that chlorophyll is a compound affecting the color of dried tea leaves by calculating the determination coefficient between greenness and chemical composition. By analyzing the characteristics of leaf color, tea soup color, chemical composition, and volatile flavor compounds of dried tea leaves and their correlation with the perceptual quality scores provided by the tea tasting panel, Wang et al. (2010) concluded that the overall tea quality score is positively correlated with chlorophyll concentration.

Accurately measuring the chlorophyll content in plants is often achieved through chemical methods, which require specific laboratory equipment and are destructive, time-consuming, and costly. However, handheld Soil and plant analyzer development (SPAD) chlorophyll meters can rapidly and non-destructively measure the chlorophyll relative content in plant leaves. Although the relationship between SPAD readings and chlorophyll content on potato leaves is weaker compared to chemical methods, the SPAD readings on wheat leaves are affected by the variety, sampling time, and space, leading to some discrepancies with the actual chlorophyll content of the plant (Uddling et al., 2007). For the leaves of potted corn, SPAD readings also change due to the decrease in leaf water content and variations in irradiance (Martínez, & Guiamet, 2004). Liu, Yang and Yang (2012) demonstrated through field-scale research that SPAD readings can estimate the chlorophyll relative content in tea leaves without being affected by time and spatial factors.

Multispectral imagery is a remote sensing technique widely applied in the field of agriculture. By capturing the spectral information reflected by plants, it provides valuable insights into crop growth and health. Curran (1983) pioneered the utilization of multispectral cameras installed on aircraft and satellites to gather multispectral reflectance data, enabling the estimation of the green leaf area index. Xiao and McPherson (2005) employed satellite remote sensing data to assess the health conditions of trees at the University of California, Davis. This assessment was conducted using three spectral bands: red, red-edge, and green. Noh et al. (2006) employed multispectral imaging sensors installed on agricultural machinery to estimate the SPAD values of maize leaves under varying nitrogen application levels. This estimation was achieved by measuring the reflectance of the maize canopy in three channels.

In recent years, remotely piloted aircraft (RPA) has been widely employed in various fields of plant growth monitoring due to its simplicity of operation, high spatiotemporal resolution, and dynamic and timely data acquisition. In recent research, Shi et al. (2022) utilized multi-source remote sensing images, including RPA imagery, and employed machine learning algorithms to assess key phenotypic parameters in tea gardens. They discovered the optimal combinations of parameters for above-ground biomass (AGB) and leaf area index (LAI) across multiple tea gardens. By utilizing visible light RPA imagery and simultaneously collecting elderberry tea leaves and tea buds to measure their nitrogen content, Wang et al. (2023) proposed a pixel-level nitrogen content prediction method based on machine learning and deep learning. However, to date, the application of RPA s in monitoring tea plant growth and predicting vegetation parameters remains relatively limited.

In the field of modeling methods, the current methods for predicting plant parameters using remote sensing data can be categorized into two types: statistical analysis and machine learning. Statistical analysis involves establishing regression models based on the linear relationship between crop parameters and spectral data. For example, Chen et al. (2020) utilized univariate and multivariate linear regression methods to select six multispectral bands and 13 commonly used vegetation indices for the estimation of cotton plant water content. Gano et al. (2021) used a least squares linear regression model to analyze phenotypic parameters of West African sorghum under two different moisture conditions using multispectral RPA remote sensing imagery-derived vegetation indices. However, statistical analysis models are primarily used for linear problems and have lower predictive capabilities for nonlinear problems.

Machine learning methods enable computers to have the ability to learn on their own. They can independently analyze and process data, thereby establishing a correlation model between plant phenotypic characteristics and remote sensing variables. Previous studies have compared machine learning models with statistical analysis models. For instance, Chen et al. (2023) estimated winter wheat canopy chlorophyll relative content using both single-factor regression and machine learning methods, demonstrating that machine learning models achieved higher accuracy. Guo et al. (2022) used a Stepwise Regression Model (SRM) to determine the optimal combination of spectral and texture indices for estimating SPAD values. Subsequently, Support Vector Machine (SVM) and Random Forest (RF) models were applied to estimate the SPAD values of maize leaves based on the optimal combination. The estimation model using SVM achieved better prediction results, with an of 0.81. Yin et al. (2023) utilized multispectral images obtained from RPA as input for various machine learning models to predict the relative chlorophyll content in the potato crop canopy. The study found that the Random Forest model was the most effective, achieving values of 0.61, 0.79, 0.83, and 0.76 for the tuber initiation, tuber bulking, starch accumulation, and overall growth stages, respectively. This performance was superior to other models at all stages.

However, these studies often lack a comprehensive feature importance selection method, relying on either all input features or solely on the relationship between a single variable and the target variable to confirm feature importance. Additionally, there is limited research in the field of tea plant remote sensing on

comparing the performance of different machine learning models for predicting chlorophyll relative content. Therefore, further discussion is needed to determine the best modeling approach for predicting chlorophyll in tea plant canopy leaves.

With the rapid development of information technology, inspired by human intelligence, social behavior of biological populations, and natural phenomena, many intelligent optimization algorithms have been invented to address complex optimization problems. In recent years, intelligent optimization algorithms have gained significant attention and widespread application in fields such as financial engineering, bioinformatics, medicine, agriculture, and RPAs. These algorithms have overcome the limitations of traditional optimization methods and are effective in saving time and effort when dealing with large and complex parameter optimization problems. Lu et al. (2022) employed the Particle Swarm Optimization (PSO) technique to optimize the Extreme Learning Machine (ELM) algorithm. They selected six vegetation indices and SPAD values for correlation analysis and established a model for estimating SPAD values based on vegetation indices. By initializing the PSO correlation coefficients with the optimal convergence values of the fitness function, they effectively addressed issues related to the randomness of ELM model weights, thresholds, and network parameters. By initializing the coefficients of the PSO with the optimal approximation derived from the convergence of fitness function values, this method addresses the issues of randomness in weights and thresholds, as well as the uncertainty in network parameters in the ELM model. Consequently, it improved the model's prediction accuracy, raising the from 0.748 to 0.856. This led to the development of a technique for estimating the relative chlorophyll content in leaves affected by red date spider mite infestation.

This approach ultimately led to the development of a method for estimating chlorophyll content in jujube leaves under the influence of mite infestation. However, overall, the application of intelligent optimization algorithms in combination with machine learning in the field of agricultural remote sensing parameter prediction is relatively limited and requires further research.

In conclusion, the main objective of this article is to achieve precise prediction of tea canopy chlorophyll relative content using multispectral images collected by RPA. The model establishment process is divided into three steps:

(1) Utilizing four machine learning models to predict the tea canopy's chlorophyll relative content and comparing their performance.

(2) Constructing a comprehensive remote sensing variable importance assessment method and comparing the model performance before and after variable selection among the chosen machine learning algorithms.

(3) Attempting to integrate intelligent optimization algorithms into the selected machine learning models to establish a tea canopy chlorophyll relative content prediction model with higher prediction accuracy.

Material and Methods

The overall experimental design, as shown in Figure 1, begins with the utilization of a RPA with a multispectral camera to capture remote sensing images of the study area. Simultaneously, a SPAD detector is used on the ground to collect data on chlorophyll relative content and record its corresponding locations. Subsequently, the multispectral images are cropped, stitched together, and background noise is removed. The regions of interest are selected, and their reflectance values are collected, followed by the calculation of vegetation indices. Next, four machine learning algorithms are employed to construct predictive models, and feature selection methods and optimization algorithms are applied to modify these models. Finally, regression evaluation metrics are used to assess the predictive performance of the different models.

Study area

The study area is located within the Jiangsu Bocha Agricultural Science and Technology Development Co., Ltd.'s high-standard tea plantation in the Taiwan Farmers Entrepreneurial Park, Hengxi Street, Nanjing, China. It is situated at 118°44' East longitude and 31°43' North latitude. The study area falls within the subtropical northern monsoon climate zone, characterized by a mild climate with an annual average temperature of 15.7 degrees Celsius. The frost-free period averages 224 days, and there is abundant rainfall with an annual average precipitation of 1072.9 millimeters. The rainy season coincides with the warm season, meeting the growth requirements of tea plants. The plantation primarily cultivates various high-quality tea varieties, including group planting, Yingshuang, Zhongcha 108, Zhongbai 1, and Zhonghuang 3. Based on row spacing, tea plant growth status, and inter-row weed density, the study area was chosen within the group planting zone, characterized by larger inter-canopy spacing and lower inter-row weed density. There were no significant differences in field management practices, such as fertilization, irrigation levels, and planting density within the study area. The study area is depicted in Figure 2.

Ground data measurement

Due to the time-sensitive nature of leaf chlorophyll relative content, leaf chlorophyll content in the tea canopy was determined using the TYS-A handheld chlorophyll meter from Beijing Zhongke Weihe Company. The working principle involves utilizing the absorption characteristics of chlorophyll in plant leaves to specific wavelengths of light, allowing the instrument to emit red and near-infrared light towards the plant leaves and measure the transmittance or reflectance of these two types of light. Since chlorophyll absorbs these two types of light to different extents, by calculating the ratio of the absorption difference between the two types of light, the instrument can determine the relative chlorophyll content in the leaves.

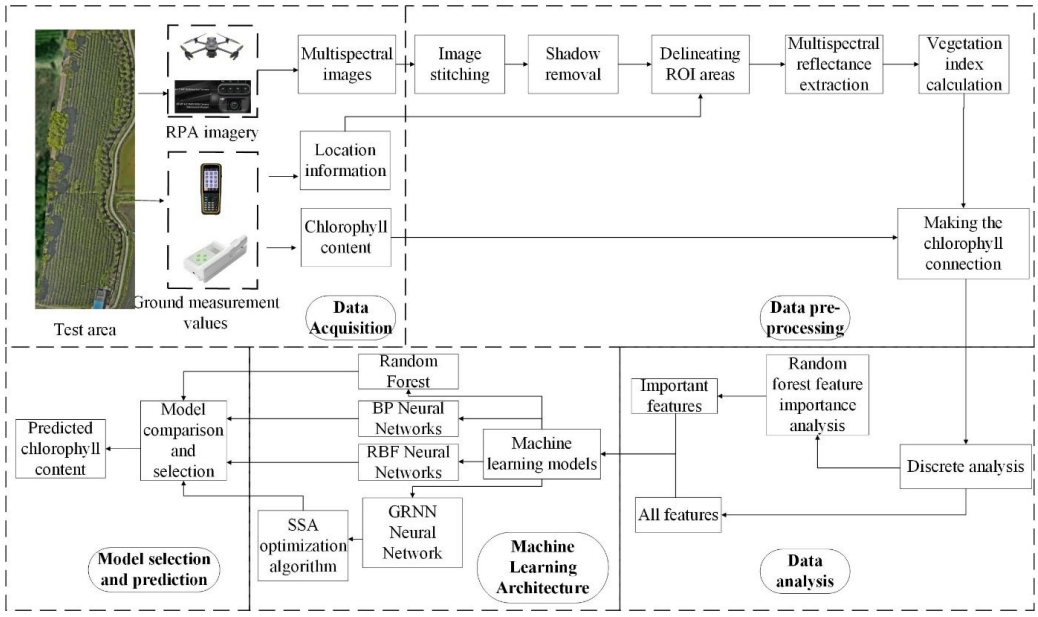


Figure 1: Experimental design for predicting chlorophyll relative content in the canopy of tea plants.

Using a 70cm×70cm sampling frame, select 10 rows of tea trees near the center of the tea planting area for sampling. In each row, choose 10 sampling plots, with each plot's center points being 4 meters apart from each other, resulting in a total of 100 experimental plots selected within the trial area. Taking into consideration the characteristics of tea leaves and the distribution of young and old leaves in the canopy, a five-point sampling method was applied at the top of the canopy. Five leaves with varying degrees of maturity, free from significant pest damage or withering, were chosen. Measurements were taken at the leaf tip, middle, and mid-lower sections of each leaf, avoiding the leaf veins. The average value of these measurements was taken as the chlorophyll relative content for that specific leaf. Subsequently, the average values of the five leaves were computed to represent the reference value for the canopy's chlorophyll relative content in each experimental plot. In total, 100 data points for chlorophyll relative content in the tea canopy were collected.

Using a handheld Global Navigation Satellite System (GNSS) to determine the specific location of the sampling plots, due to the traditional GNSS positioning accuracy of about 3 meters, we chose the Huace T7 smart measurement system equipped with a Real Time Kinematic (RTK) module produced by Beijing Huachen Beidou Information Technology Co., Ltd. Compared to traditional GNSS measurement systems, the RTK module can provide centimeter-level positioning accuracy, which is crucial for applications requiring high precision. Additionally, the RTK module is more effective in dealing with multipath effects and obstructions compared to other traditional positioning technologies. This means that even in obstructed environments, it can still maintain high-precision positioning. All equipment used in ground measurements is shown in Figure 3. To further ensure accurate acquisition of the sampling frame's location information, we use the center of each sampling frame as the sampling point and record the location information of its ground projection point. During the sampling process, we use a tape measure to ensure that the distance between the centers of adjacent sampling frames is 4 meters.

RPA Image collection and image preprocessing

Data collection

The acquisition of multispectral remote sensing data was carried out using the DJI Mavic 3M quadcopter. This system integrates a multispectral camera with a visible light camera, enabling the simultaneous capture of visible light images and images in four spectral channels. The visible light camera has a resolution of 20 million pixels, a 4/3 Complementary Metal Oxide Semiconductor (CMOS) sensor, and a mechanical shutter. The four 5-million-pixel monochrome sensors cover the following spectral bands: green (G, center wavelength 560 nm, bandwidth 16 nm), red (R, center wavelength 650 nm, bandwidth 16 nm), red edge (RE, center wavelength 730 nm, bandwidth 16 nm), and near infra-red (NIR, center wavelength 860 nm, bandwidth 26 nm).

The remote sensing imagery was acquired on April 9, 2023, between 10:30 and 13:50. During the capture period, the study area experienced partly cloudy weather with ample sunlight, a temperature of 23 degrees Celsius, an average wind speed of 3.9 kilometers per hour, and a solar zenith angle of 66 degrees at noon. Flight paths were planned within the experimental site boundaries. The DJI Mavic 3M RPA is equipped with a light

intensity sensor, allowing simultaneous collection of illumination data during the imaging process for initial radiometric correction. To avoid the impact of cloud cover obscuring sunlight on remote sensing images, throughout the entire RPA photography process, special attention was paid to ensure that the research area was always under direct sunlight. To further calibrate reflectance values, multispectral images of three polytetrafluoroethylene (PTFE) calibration panels with sizes of 200mm×200mm and radiance reflectance values of 25%, 50%, and 75% were captured prior to the main data acquisition.

During data collection, the RPA's flight altitude was set at 20 meters, with a speed of 5 meters per second. The Ground Sampling Distancen (GSD) was 3 centimeters per pixel, and automatic capture mode was employed. Overlap percentages in the forward and side directions were set at 80% and 70%, respectively. To enhance image stitching accuracy in post-processing, four ground control points were established within the study area, and their coordinates were measured using a GNSS geodetic receiver equipped with an RTK module for the materialization of the terrestrial reference for georeferencing purposes.

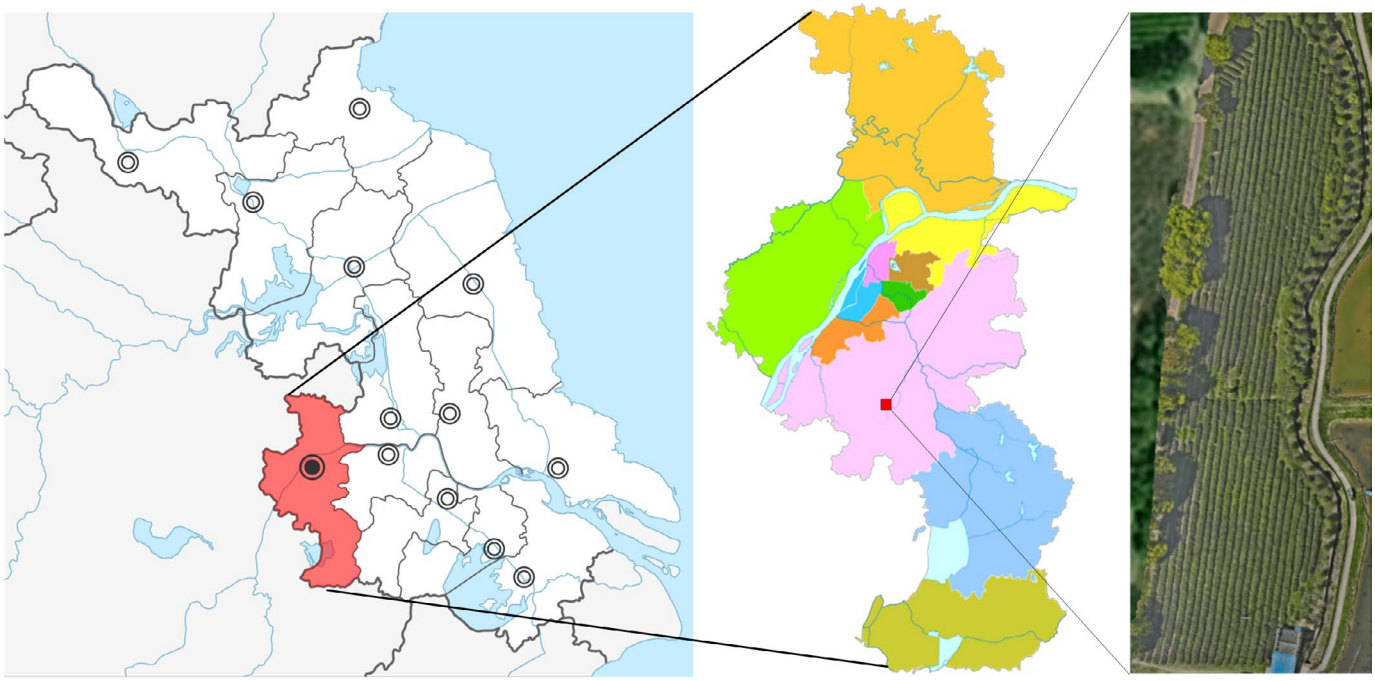
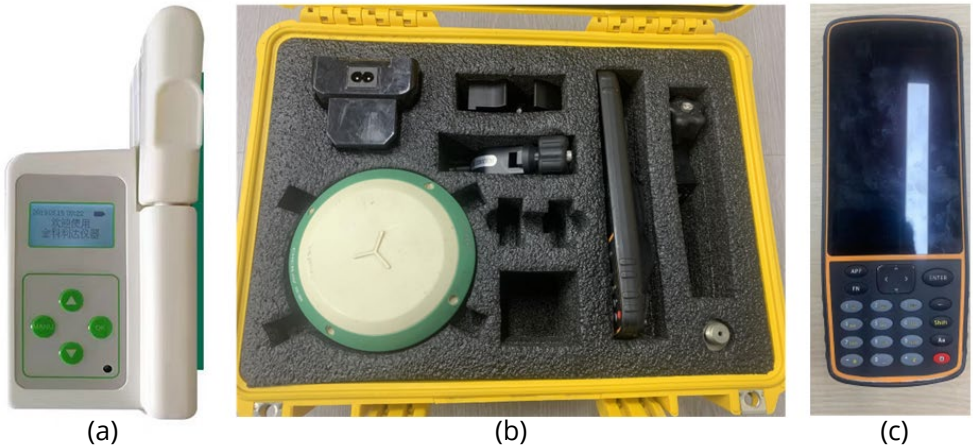


Figure 2: The geographical position of the study site.



(a) (b) (c)

Figure 3: Equipment used for ground measurement. (a) Handheld chlorophyll meter (b) GNSS-RTK positioning module (c) GNSS Handheld Notebook.

Image preprocessing

After the data collection was completed, the DJI Terra software was used to perform the synthesis and stitching of multispectral image data and visible light images, as well as radiometric calibration. DJI Terra software can automatically match the image positions based on the information contained in the images. First, 466 sets of measured images were imported, and then the previously set ground control points were imported as reference points for further rectification of images captured in different spectral bands. Images of PTFE calibration boards with different reflectances, captured by the multispectral camera at a height of 2.3 meters, were selected. The software sets the reflectance of the calibration board to 25%, 50%, and 75% respectively. The agricultural multispectral modeling mode was chosen. The software reconstructs the two-dimensional multispectral model through aerial triangulation and automatically performs position and radiometric calibration. Finally, it completes the stitching, resulting in a Tag Image File Format (TIF) format file containing reflectance information for four spectral bands.

Shadow removal

There is a significant difference in reflectance between the soil and the tea plant canopy. Additionally, due to the complex and three-dimensional structure of the tea plant canopy, some areas within the canopy may exhibit withered leaves. As a result, the acquired images often contain both soil background and canopy shadows, which have a negative impact on the extracted reflectance, leading to a reduction in modeling and inversion accuracy. According to Meng (2023), in multispectral remote sensing images, the Green Normalized Difference Vegetation Index (GNDVI) can effectively remove these two types of interference and significantly extract the areas of tea plant canopy leaves. Therefore, GNDVI was selected as the mask for the experimental area images. By setting the pixel values to zero for areas with values below the threshold, soil background and canopy shadows are eliminated. The calculation method for GNDVI is provided in Equation 1. Based on experimental results, the threshold is set at 0.135.

$$GNDVI = ((NIR) - G)/((NIR) + G) \quad (1)$$

Extraction of raw spectral reflectance and vegetation index calculation

Single-band image data is obtained through a multispectral camera and is processed through image registration using DJI Terra software to perform tasks such as image synthesis, stitching, and radiometric calibration, resulting in a TIF format file. These stitched images are imported into ENVI 5.6. Soil and shadow-affected bands are removed, and with the aid of GNSS measurements taken in the field, the Regions of Interest (ROI) are extracted by manually delineating the imagery of each rectangular sampling frame. To calibrate the positions of the ROIs accurately, the location of the center point within each rectangular ROI is checked individually during selection to ensure it aligns with the geographic location previously measured in the field. This process yields the average spectral reflectance of all vegetation canopy pixels within the ROIs for each of the four bands, which is subsequently considered as the raw spectral reflectance of the tea plant canopy in each respective subarea, as illustrated in Figure 4.

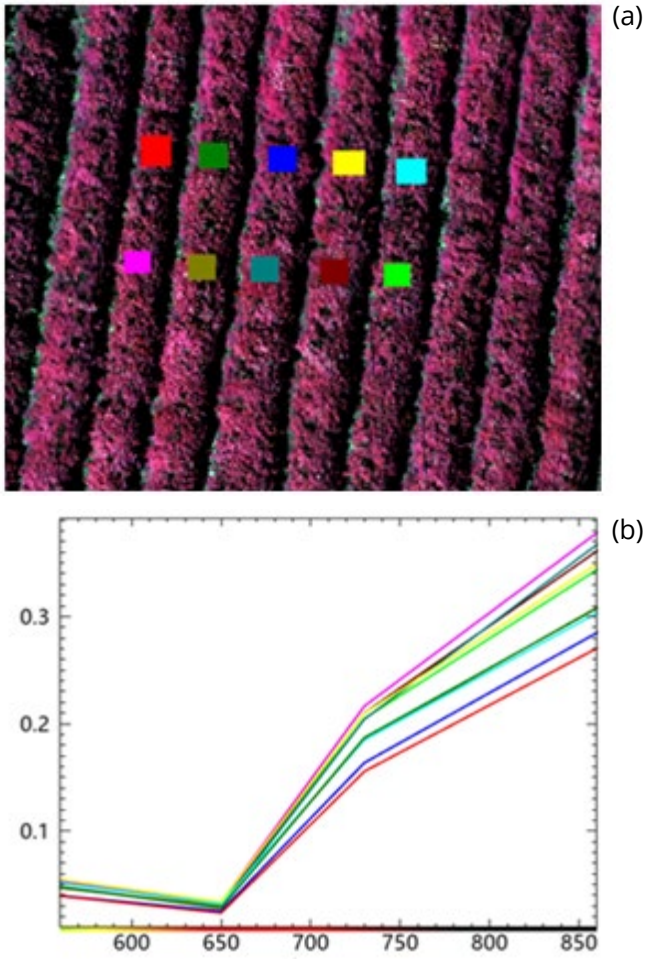


Figure 4: Selection of ROIs and extraction of reflectance values in the multispectral image. (a) Delimitation of the samples; (b) Spectral response of the samples.

The multispectral images, encompassing four spectral bands, offer the potential for generating a variety of vegetation indices using different computation methods. A total of 37 commonly used vegetation indices closely related to plant growth status were selected for monitoring the chlorophyll relative content in the tea plant canopy. Index formulas, as presented in Table 1, were constructed using ENVI 5.6 software to calculate the corresponding vegetation indices for the regions of interest.

Table 1: Vegetation Indices Used in the Study.

Vegetation Index	Formula	Source
Chlorophyll Index-green (CI-GREEN)	$\frac{NIR}{G}-1$	(Gitelson, Gritz, & Merzlyak, 2003)
Chlorophyll Index-red (CI-RED)	$\frac{NIR}{R}-1$	(Clevers, Kooistra, & van den Brande, 2017)
Chlorophyll Index-red edge (CI-REDEDGE)	$\frac{NIR}{RE}-1$	(Gitelson, Gritz, & Merzlyak, 2003)
Content Validity Index (CVI)	$\frac{NIR}{G}\times\frac{R}{G}$	(Vincini, Frazzi, & D'Alessio, 2008)
Dynamic Valgus (DVI)	$NIR-R$	(Rouse et al., 1974)
DVI-REG	$NIR-RE$	(Rouse et al., 1974)
Enhanced Vegetation Index 2 (EVI2)	$2.5\times\frac{(NIR-R)}{(NIR+2.4R+1)}$	(Jiang et al., 2008)
Enhanced Vegetation Index2 -2 (EVI2-2)	$2.4\times\frac{(NIR-R)}{(NIR+R+1)}$	(Jiang et al., 2008)
Green Normalized Difference Vegetation Index (GNDVI)	$\frac{(NIR-G)}{(NIR+G)}$	(Gitelson, Gritz, & Merzlyak, 2003)
Green Optimized Soil Adjusted Vegetation Index (GOSAVI)	$1.16\times\frac{(NIR-G)}{(NIR+G+0.16)}$	(Liu, Pattey, & Jego, 2012)
Green Red Vegetation Index (GRVI)	$\frac{(G-R)}{(G+R)}$	(Zhang et al., 2018)
Leaf Chlorophyll Index (LCI)	$\frac{(NIR-RE)}{(NIR-R)}$	(Xiao et al., 2014)
Modified Chlorophyll Absorption in Reflectance Index (MCARI)	$(RE-R)-0.2\times(RE-G)\times\frac{RE}{R}$	(Daughtry et al., 2000)
Modified Chlorophyll Absorption in Reflectance Index 1 (MCARI1)	$1.2\times\left[2.5(NIR-R)-1.3(NIR-G)\right]$	(Haboudane et al., 2004)
Modified Chlorophyll Absorption in Reflectance Index 2(MCARI2)	$\frac{\left(3.75(NIR-R)-1.95(NIR-G)\right)}{\sqrt{(2NIR+1)^2-6\left(NIR-5\sqrt{R}\right)-0.5}}$	(Haboudane et al., 2004)
Modified Non-Linear Index (MNLI)	$\frac{\left(1.5NIR^2-1.5G\right)}{\left(NIR^2+R+0.5\right)}$	(Gong et al., 2003)
Modified Simple Ratio (MSR)	$\frac{(NIR/R-1)}{\sqrt{NIR/R+1}}$	(Chen, 1996)
Modified Simple Ratio-Rededge (MSR-REDEDGE)	$\frac{(NIR/RE-1)}{\sqrt{NIR/RE+1}}$	(Chen, 1996)

Continue...

Table 1: Continuation.

Normalized Difference Red Edge (NDRE)	$\frac{(NIR-RE)}{(NIR+RE)}$	(Gitelson, & Merzlyak, 1997)
Normalized Difference Red/Green Redness Index (NDREI)	$\frac{(RE-G)}{(RE+G)}$	(Hassan et al., 2018)
Normalized Area Vegetation Index (NAVI)	$1-\frac{R}{NIR}$	(Carmona, Rivas, & Fonnegra, 2015)
Normalized Difference Vegetation Index (NDVI)	$\frac{(NIR-R)}{(NIR+R)}$	(Rouse et al., 1974)
Optimized Soil Adjusted Vegetation Red Index (OSAVI-RED))	$1.6\times\frac{(NIR-R)}{(NIR+R+0.16)}$	(Rondeaux, Steven, & Baret, 1996)
Optimized Soil Adjusted Vegetation Green Index (OSAVI-GREE)	$1.6\times\frac{(NIR-G)}{(NIR+G+0.16)}$	(Rondeaux, Steven, & Baret, 1996)
Optimized Soil Adjusted Vegetation Rededge Index (OSAVI-RE)	$1.6\times\frac{(NIR-RE)}{(NIR+RE+0.16)}$	(Rondeaux, Steven, & Baret, 1996)
Renormalized Difference Vegetation Index (RDVI)	$\frac{(NIR-R)}{\sqrt{NIR+R}}$	(Roujean, & Breon, 1995)
Renormalized Difference Vegetation red Index (RDVI-REG)	$\frac{(NIR-RE)}{\sqrt{NIR+RE}}$	(Roujean, & Breon, 1995)
Red-Edge Triangulated Vegetation Index (RTVI-Core)	$100\times(NIR-RE)-10\times(NIR-G)$	(Walsh et al., 2018)
Ratio Vegetation Index (RVI)	$\frac{NIR}{R}$	(Rouse et al., 1974)
Soil-Adjusted Vegetation Index (SAVI)	$1.5\times\frac{(NIR-R)}{(NIR+R+0.5)}$	(Huete, 1988)
Soil-Adjusted Vegetation green Index (SAVI-GREEN)	$1.5\times\frac{(NIR-G)}{(NIR+G+0.5)}$	(Verrelst et al., 2008)
Simplified Canopy Chlorophyll Content Index (S-CCCI)	$\frac{NDRE}{NDVI}$	(Raper, & Varco, 2015)
Simple Ratio Index (SR-REG)	$\frac{NIR}{RE}$	(Walsh et al., 2018)
Transformed Chlorophyll Absorption in Reflectance Index (TCARI)	$3\times\left[(RE-R)-0.2(RE-G)\times\left(\frac{RE}{R}\right)\right]$	(Haboudane et al., 2002)
TCARI/OSAVI	$\frac{TCARI}{OSAVI}$	(Haboudane et al., 2002)
Triangular Vegetation Index (TVI)	$\frac{\left[120(NIR-G)-200(R-G)\right]}{2}$	(Broge, & Leblanc, 2001)
Wide Dynamic Range Vegetation Index (WDRVI)	$\frac{(0.2NIR-R)}{(0.2NIR+R)}$	(Gitelson, 2013)

Machine learning modeling

The algorithms used in this study were all implemented in the Python 3.8 environment and a Dell server containing a 16-core Intel processor was used for experiments. Four different machine learning methods, namely RF, Backpropagation neural network (BPNN), Radial Basis Function network (RBFN), and Generalized Regression Neural Network (GRNN), were employed to predict the chlorophyll relative content in the tea plant canopy. It's worth noting that the GRNN algorithm has been less commonly applied in previous studies for predicting vegetation parameters from remote sensing imagery. The GRNN algorithm is based on a radial basis network and requires only one network parameter, namely the smoothing factor σ, to be determined during the calculation process. In order to reduce redundancy between the original spectral reflectance and vegetation index feature sets and further lower feature dimensionality, we conducted experiments to assess the importance of multidimensional spectral reflectance and vegetation indices in the entire chlorophyll relative content prediction model by introducing Random Forest. Subsequently, in our experiments, we attempted to optimize GRNN by incorporating the Sparrow Search algorithm (SSA) algorithm to obtain a more accurate determination of the optimal smoothing factor σ, with the goal of establishing a more precise predictive model for tea plant canopy chlorophyll relative content.

GRNN network

GRNN is a feed-forward neural network based on radial basis functions. It has a simple structure and requires only one network parameter to be determined during the computation, namely the smoothing factor σ (Specht, 1991). GRNN has low computational complexity, requires fewer training parameters, exhibits high training efficiency, low error rates, and fast convergence speed. As a result, it has been widely used in various fields for building machine learning models with small data sets.

The GRNN consists of four layers, namely the input layer, pattern layer, summation layer, and output layer. Each layer is made up of a number of neurons. $X = [x_1, x_1, \ldots, x_n]^T$ stands for the network's input vector, and $Y = [y_1, y_1, \ldots, y_n]^T$. The network's output vector is denoted by the symbol T. The GRNN network's structure is shown in Figure 5.

Input layer: The input layer is where the signals enter the network. The size of the input signal vector affects how many neurons are present in the input layer. The input layer neurons just send the signals that are received to the pattern layer without any signal processing.

Pattern Layer: The pattern layer has the same number of neurons as samples in the input model. Each neuron represents a distinct sample. The Green's function is used by neurons in the pattern layer to process signals coming from the input layer and send them on to the summation layer. Equation 2 is the expression for the pattern layer's neuron transfer function.

$$P_i = \exp\left[-\frac{(X - X_i)^T (X - X_i)}{2\sigma^2}\right] \quad i = 1, 2, \cdots, n \tag{2}$$

Where X is the learning sample corresponding to the i-th neuron and X_i the network input variable. Using the exponential form of the Euclidean distance squared between the learning sample X_i and the input variable X, which is defined as $D_i^2 = (X - X_i)^T (X - X_i)$, the i-th neuron in the pattern layer computes the output value.

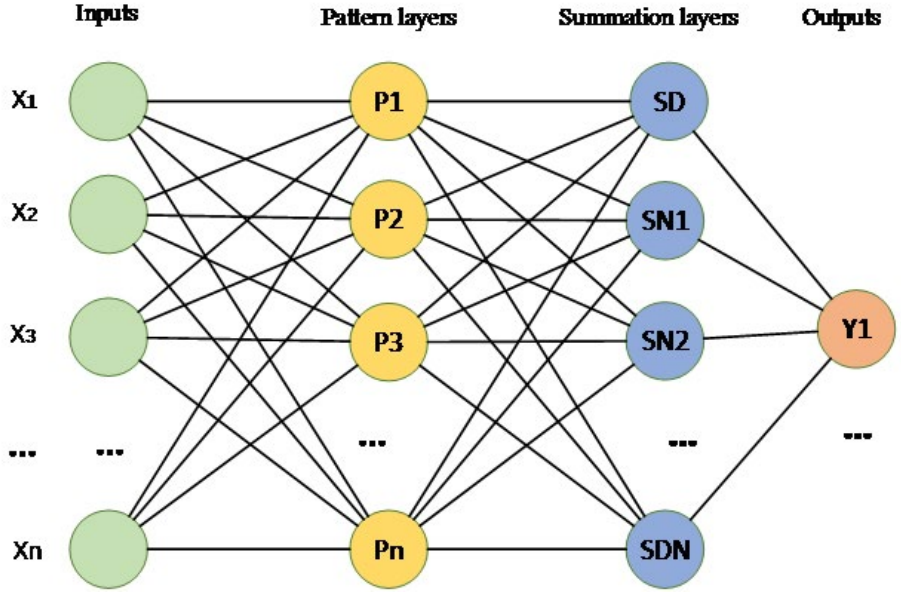


Figure 5: Structure of the GRNN Network.

Summation Layer: The summation layer consists of two separate types of neurons that use various summing techniques. The first type of neuron performs summation by computing the arithmetic sum of the output values from every neuron in the pattern layer while setting the connection weights between the summation layer and the pattern layer to 1. Equation 3 provides the summation expression for this sort of neuron, while Equation 4 expresses its related transfer function.

$$P_i = \sum_{i=1}^{n} \exp\left[-\frac{(X - X_i)^T (X - X_i)}{2\sigma^2}\right] \quad (3)$$

$$S_D = \sum_{i=1}^{n} P_i \quad (4)$$

The weighted total of the output values from every neuron in the pattern layer is computed by the second kind of summation neuron. The symbol y_{ij} stands for the weight of the link between the j-th neuron in the summation layer and the i-th neuron in the pattern layer. Equation 5 provides the summation expression for this sort of neuron, while Equation 6 expresses its related transfer function.

$$P_i = \sum_{i=1}^{n} Y_i \exp\left[-\frac{(X - X_i)^T (X - X_i)}{2\sigma^2}\right] \quad (5)$$

$$S_{Nj} = \sum_{i=1}^{n} y_{ij} P_i \qquad j = 1, 2, \cdots, k \quad (6)$$

Output Layer: The output layer of the neural network generates the output values for each sample, and the output parameter's dimension k determines how many neurons there are in the output layer. The summation layer's neurons' output values are divided into groups by the output layer. The output value of neuron *j*, which is denoted by Equation 7, is the output value of the network.

$$y_i = \frac{S_{Nj}}{S_D} \qquad j = 1, 2, \cdots, k \quad (7)$$

Vegetation index importance assessment

The original dataset consists of 100 samples, with each sample having four reflectance values and a total of 37 vegetation indices, resulting in a parameter space of 41 dimensions. To reduce redundancy in the original spectral reflectance and vegetation index feature set and further lower the feature dimensionality, we introduced random forest feature screening for the importance assessment of the multidimensional spectral reflectance and vegetation indices.

The main steps for evaluating feature importance using random forest are as follows:

steps 1: Utilize the Bootstrap approach to randomly sample the original dataset with replacement. Each sampling consists of 66 samples, and the remaining samples form the Out-of-Bag (OOB) set.

steps 2: For each sampled dataset, randomly select M1 (M1 < 41) features as input for training a decision tree. Construct the decision tree and, at each node of the tree, evaluate the best feature for splitting the samples based on these selected features.

steps 3: Repeat steps 1 and 2 K times to generate K decision trees, forming a random forest.

To assess the feature importance using random forest, many studies adopt the GINI coefficient method. However, this method may not perform well when dealing with datasets containing high-cardinality features. Therefore, in this study, we employ a feature ranking approach to determine the importance of different features. The specific method is described as follows.

In random forest, for each tree, the prediction error on the OOB test set is recorded as the mean squared error (MSE), as shown in the Equation 8.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (8)$$

In the formula, n is the number of samples in the OOB set, y_i is the actual value, and $\hat{y}_i$ is the predicted value. Then, the same process is performed after permuting each predictor variable. The difference is computed as the average across all trees, and the data is normalized by dividing it with the standard deviation of the differences. If the standard deviation of the differences for a variable is equal to zero, no division is performed because in this case, the average would be nearly zero. The larger the difference in MSE after permutation, the more important the variable is considered to be.

For each tree, the elements used to build the tree are randomly permuted in the OOB set. A new MSE is calculated, and the importance value of the variable is computed using the Equation 9.

$$\bar{\delta}_j = \frac{1}{B} \sum_{b=1}^{B} \left(MSE - MSE_{permuted_j}\right) = \frac{1}{B} \sum_{b=1}^{B} \delta_{bj} \quad (9)$$

In the formula, b represents each tree and j represents each variable. $\bar{\delta}_j$ is the average value across all trees B using variable j. The final importance of a feature is obtained by normalizing it with the standard error as shown in the Equation 10.

$$\%IncMSE = \frac{\bar{\delta}_{bj}}{\sigma_{\delta bj} / \sqrt{B}} \quad (10)$$

In the formula, $\sigma_{\delta bj}$ represents the standard deviation of $\bar{\delta}_{bj}$, and %IncMSE indicates the increase in mean squared error associated with each variable. A higher %IncMSE indicates a more important variable.

Sparrow search algorithm

According to Xue and Shen (2020), the SSA is a swarm optimisation method that simulates the foraging habits of sparrow populations, which include three different sorts of groups: finders, followers, and sentinels. Finders locate and mark food sources as global optima, continuously updating their positions to escape local optima and achieve better global search capability. Additionally, a proportion of sentinels is set to ensure the safety of the population, providing the algorithm with robustness. The combination of strong robustness and global search capability in this swarm intelligence optimization algorithm is crucial for determining the optimal value of the smoothing factor σ in the GRNN model and improving prediction accuracy. Therefore, in this study, SSA is selected as the method for optimizing the parameter value. The flowchart of SSA applied in the GRNN network is illustrated in Figure 6.

The diverse population updating process of SSA is illustrated as follows:

Update the position of the discoverer, as shown in Equation 11.

$$X_{ij}(t+1)=\begin{cases} X_{ij}(t)\cdot\exp\left(-\dfrac{i}{\alpha\cdot T}\right), R_2 < T_s \\ X_{ij}(t)+Q\cdot L, R_2 \geq T_s \end{cases}, \quad (11)$$

Within the given equation, t represents the current iteration count, while T denotes the maximum iteration count. $X_{ij}(t)$ represents the current location of the i -th individual discoverer. A random number α is drawn from the range R_2 (R_2 ϵ [0,1]) as the warning value, and another random number α is selected from the range T_s (T_s ϵ [0.5,1.0]) as the safety value. Q is a random variable following a normal distribution, and L is a 1×d matrix consisting of ones. If R_2 is less than T_s, it indicates that the foraging environment of the discoverer is safe. In such cases, the individual should depart from the current region and explore other secure areas for foraging.

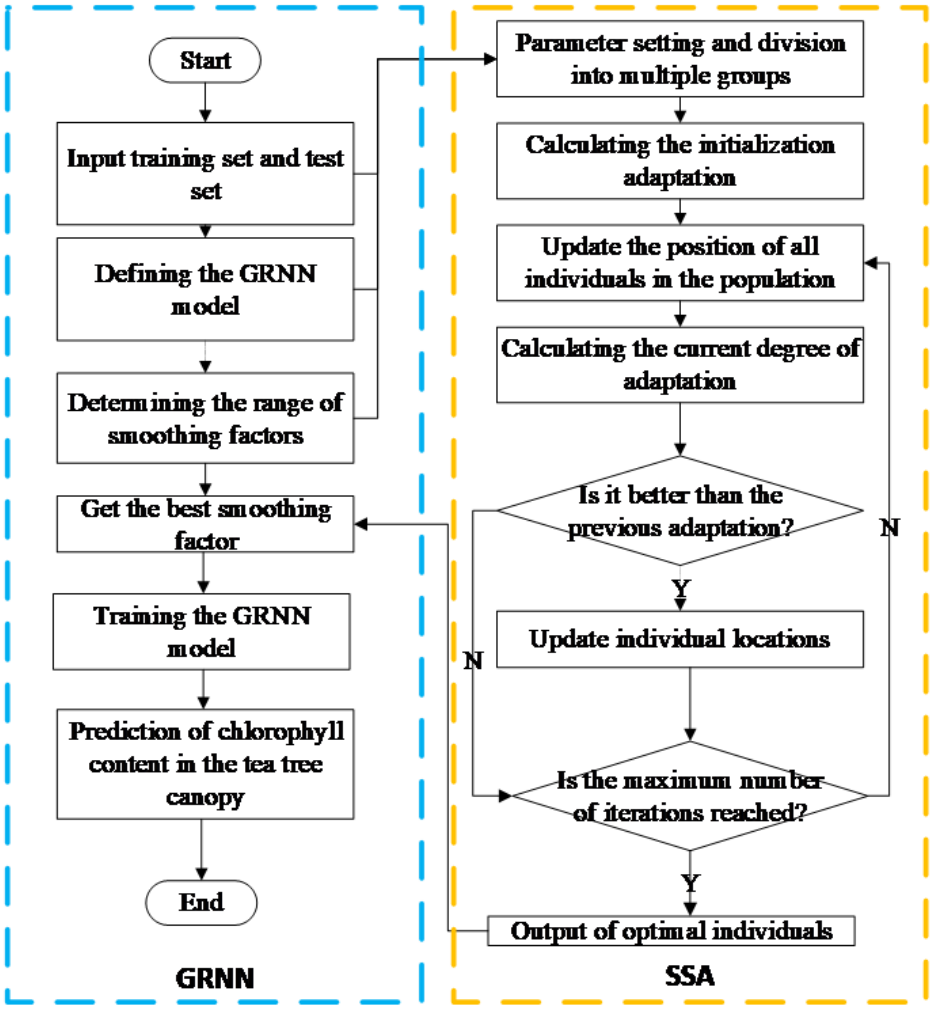


Figure 6: SSA-GRNN Model Flowchart.

Update the position of the followers, as shown in Equation 12.

$$X_{ij}(t+1)=\begin{cases} Q\cdot\exp\left(\dfrac{X_w(t)-X_{ij}(t)}{i^2}\right), i>\dfrac{n}{2}, \\ X_b(t+1)+|X_{ij}(t)-X_b(t+1)|\cdot A^{+}\cdot L, i\le\dfrac{n}{2} \end{cases} \quad (12)$$

Within the given equation, X_b represents the optimal position of the discoverer, while X_w represents the current worst position in the global context. A^+ represents the Moore-Penrose pseudo-inverse of A^T (obtained as AA^T), the matrix A is a 1×d vector where each element is randomly assigned a value of either 1 or -1. If i is greater than $n/2$, it indicates that the follower with lower energy cannot find food and needs to fly to other areas for foraging. Otherwise, the follower will move towards the set of optimal foraging positions.

Update the position of the sentinels, as shown in Equation 13.

$$X_{ij}(t+1)=\begin{cases} X_{best}(t)+\beta|X_{ij}(t)-X_{best}(t)|, f_i>f_g, \\ X_{ij}(t)+K\left(\dfrac{|X_{ij}(t)-X_w(t)|}{(f_i-f_w)+\in}\right), f_i=f_g \end{cases} \quad (13)$$

In the given equation, X_{best} represents the current best position in the global context. The step size is controlled by the parameter β. K is used to control the direction and step size of individual movement, taking arbitrary values within the range [-1, 1]. f represents the fitness value of the ith individual, f_g denotes the current global best fitness value, and f_w represents the current worst fitness value. ϵ is a small random number close to zero, introduced to avoid division by zero. If f_i is equal to f_g, it indicates that the sparrow is currently in the best position, allowing for random search in the nearby region. Otherwise, the sparrow is highly susceptible to attacks.

Through such a design, the SSA ingeniously draws inspiration from the foraging strategy of sparrows in nature, transforming it into an efficient swarm intelligence optimization method. This method not only effectively simulates the dynamic interaction among sparrows in different roles, such as the discoverer, follower, and scout, but also enables intelligent updating of individual positions on this basis, thereby facilitating the algorithm in finding the global optimum. Especially in dealing with complex optimization problems, such as the optimization of the smoothing factor σ in the GRNN model, SSA demonstrates its strong global search capability and robustness.

Regression evaluation design

To evaluate the performance of the tea plant canopy chlorophyll relative content prediction model, four types of parameters were chosen to calculate the error between the predicted chlorophyll relative content values and the actual values: coefficient of determination (R^2), MSE, root mean squared error (RMSE), and mean absolute error (MAE). A larger value of R^2 was utilised as the criterion, with a better inversion effect being indicated by a higher value. The MSE measures the total model error as the average squared difference between the predicted and actual values. The RMSE, which measures the overall prediction error while preserving the same scale as the initial observations, is the square root of the MSE. The average absolute difference between the projected and actual values is calculated by the MAE, giving a clear indication of the average deviation between the two. Equation 14, 15, 16 and 17, where n stands for the number of samples, y_i is the actual measured value, and $\hat{y}_i$ is the predicted value of the model, provides the equations for R^2, MSE, RMSE, and MAE, respectively:

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (14)$$

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)2 \quad (15)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (16)$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \quad (17)$$

Results and Discussion

Analysis of chlorophyll relative content in the tea canopy

Firstly, an analysis of the chlorophyll relative content in the tea canopy leaves collected will be conducted. In the selected sampling area of the experiment, there is a row of naturally growing trees on the west side, which results in some tea plants in the same tea garden being exposed to full sunlight conditions while others remain under partial shade in the afternoon. Its distribution is shown in Figure 7.

During the measurements, it was observed that there was a significant spatial distribution difference in the chlorophyll relative content within the tea canopy. The box plot depicting the distribution of chlorophyll relative content is shown in Figure 8.

It can be observed that the variations in the sampling area have a significant impact on the chlorophyll relative content in the leaf canopy. Designated the areas with shading at 2:00 PM as having shading conditions. From the figure, it can be seen that tea plants growing in shaded environments in the afternoon exhibit significantly higher chlorophyll relative content compared to those growing in full sunlight conditions. Chlorophyll content

in tea leaves is an important indicator for evaluating tea quality. Therefore, our measurements also confirm some previous studies, indicating that moderate low-light stress contributes to increased chlorophyll relative content in tea plants, thereby enhancing the quality of tea products (Sonobe, Sano, & Horie, 2018; Wang et al., 2010).

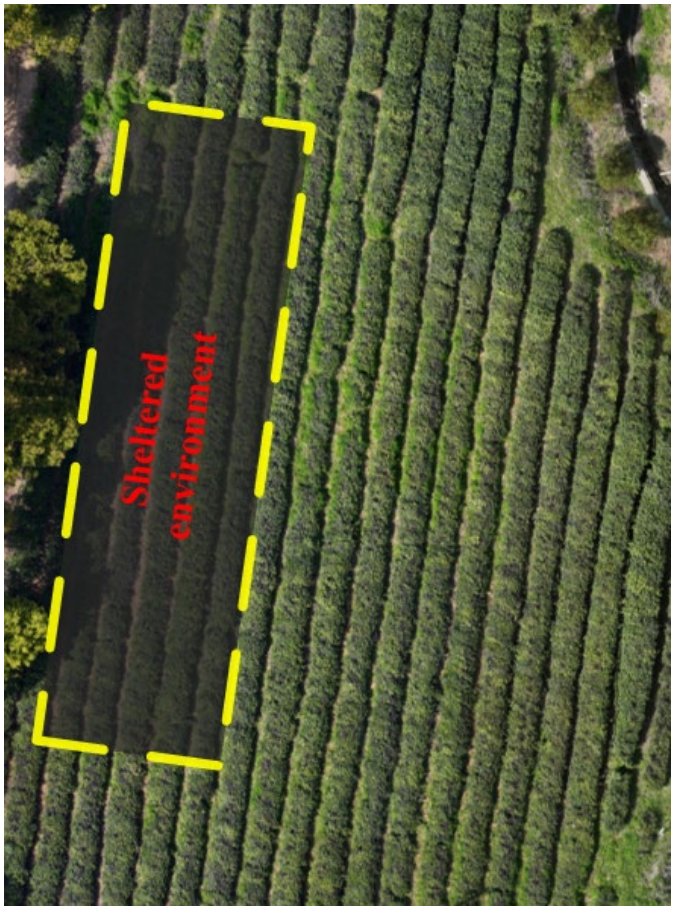


Figure 7: Distribution in a Sheltered Environment.

Although moderate shading benefits the enhancement of tea plant's chlorophyll relative content, most tea gardens still adopt open-field cultivation practices and do not deliberately employ shading measures. Therefore, in this study, we collected the data to establish a more widely applicable predictive model for chlorophyll relative content in the tea canopy.

The total sample size of tea canopy leaf chlorophyll relative content collected in the experimental plot is 100. To ensure the reliability of the data for constructing the tea canopy leaf chlorophyll relative content prediction model, we divided the collected data into seven equal parts based on their range, as shown in Figure 9. This was done to ensure that the collected values had a significant degree of dispersion and were distributed fairly evenly across the data segments. Subsequently, all sample data were randomly divided into two groups at a 4:1 ratio. Eighty samples were used as training data for analysis and modeling, while the remaining 20 samples were used as test data to verify the accuracy of the constructed model. Table 2 displays the standard deviation and coefficient of variation for the training samples, test samples, and the overall sample, showing similar levels of dispersion and confirming the reliability of the data partitioning.

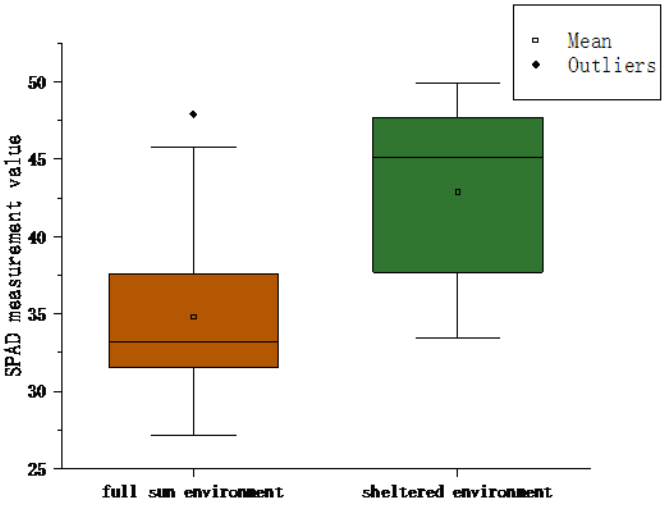


Figure 8: Distribution of chlorophyll relative content measurements in tea canopy under different illumination environments.

Ranking and selection of remote sensing variables based on importance

We employed the RF method to calculate and rank the feature importance of the 42 remote sensing variables. This ranking provided a reference for variable selection before constructing the model. Figure 10 illustrates the relationship between RF variable importance results and the predicted chlorophyll relative content. We applied a threshold for variable selection based on importance, following previous experiments in other fields (Genuer, Poggi, & Tuleau-Malot, 2010; Epifanio, 2017). Features were sorted in descending order of importance and accumulated until their cumulative importance reached 85% (Zhu et al., 2022). Based on this criterion, a threshold of 1.3% importance was established, resulting in the selection of 18 optimal remote sensing variables. These variables are MCARI2, MCARI, re, r, g, CVI, GRVI, NDREI, S-CCCI, RDVI-REG, OSAVI-REG, TCARI/OSAVI, TCARI, CI-GREEN, OSAVI, RTVI-CORE, GNDVI, and DVI-REG. Despite comprising only 43% of the total parameters, the sum of their importance for predicting final chlorophyll content reached 85.1%. This underscores the effectiveness of remote sensing variable selection in optimizing modeling features.

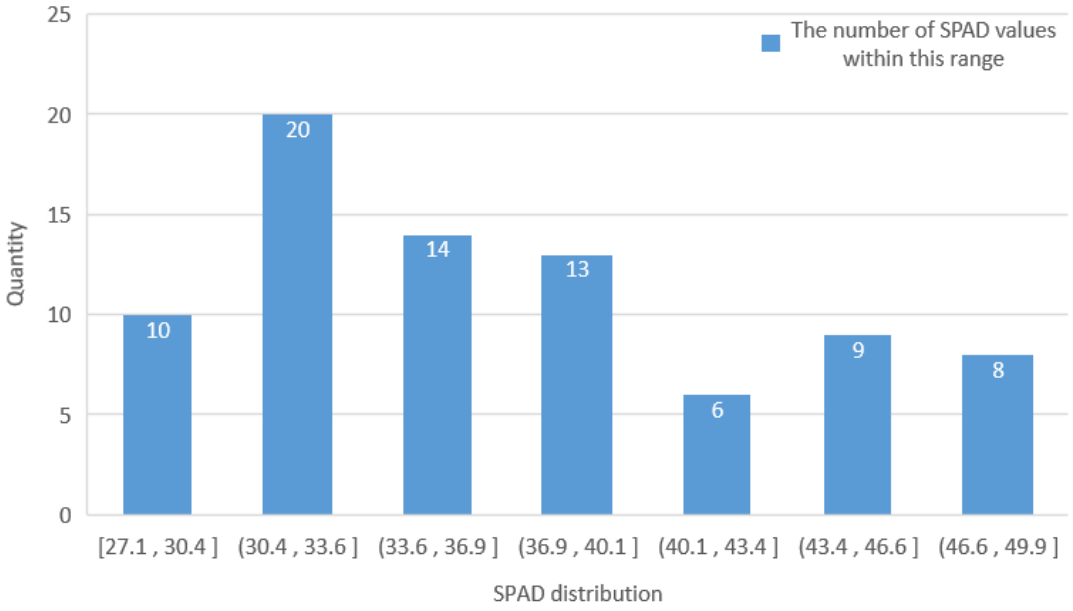


Figure 9: Distribution of SPAD data collected in the field.

Table 2: Statistics of SPAD values of tea plant canopy.

Data Type	Sample size	Minimum value	Maximum value	Average value	Standard deviation	Coefficient of variation(%)
Training set	80	27.14	49.9	37.08	6.05	16.30
Test set	20	27.46	48.6	37.39	6.86	18.35
All	100	27.14	49.9	37.03	6.14	16.6

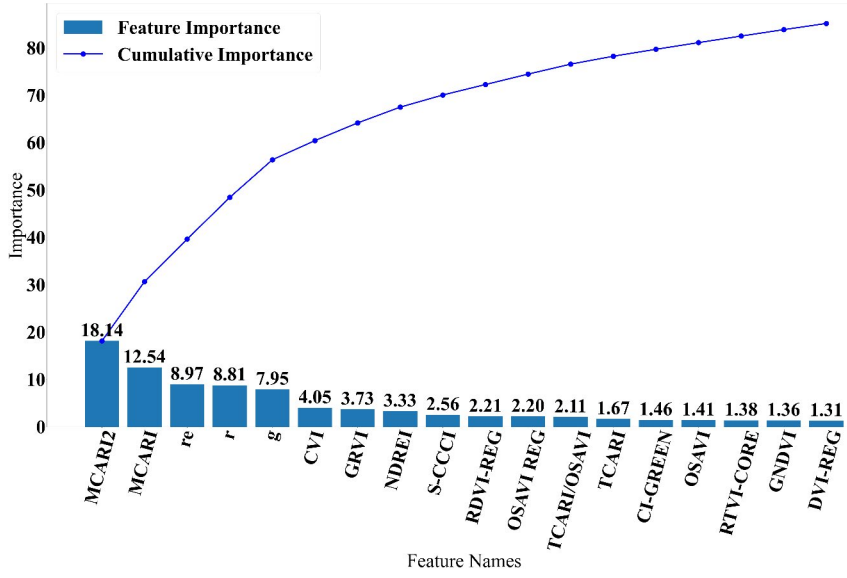


Figure 10: Importance assessment of remote sensing variables with importance greater than 1.3%.

Performance evaluation of machine learning models

The test results after training various models are shown in Table 3. Among the models that use all remote sensing variables as input, the predictive results of the four networks are illustrated in Figure 11. In terms of accuracy, from highest to lowest, they are GRNN, BPNN, RF, and RBFN. GRNN exhibits a slightly higher predictive accuracy compared to the other models, with an R^2 value of 0.59. It has improved by 0.05 compared to the second-best model, BPNN. This improvement can be attributed to the fact that, especially when dealing with a limited sample size, GRNN demonstrates a stronger predictive ability compared to other machine learning methods (Izonin et al., 2021). However, this advantage is not very pronounced, and it's worth noting that the predictive accuracy of all machine learning models is not particularly high. Upon analysis, we believe that for small-sample datasets, introducing too many features as input can pose challenges during the machine learning training process, thereby affecting predictive accuracy.

In the experiment where important features filtered by RF were used as model inputs, we decided not to use RF for regression testing anymore, as the performance of the RF regression model heavily relies on its assessment of feature importance. Since the most significant features have already been filtered out using the RF model, the potential for performance improvement in the RF regression model with these features has reached saturation, and continuing to use it would not bring significant improvements to model performance. Therefore, we proceeded with experiments using three different networks: BPNN, RBFN and GRNN. The experimental results for the GRNN model were as follows: R^2 = 0.77, MSE = 10.44, RMSE = 3.23, and MAE = 2.20. For

Table 3: Comparison of parameter settings and predictive accuracy among machine learning models.

Model selection		Parameter settings	R^2	MSE	RMSE	MAE
RF		Bootstrap = True Maximum depth = 9 Number of trees =100	0.53	21.8	4.67	3.76
Enter all remote sensing variables	BPNN	Hidden layer dimension =14 Optimizers=Adam Activation functions = Sigmoid Early Stop Method=True Epochs=5000	0.54	21.41	4.63	3.46
	RBFN	Hidden layer dimension =10 Optimizers=Adam Early Stop Method=True Epochs=5000	0.51	22.48	4.74	3.68
	GRNN	Smoothing factor=0.7	0.59	18.92	4.34	3.72
	SSA-GRNN	Number of individuals = 30 Number of iterations = 30 Dimensionality = 1 Lower bound = 0.01 Upper bound = 2	0.61	18.62	4.31	2.57
Enter important remote sensing variables	BPNN	Hidden layer dimension =12 Optimizers=Adam Activation functions = Sigmoid Early Stop Method=True Epochs=5000	0.58	19.26	4.39	3.73
	RBFN	Hidden layer dimension =6 Optimizers=Adam Early Stop Method=True Epochs=5000	0.66	15.45	3.93	3.27
	GRNN	Smoothing factor=0.55	0.77	10.11	3.18	2.42
	SSA-GRNN	Number of individuals = 30 Number of iterations = 30 Dimensionality = 1 Lower bound = 0.01 Upper bound = 2	0.84	7.04	2.65	1.92

the RBFN model, the metrics were R^2 = 0.66, MSE = 15.45, RMSE = 3.93, and MAE = 3.27, which were higher than those of BPNN but considerably lower than GRNN. After using the important remote sensing variables as inputs, the BPNN model showed only a slight improvement in prediction accuracy. The different machine learning algorithms' prediction results are shown in Figure 12, where models using important variables as inputs (BPNN, RBFN, and GRNN) exhibited increases in R^2 of 0.04, 0.15, and 0.16, respectively, compared to models using all variables as inputs.

Figure 13 displays the residual plots for the three machine learning algorithms using both the full set of remote sensing variables and the important remote sensing variables as inputs. To facilitate observation, the test set was arranged in ascending order of actual measurement values after obtaining the prediction results.

It is evident that after training with the selected remote sensing variables as input data, the prediction accuracy of the BPNN improved slightly, while the RBFN and GRNN networks showed a significant improvement. Both the RBFN and GRNN networks use radial basis functions as activation functions and do not require data normalization during input. Therefore, each feature has a considerable impact on prediction accuracy, as supported by a study by Binh Thai et al. (2018), which emphasizes the significant impact of feature selection on machine learning models using radial basis functions as activation functions.

Random Forest effectively selected the most influential features on the overall model and eliminated the interference of redundant data on model accuracy during the modeling process, resulting in a significant improvement in the prediction accuracy of these two models. The BPNN model requires data normalization during training and relies more on adaptive adjustment of feature weights through backpropagation. Therefore, it exhibits some robustness against redundant data, with factors such as the choice of the number of hidden layer nodes (Shen, Wang, & Gao, 2008) and activation functions (Bai, Zhang, & Hao, 2009) being more critical for BPNN performance. Thus, the feature importance selection by Random Forest only marginally improved prediction accuracy for the BP network.

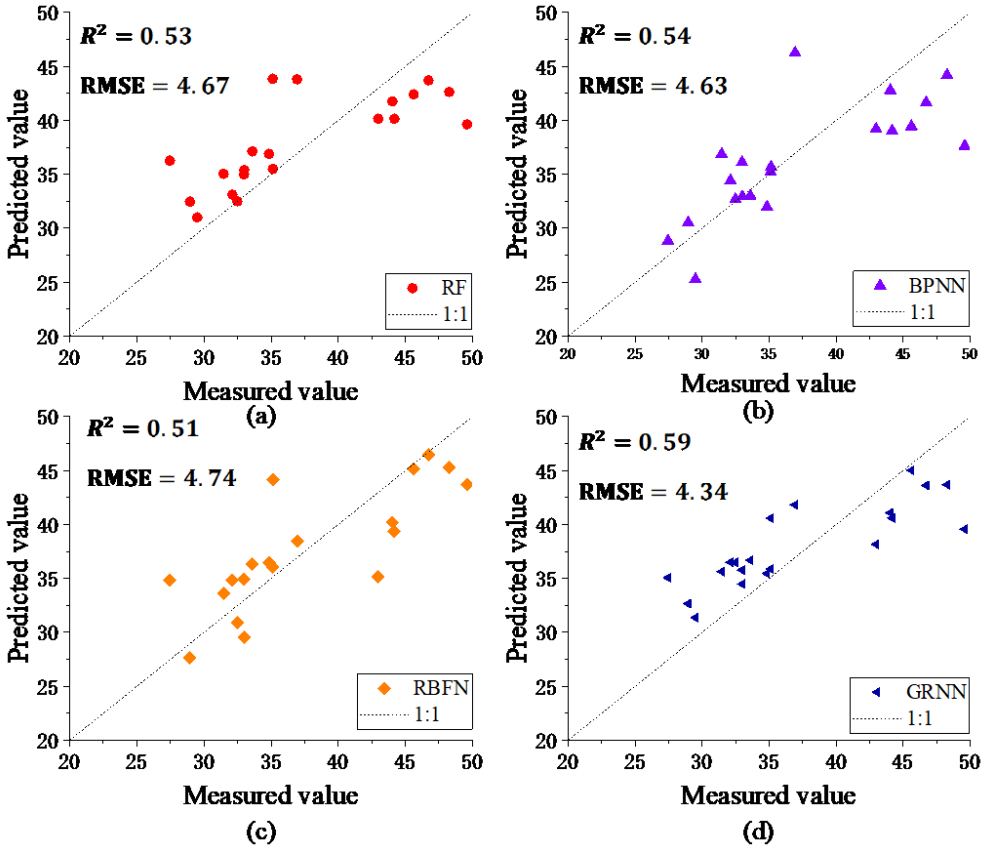


Figure 11: Predictive performance of various machine learning models with all remote sensing variables as inputs. (a) RF, (b) BPNN, (c) RBFN, (d) GRNN.

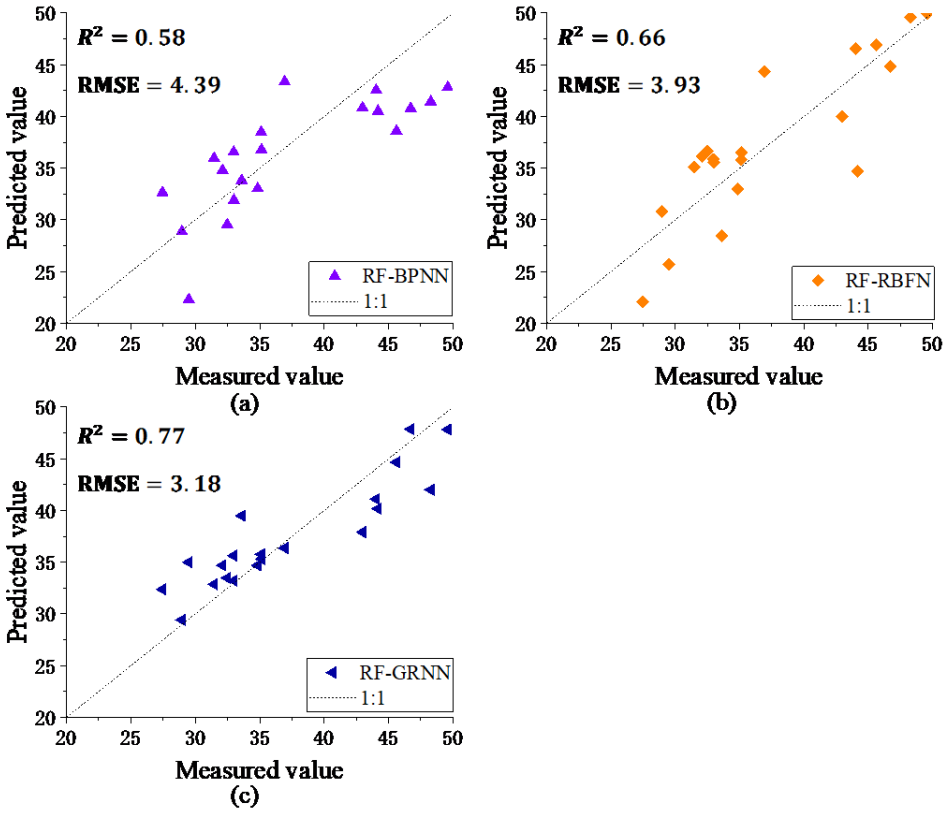


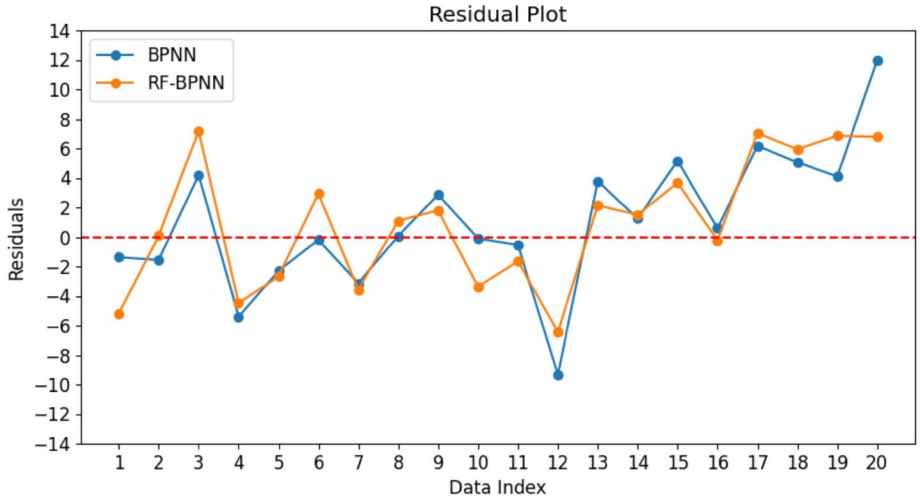
Figure 12: Predictive performance of various machine learning models with important remote sensing variables as inputs. (a) BPNN, (b) RBFN, (c) GRNN.

In previous studies, scholars relied more on individual correlation analysis for selecting remote sensing variables when predicting various plant phenotypic parameters (Chen et al., 2023; Guo et al., 2022). Although this method could identify features related to the target parameters to some extent, it overlooked the interactions and combined effects between variables, potentially leading to the omission of important features or the inclusion of too much unnecessary information. In contrast, the RF feature selection method adopted in this study evaluates the importance of each variable in the decision tree construction process comprehensively. It considers not only the correlation between individual variables and the target variable but also automatically takes into account the interactions between variables. This approach can more effectively identify the feature set that contributes most to the predictive model, thereby improving the model's accuracy and generalizability. A series of experiments also confirmed that this step could reduce the model's training difficulty while simultaneously increasing the high accuracy of machine learning models in predicting the relative chlorophyll content in the tea tree canopy.

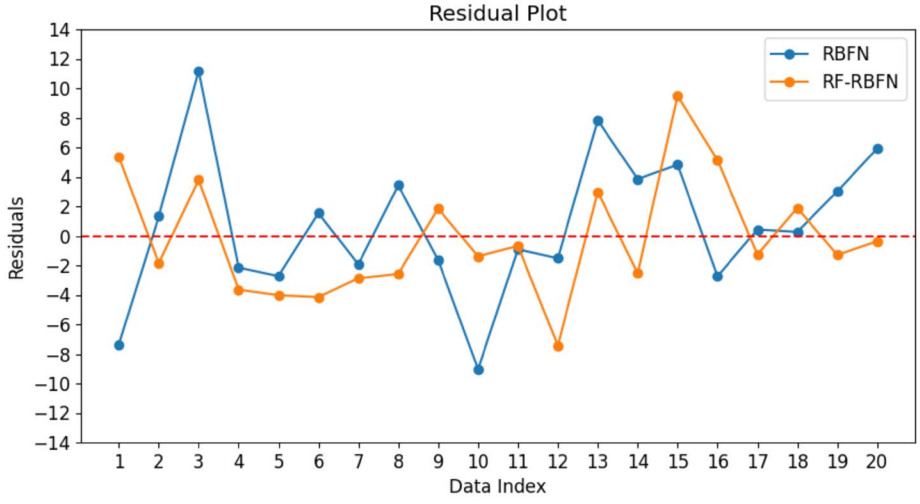
Analysis of prediction performance and stability of the RFSSA-GRNN model for canopy chlorophyll content in tea plants

In recent years, group intelligent algorithms have been widely applied in various fields due to their fast convergence speed and simplicity in computation (Tang, Liu, & Pan, 2021). However, their applications in predicting vegetation parameters using remote sensing data have been relatively limited. In this study, we explored this area by optimizing the previously selected GRNN model with the application of SSA. We used both all remote sensing variables and important remote sensing variables as model inputs to predict the test dataset. The results are shown in Figure 14.

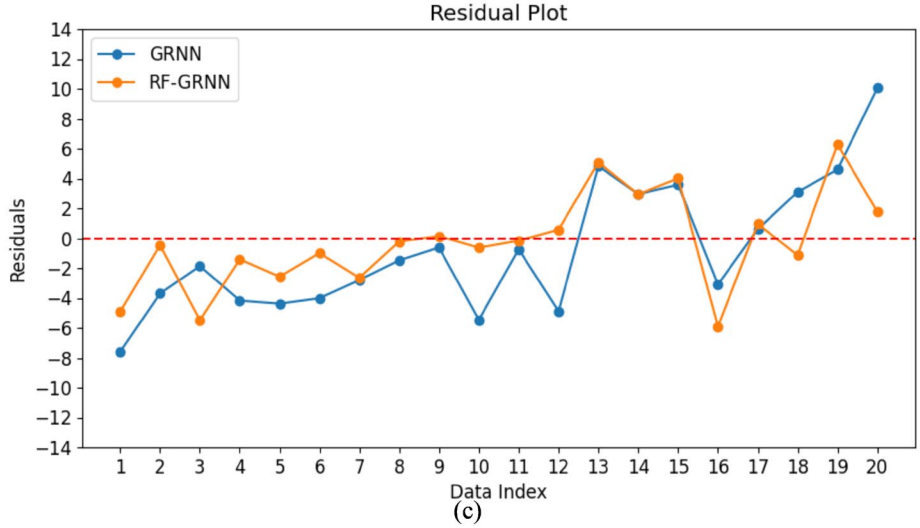
Whether using all remote sensing variables as inputs or important remote sensing variables as inputs, both models achieved higher R^2 values compared to other machine learning models with the same inputs. When using all remote sensing variables as inputs, SSA found the optimal smoothing factor σ to be 0.267. The overall accuracy of the predictive model built with this factor exceeded that of various other machine learning models previously used. The R^2, MSE, RMSE, and MAE were 0.61, 18.62, 4.31, and 2.57, respectively. Compared to the GRNN model with the same remote sensing variable inputs, the R^2 improved by 0.02.



(a)



(b)



(c)

Figure 13: Residual analysis of predictive results for three machine learning models under different input scenarios. (a) BPNN, (b) RBFN, (c) GRNN.

When selecting important remote sensing variables as inputs, the RFSSA-GRNN model demonstrated excellent predictive performance. SSA identified the optimal smoothing factor as 0.174, resulting in R^2, MSE, RMSE, and MAE values of 0.84, 7.04, 2.65, and 1.92, respectively. The R^2 improved by 0.07 compared to the GRNN model using important remote sensing variables as inputs. By comparing the predictive performance on a separate test set, it was found that its predictive accuracy significantly surpasses the results of previous models.

To further validate the predictive performance of the RFSSA-GRNN model, we conducted a comparison with the RF-GRNN and SSA-GRNN models. Figure 15 presents a residual analysis when employing the RFSSA-GRNN model and the RF-GRNN model. When the number of samples to be predicted is small, the smoothing factor chosen by the RF-GRNN is also obtained through trial and error as a better parameter, resulting in the predictive results of the two models produced after training being quite similar. However, when dealing with larger values of samples to be predicted, the incorporation of the SSA algorithm can further optimize the selection of the smoothing factor, enhancing the stability of the model's predictions.

Figure 16 illustrates the residual analysis of the RFSSA-GRNN model and the SSA-GRNN model's predictions. It can be observed that when the relative chlorophyll content to be predicted falls within a moderately medium range, both models achieve good results, and the prediction outcomes are quite similar. This indicates that SSA is capable of assisting the GRNN network in training models with high accuracy under varying input parameters. However, when the SPAD values to be predicted are either extremely high or extremely low, using selected important remote sensing parameters as inputs can still, to some extent, eliminate the interference of redundant features on model training, thereby improving the model's predictive accuracy on the relative chlorophyll content.

The fitness function is one of the crucial metrics for evaluating the network during the process of network training. In this experiment, we used the sum of mean squared errors between the training set and the test set at each training round as the fitness function. Figure 17 illustrates the variations in the fitness function for different remote sensing variable inputs. It can be observed that in both input scenarios, the SSA model reached a stable state by the 7th iteration, demonstrating the stability and rapid convergence of the SSA model during training.

Unlike the trial-and-error approach used to obtain the optimal values for parameters in other models in this experiment, the SSA-GRNN model autonomously selects the best parameters by setting only the dimensionality and upper and lower bounds. This effectively eliminates the extensive manual work required to determine parameters during the model establishment process. Additionally, the fitness function of the SSA model using important variables as input is lower than that of the SSA model using all variables as input. This further confirms that variable selection contributes to achieving better predictive performance in the prediction model.

Intelligent optimization algorithms are currently underutilized in predicting the relative chlorophyll content of plant canopies. Lu et al. (2022) applied the PSO optimization algorithm to enhance the ELM model for predicting the relative chlorophyll content in jujube leaves affected by mite infestation. We have improved upon this by introducing the SSA to optimize the smoothing factor of the GRNN model. The application of the SSA algorithm not only enhances the model's predictive accuracy but also simplifies the complexity and randomness associated with manual tuning, making the optimization process more efficient and stable. This advancement could be crucial for future large-scale remote sensing data prediction models, as it facilitates more automated and intelligent decision support in applications like monitoring the relative chlorophyll content in the canopy of tea trees.

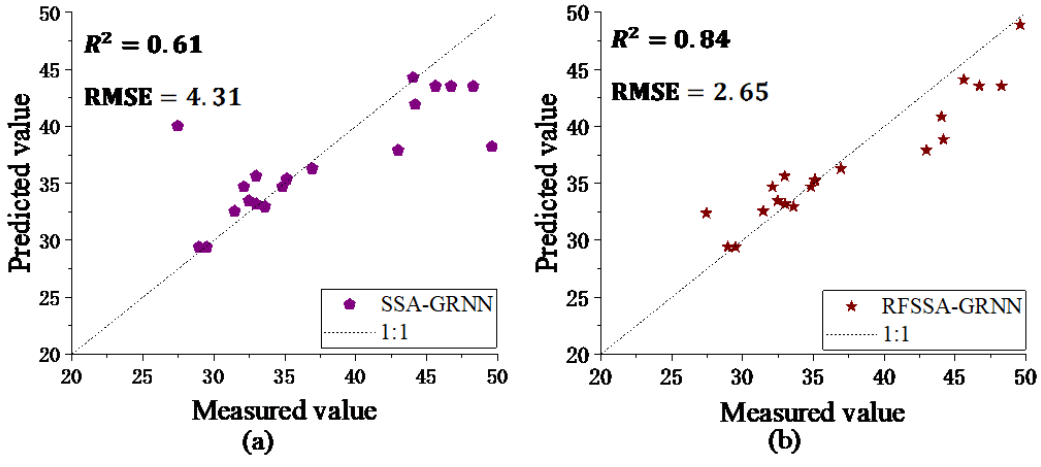


Figure 14: Prediction of chlorophyll values by the SSA-GRNN model. (a) Input with all remote sensing variables. (b) Input with important remote sensing variables.

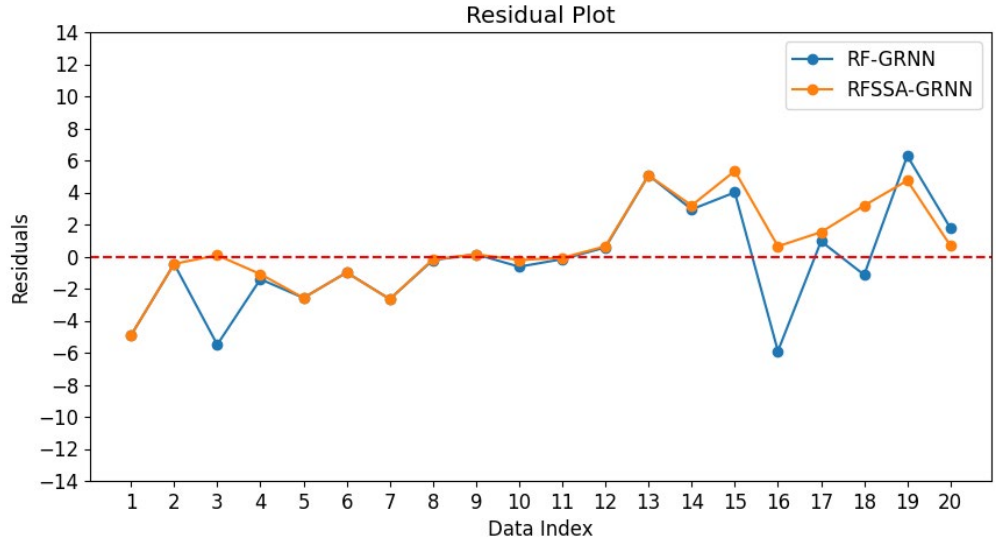


Figure 15: Residual Analysis of GRNN and SSAGRNN Models Using Important Remote Sensing Variables as Inputs.

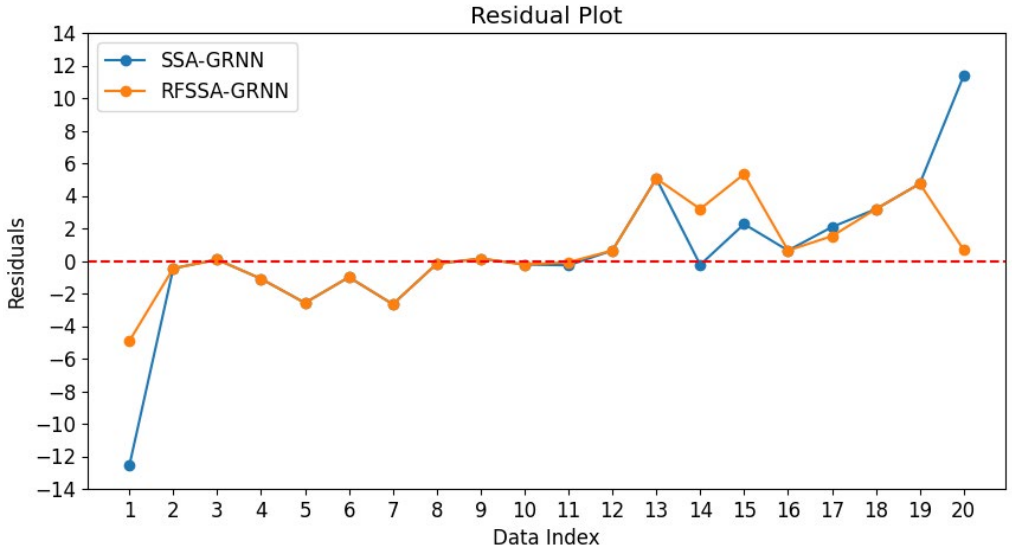


Figure 16: Residual Analysis of SSAGRNN Model Predictions Under Two Different Input Scenarios.

In this study, we introduce a novel approach for rapidly and non-destructively monitoring the growth status of tea trees by combining multispectral RPA imagery with machine learning models to predict the chlorophyll content of the tea tree canopy. The chlorophyll content in tea leaves is closely related to their growth status and the quality of the final product. By developing the RFSSA-GRNN model for predicting the relative chlorophyll content in the tea tree canopy, we can utilize multispectral RPA for non-destructive monitoring of chlorophyll content, facilitating accurate assessments of the health status of tea trees. This enables the timely identification of issues in plant growth, allowing for appropriate management and intervention measures to be taken.

While remote sensing technology has been used to assess the relative chlorophyll content of other plants and monitor the AGB and LAI of tea trees (Yin et al., 2023; Shi et al., 2022), the application of RPA remote sensing imagery for monitoring the relative chlorophyll content of tea crops is not yet common. Wahono et al. (2021) conducted preliminary exploration, predicting the relative chlorophyll content of tea trees using visible light images and linear regression equations. Our study further expands this work by utilizing multispectral images and introducing more complex machine learning algorithms, comparing four different machine learning models. We discovered the potential of the GRNN network for accuracy in prediction,

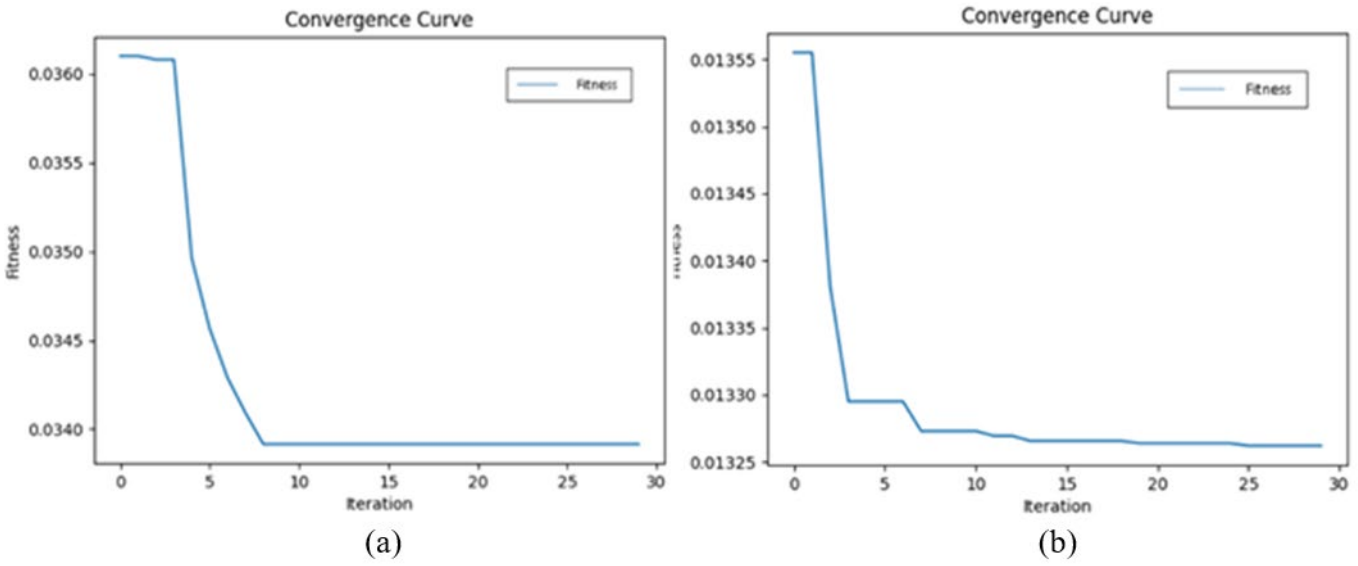


Figure 17: Changes in fitness function during the training process of the SSA-GRNN Model. (a) Input with all remote sensing variables (SSA-GRNN). (b) Input with important remote sensing variables (RFSSA-GRNN).

and based on subsequent RF for remote sensing variable selection and SSA for optimization of the GRNN smoothing factor, we established an RFSSA-GRNN model for predicting the relative chlorophyll content in the tea tree canopy. Experimental results show that this model can accurately predict the relative chlorophyll content in the tea tree canopy in field conditions.

In this study, there are certain limitations to be acknowledged. Firstly, the sample size is relatively small. Although existing experiments have demonstrated the high accuracy of the chlorophyll content prediction in the canopy leaves of the current tea plant variety proposed in this paper, further research is needed to assess the predictive accuracy on different tea plant varieties at different time periods. Additionally, tea plants are often subjected to multiple stresses during their growth, such as nitrogen deficiency and drought (Lv et al., 2021). Therefore, future research could focus on how to use machine learning models to simulate chlorophyll changes when these stresses occur, thus mitigating the damage caused by various stressors to tea plant growth.

Conclusions

Based on RPA multispectral remote sensing data and machine learning, this study compared the prediction effects of four machine learning models and selected GRNN. By integrating the use of RF for remote sensing variable selection and the SSA optimization algorithm to optimize GRNN, the RFSSA-GRNN model for predicting the relative chlorophyll content of the tea tree canopy was proposed. This model achieved good prediction results in experiments, facilitating the rapid and effective monitoring of the growth status of tea trees.

Author Contribution

Conceptual idea: Zhou, Q.Y.; Zhang, Y.H.; Methodology design: Zhang, J.C.; XING, W.; Data collection: Zhang, J.C.; WEI, T.W.; Data analysis and interpretation: Zhou, Q.Y.; Zhang, J.C., and Writing and editing: Zhang, J.C.; WEI, T.W.; Wang, J.

Acknowledgments

We would like to express our gratitude to the Talent Research Program of Anhui Agricultural University for providing financial support under grant numbers rc482003. Additionally, we extend our thanks to Jiangsu Bocha Agricultural Science and Technology Development Co., Ltd. for providing the experimental site and to Mr. Guo Chao from Nanjing Yunyue Information Technology Co., Ltd. for his technical support related to RPA.

References

Bai, Y., Zhang, H.-X., & Hao, Y. (2009). The performance of the backpropagation algorithm with varying slope of the activation function. *Chaos, Solitons & Fractals*, *40*:69-77.

Binh Thai, P. et al. (2018). A hybrid machine learning ensemble approach based on a radial basis function neural network and rotation forest for landslide susceptibility modeling: A case study in the himalayan area, India. *International Journal of Sediment Research*, *33*(2):157-170.

Broge, N. H., Leblanc, E. (2001). Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. *Remote Sensing of Environment*, *76*(2):156-172.

Carmona, F., Rivas, R., & Fonnegra, D. C. (2015). Vegetation index to estimate chlorophyll content from multispectral remote sensing data. *European Journal of Remote Sensing*, *48*(1):319-326.

Chen, J. M. (1996). Evaluation of vegetation indices and a modified simple ratio for boreal applications. *Canadian Journal of Remote Sensing*, *22*(3):229-242.

Chen, S. et al. (2020). Retrieval of cotton plant water content by uav-based vegetation supply water index (vswi). *International Journal of Remote Sensing*, *41*(11):4389-4407.

Chen, X. et al. (2023). Estimation of winter wheat canopy chlorophyll content based on canopy spectral transformation and machine learning method. *Agronomy*, *13*(3):783.

Clevers, J., Kooistra, L., & Van Den Brande, M. (2017). Using sentinel-2 data for retrieving lai and leaf and canopy chlorophyll content of a potato crop. *Remote Sensing*, *9*(5):405.

Curran, P. (1983). Multispectral remote sensing for the estimation of green leaf area index. *Philosophical Transactions of the Royal Society of London Series A, Mathematical and Physical Sciences*, *309*(1508):257-270.

Daughtry, C. S. T. et al. (2000). Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. *Remote Sensing of Environment*, *74*(2):229-239.

Epifanio, I. (2017). Intervention in prediction measure: A new approach to assessing variable importance for random forests. *BMC Bioinformatics*, *18*(1):230.

Gano, B. et al. (2021). Using uav borne, multi-spectral imaging for the field phenotyping of shoot biomass, leaf area index and height of west african sorghum varieties under two contrasted water conditions. *Agronomy-Basel*, *11*(5):850.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14):2225-2236.

Gitelson, A. A. (2013). Remote estimation of crop fractional vegetation cover: The use of noise equivalent as an indicator of performance of vegetation indices. *International Journal of Remote Sensing*, *34*(17):6054-6066.

Gitelson, A. A., Gritz, Y., & Merzlyak, M. N. (2003). Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *Journal of Plant Physiology*, *160*(3):271-282.

Gitelson, A. A., Merzlyak, M. N. (1997). Remote estimation of chlorophyll content in higher plant leaves. *International Journal of Remote Sensing*, *18*(12):2691-2697.

Gong, P. et al. (2003). Estimation of forest leaf area index using vegetation indices derived from hyperion hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(6):1355-1362.

Guo, Y. et al. (2022). Machine learning-based approaches for predicting spad values of maize using multi-spectral images. *Remote Sensing*, *14*(6):1337.

Haboudane, D. et al. (2004). Hyperspectral vegetation indices and novel algorithms for predicting green lai of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sensing of Environment*, *90*(3):337-352.

Haboudane, D. et al. (2002). Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote Sensing of Environment*, *81*(2-3):416-426.

Hassan, M. A. et al. (2018). Time-series multispectral indices from unmanned aerial vehicle imagery reveal senescence rate in bread wheat. *Remote Sensing*, *10*(6):809.

Huete, A. R. (1988). A soil-adjusted vegetation index savi. *Remote Sensing of Environment*, *25*(3):295-310.

Izonin, I. et al. (2021). A grnn-based approach towards prediction from small datasets in medical application. *Procedia Computer Science*, *184*:242-249.

Jiang, Z. et al. (2008). Development of a two-band enhanced vegetation index without a blue band. *Remote Sensing of Environment*, *112*(10):3833-3845.

Krause, G. H., & Weis, E. (1991). Chlorophyll fluorescence and photosynthesis: The basics. *Annual Review of Plant Physiology and Plant Molecular Biology*, *42*(1):313-349.

Liu, J., Pattey, E., & Jego, G. (2012). Assessment of vegetation indices for regional crop green lai estimation from landsat images over multiple growing seasons. *Remote Sensing of Environment*, *123*:347-358.

Liu, Z. A., Yang, J. P., & Yang, Z. C. (2012). Using a chlorophyll meter to estimate tea leaf chlorophyll and nitrogen contents. *Journal of Soil Science And Plant Nutrition*, *12*(2):339-348.

Lu, J. et al. (2022). Inversion of chlorophyll content under the stress of leaf mite for jujube based on model pso-elm method. *Frontiers in Plant Science*, *13*:1009630.

Lv, Z. et al. (2021). Research progress on the response of tea catechins to drought stress. *Journal of the Science of Food and Agriculture*, *101*(13):5305-5313.

Martínez, D., & Guiamet, J. (2004). Distortion of the spad 502 chlorophyll meter readings by changes in irradiance and leaf water status. *Agronomie*, *24*(1):41-46.

Meng, Q. (2023). Evaluation technology of urban green space with remote sensing. In Q. Meng. *Remote sensing of urban green space*. Singapore: Springer Nature Singapore, (pp.207-237).

Miao, S. et al. (2022). Extraction methods, physiological activities and high value applications of tea residue and its active components: A review. *Critical Reviews in Food Science and Nutrition*, *63*(33):12150-12168.

Noh, H. et al. (2006). A neural network model of maize crop nitrogen stress assessment for a multi-spectral imaging sensor. *Biosystems Engineering*, *94*(4):477-485.

Pan, S. Y. et al. (2022). Tea and tea drinking: China's outstanding contributions to the mankind. *Chinese Medicine*, *17*(1):27.

Raper, T. B., & Varco, J. J. (2015). Canopy-scale wavelength and vegetative index sensitivities to cotton growth parameters and nitrogen status. *Precision Agriculture*, *16*(1):62-76.

Rondeaux, G., Steven, M., & Baret, F. (1996). Optimization of soil-adjusted vegetation indices. *Remote Sensing of Environment*, 55(2):95-107.

Roujean, J.-L., & Breon, F.-M. (1995). Estimating par absorbed by vegetation from bidirectional reflectance measurements. *Remote Sensing of Environment*, *51*(3):375-384.

Rouse, J. W. et al. (1974). Monitoring vegetation systems in the great plains with erts. *NASA. Goddard Space Flight Center 3d ERTS-1 Symptoms*, *351*(1):309.

Shen, H. Y., Wang, Z., & Gao, C. Y. (2008). Determining the number of bp neural network hidden layer units. *Journal of Tianjin University of Technology*, *24*:13-15, 2008.

Shi, Y. et al. (2022). Using unmanned aerial vehicle-based multispectral image data to monitor the growth of intercropping crops in tea plantation. *Frontiers in Plant Science*, *13*:820585.

Sonobe, R., Sano, T., & Horie, H. (2018). Using spectral reflectance to estimate leaf chlorophyll content of tea with shading treatments. *Biosystems Engineering*, *175*:168-182.

Specht, D. F. (1991). A general regression neural network. *IEEE Trans Neural Netw*, *2*(6):568-576.

Tang, J., Liu, G., & Pan, Q. (2021). A review on representative swarm intelligence algorithms for solving optimization problems: Applications and trends. *IEEE/CAA Journal of Automatica Sinica*, *8*(10):1627-1643.

Uddling, J. et al. (2007). Evaluating the relationship between leaf chlorophyll concentration and spad-502 chlorophyll meter readings. *Photosynthesis Research*, *91*(1):37-46.

Verrelst, J. et al. (2008). Angular sensitivity analysis of vegetation indices derived from chris/proba data. *Remote Sensing of Environment*, *112*(5):2341-2353.

Vincini, M., Frazzi, E., & D'alessio, P. (2008). A broad-band leaf chlorophyll vegetation index at the canopy scale. *Precision Agriculture*, *9*(5):303-319.

Wahono, D. I. et al. (2021). Comparing visible light based vegetation index and chlorophyll meter to estimate chlorophyll and nitrogen content of tea (*camellia sinensis* l. Kuntze) leaves. *Annals of the Romanian Society for Cell Biology*, *25*(1):5033-5043.

Walsh, O. S. et al. (2018). Assessment of uav based vegetation indices for nitrogen concentration estimation in spring wheat. *Advances in Remote Sensing*, *7*(2):19.

Wang, J. et al. (2023). Research on rapid and low-cost spectral device for the estimation of the quality attributes of tea tree leaves. *Sensors*, 23(2):571.

Wang, K. et al. (2010). Analysis of chemical components in oolong tea in relation to perceived quality. *International Journal of Food Science and Technology*, *45*(5):913-920.

Wang, L. F. et al. (2004). The compounds contributing to the greenness of green tea. *Journal of Food Science*, *69*(8):S301-S305.

Wang, Y. et al. (2019). Rapid prediction of chlorophylls and carotenoids content in tea leaves under different levels of nitrogen application based on hyperspectral imaging. *Journal of the Science of Food and Agriculture*, *99*(4):1997-2004.

Xiao, Q., Mcpherson, E. G. (2005). Tree health mapping with multispectral remote sensing data at uc davis, california. *Urban Ecosystems*, 8(3-4):349-361.

Xiao, Y. F. et al. (2014). Sensitivity analysis of vegetation reflectance to biochemical and biophysical variables at leaf, canopy, and regional scales. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(7):4014-4024.

Xue, J., & Shen, B. (2020). A novel swarm intelligence optimization approach: sparrow search algorithm. *Systems Science & Control Engineering*, *8*(1):22-34.

Yin, H. et al. (2023). Multi-temporal uav imaging-based mapping of chlorophyll content in potato crop. *PFG-Journal of Photogrammetry Remote Sensing and Geoinformation Science*, *91*(2):91-106.

Zhang, L. G. et al. (2018). Density weighted connectivity of grass pixels in image frames for biomass estimation. *Expert Systems With Applications*, *101*:213-227.

Zhu, Y. et al. (2022). Image classification method of cashmere and wool based on the multi-feature selection and random forest method. *Textile Research Journal*, *92*(7-8):1012-1025.