



# Spectroscopic determination of chlorophyll content in sugarcane leaves for drought stress detection

Jingyao Gai<sup>1</sup> · Jingyong Wang<sup>1</sup> · Sasa Xie<sup>2</sup> · Lirong Xiang<sup>3</sup> · Ziting Wang<sup>1,4</sup>

Accepted: 19 September 2023 / Published online: 13 November 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

Drought is a major abiotic stress that affects the productivity of sugarcane worldwide. Water deficiency during sugarcane growth will lead to a reduction in leaf pigment content, such as chlorophyll, known as chlorosis. Although changes in spectral reflectance signature were identified a conspicuous sign of chlorophyll content changes caused by drought stress, the quantitative relationships between leaf chlorophyll content and spectral reflection signatures are still poorly explored. In this study, we present our contribution in systematically establishing a model for estimating leaf chlorophyll content in drought-affected sugarcane using VIS/NIR reflectance spectroscopy and characteristic band extraction techniques. Leaves of sugarcane plants at early elongation stage under different controlled irrigation conditions were used for spectra data collection, and the chlorophyll contents were collected with standard analytical methods. Different characteristic band extraction techniques and regression models were compared and discussed to obtain a chlorophyll content estimation model with the best performance. As the quantitative results, the combination of characteristic bands extracted by the successive projection algorithm (SPA) with a Stacking regression model achieved a high chlorophyll content estimation performance ( $R^2=0.9834$ ,  $RMSE=0.0544$  mg/cm<sup>2</sup>) with only 4.3% of original spectral variables as inputs. This study provides a theoretical basis for accurate and non-invasive drought stress level estimation in large-scale cultivation.

**Keywords** VIS/NIR spectroscopy · Sugarcane · Leaf chlorophyll content · Machine learning · Drought stress

## Introduction

Sugarcane (*Saccharum Officinarum* L.) is the largest sugar crop and the second largest bioenergy crop in the world. Most of the world's sugarcane is grown in tropical regions of countries including Brazil, India, and China, where precipitation is extremely unstable and the sugarcane plants can suffer from drought stress (Inman-Bamber & Smith, 2005). Among the five major developmental stages of sugarcane growth: sprouting, seedling, tillering, elongation (i.e. nutritional growth stage), and maturity, the

---

Jingyong Wang and Sasa Xie have contributed equally to this work.

---

Extended author information available on the last page of the article

elongation stage is unanimously considered to be the key growth stage (Yadav et al., 2020; Zhao & Li, 2015). At elongation stage, drought stress leads to a series of physiological disorders such as discoloration, folding, tillers and leaf area reduction of sugarcane leaves, which eventually reduces the yield and quality of sugar. The biological mechanism of drought response is that under drought conditions, the stomatal conductance of leaves decreases, which hinders the synthesis of chlorophyll and accelerates the decomposition of chlorophyll, thus hindering photosynthesis in leaves. Therefore, early detection of drought stress can provide theoretical guidance for the diagnosis of crop nutrient conditions as well as nutrient management in the field.

Chlorophyll content is considered to be an indicator of drought stress (as well as nitrogen stress, disease stress, etc.) (Peng et al., 2019). Although high chlorophyll content measurement accuracy can be achieved with traditional chemical methods, they are often complex, costly, and time-consuming, making them unsuitable for large-scale plant monitoring in the field. Recently developed spectroscopic technologies offer the advantages of non-destructive, time efficient and highly automated in detecting characterizing material structures and compositions (DAIE, 1992; Gitelson & Merzlyak, 1996). For most plants species, decreased chlorophyll content in leaves were found leads to the spectral reflection pattern changes (Zarco-Tejada et al., 2000), i.e. an increase in overall spectral reflectance; an increase in the proportion of reflected red and blue light, and a decrease in green light.

Scientists have extensively investigated rapid, non-destructive chlorophyll content measurement techniques based on plant spectroscopy and spectral analysis technology (Inoue et al., 2016; Wang et al., 2018). The existing spectroscopic-based chlorophyll content measurement techniques can be divided into three categories: the first category uses leaf or canopy radiation transmission models, including PROSAIL, SAIL, PROSPECT, etc. (Jacquemoud & Baret, 1990; Verhoef, 1984). Although a high measurement accuracy can be achieved, it is less frequently used due to the complexity of the radiation transmission model.

The second category for chlorophyll content estimation is using vegetation indices (VI), which were defined as linear combination or ratios of the reflectance intensities in two or three bands (Sims & Gamon, 2002). Due to the effectiveness in minimizing the potential influence of scale factors and enhancing spectral features of vegetation, VIs were widely adopted in remote sensing and photogrammetry for plant science research, especially in crop growth information monitoring. Many VIs have been reported to be effective in estimating biochemical parameters of plants. For instance, Normalized difference vegetation index (NDVI), Soil-Adjusted vegetation index (SAVI) and Ratio vegetation index (RVI) were reported effective in chlorophyll content estimation of maize (Chaoyang et al., 2010). There are studies focusing on clarifying the response differences of VIs to environmental disturbances of various kinds, therefor screening and combining different VIs to construct more robust chlorophyll content estimation models. For instance, Li et al. (2015) constructed a model for chlorophyll content estimation of sugarcane leaves at tillering stage. NDVI and red edge position (REP) were selected as variables, and a determination coefficient of 0.7386 was achieved. Qiao et al. (2022) constructed a partial least square (PLS) model for chlorophyll content estimation for maize crop using remote sensing data. The model used VISs screened with rand frog method (RFM) as variables, and achieved an *RMSE* of 2.361 mg/L. And Zhang et al. (2022) extracted spectral indices from hyperspectral images, and achieved a  $R^2$  ranging from 0.77 to 0.78 for the prediction of leaf chlorophyll content. However, most of the VI-based algorithms were designed for remote sensing data with limited wavebands available, and the dimensionality reduction nature of VIs may

discard significant information reflecting the physiological effects corresponding to chlorophyll content changes, thus decrease the estimation accuracy.

The third category uses machine learning technology to establish regression models of spectral reflectance signatures and plant physiological parameters using full spectra. In recent years, different algorithms have been developed to combine spectral reflectance features with machine learning techniques for the estimation of leaf physiological parameters, including chlorophyll content, moisture content, nitrogen content, etc. (Krishna et al., 2019; Liu et al., 2020; Wang et al., 2020). Typically, the collected spectra are in large quantities and also contain noisy, redundant and interfering information. Therefore, the spectra are usually pre-processed first to remove the noise. Commonly used pre-processing methods include: Savitzky–Golay smoothing (SG) (Yoosefzadeh-Najafabadi et al., 2021), multiplicative scatter correction (MSC) (Sonobe et al., 2021), standard normal variate (SNV) (Yamashita et al., 2020). Then, in order to eliminate redundant and interfering information, dimensionality reduction algorithms are usually applied in the form of extracting characteristic bands (El-Hendawy et al., 2021; Hoepfner et al., 2020; Liang et al., 2018; Wang et al., 2019; Zhu et al., 2020), which are less correlated and preserve majority of useful information in the estimation of leaf physiological parameters. Commonly used characteristic band selection methods include correlation coefficient method, successive projections algorithm (SPA) (Xuan et al., 2021), and random forest (RF) (Abdel-Rahman et al., 2013), etc. Finally, the extracted characteristic bands are used as the input of machine learning to predict the physiological parameters of plants. Commonly used machine learning methods include Back Propagation Neural Networks (BPNN) (Khoshhal & Mokarram, 2012), K-Nearest Neighbors (KNN) (Osco et al., 2020), Support Vector Regression (SVR) (Bhadra et al., 2020), etc.

Although significant progress has been made in the study of drought monitoring and parameter inversion using spectral data for different plant species, the spectral response of sugarcane plants to drought during the early elongation stage is seldom studied, and the quantitative relationships between leaf chlorophyll content and spectral reflection properties of sugarcane plants are still poorly discovered.

In this paper, we present our work on non-destructive estimation of chlorophyll content of sugarcane leaves under drought stress using VIS/NIR spectroscopy. The contribution of this study is that we explored the quantitative relationships between leaf chlorophyll content and spectral reflection properties of sugarcane plants under drought stress, which has not been extensively investigated before. Another methodological contribution is that we systematically investigated the performance of different existing pre-processing, dimensionality reduction, and regression techniques in establishing a model for estimating leaf chlorophyll for sugarcane plants under drought stress, which has not been reported in other studies. The specific objectives of this study were to:

- (1) Analyze the spectral changes of sugarcane leaves in the early elongation stage under different drought stress levels
- (2) Explore the optimal characteristic spectral bands and regression model for spectroscopy-based chlorophyll content estimation, and establish an optimal estimation model
- (3) Evaluate the performance of the proposed chlorophyll content estimation algorithm, and compare the performance of the proposed model with conventional VI-based methods
- (4) Validate generalizability of the proposed model, as well as the potential for monitoring the degree of drought stress in sugarcane.

## Experimental materials and data collection

### Experimental materials and experimental design

The experimental materials in this study were prepared at the State Key Laboratory of Protection and Utilization of Subtropical Agricultural Biological Resources of Guangxi University. Four sugarcane cultivars were involved, namely Zhongzhe No. 2 (ZZ2), Zhongzhe No. 9 (ZZ9), TD11 (i.e., GN18, use TD instead in this article), Funong 95-1702 (FN). Among them, both cultivars ZZ2 and ZZ9 were from the same parents (ROC25 × Yuncan 89-7), with ZZ2 is a relatively drought-sensitive variety whereas ZZ9 is a drought-resistant variety. FN is a drought-sensitive wild-type variety. TD is a transgenic strain as the recipient of the Ea-DREB2B gene introduced from FN, which exhibits greater drought tolerance with higher photosynthetic capacity and ROS scavenging ability relative to FN.

Plants of each cultivar were divided into five groups including four drought treatment groups (W1, W2, W3, W4) and one control group (CK). CK indicates normal water treatment (80% of the field capacity), W1 indicates mild drought (60% of the field capacity), W2 indicates moderate drought (45% of the field capacity), W3 indicates severe drought (35% of the field capacity), and W4 indicates extreme drought (less than 30% of the field capacity).

The experiments were carried out at the teaching and research base of the College of Agronomy, Guangxi University, China (22°51'N, 180°17'E) from September to November, 2021. As the pre-planting operations, the sugarcanes were cut into single-bud stems, soaked in carbendazim for 30 min, and cultivated at 28 °C for germination. Stems with consistent budding were selected for transplantation.

The plants was divided into two experiments, experiment 1 was to transplant stems into plastic pots and placed in a greenhouse, and experiment 2 was to transplant stems into an experimental field consisted of 20 plots, in which one plot was designed for each cultivar in each experimental group, respectively. The plants in both experiments were exposed to natural variations on photoperiod throughout the experiment, and the soil moisture was controlled according to the assigned group. For experiment 1, and the soil moisture was controlled daily by weighing the pots and replenishing them to maintain the target field holding capacity. For experiment 2, the soil moisture content was monitored daily through a soil moisture meter (TDR150, Spectrum Technologies, Inc, USA), and the field was replenished when the soil moisture content fell below the target value of the group.

### Measurement methods

From all sugarcane plants involved, leaves were sampled for data acquisition during the elongation stage. For both experiments, 4 plants were selected for each cultivar and each group for data acquisition. For experiment 1, the top 5 leaves of each sugarcane plant were collected for data acquisition, resulting in a total of (5 leaves) × (4 plants) × (4 cultivars) × (5 groups) = 400 samples for sugarcane plants grown in greenhouse. For experiment 2, the top 3 leaves were collected, resulting in a total of (3 leaves) × (4 plants) × (4 cultivars) × (5 groups) = 240 samples for sugarcane plants grown in the field. The leaf VIS/NIR spectral data and the corresponding leaf chlorophyll content were measured right after leaf collection.

## VIS/NIR spectral data acquisition

A CID portable CI-710 optical fiber spectrometer (CID Bio-Science Inc., Camas, WA, USA) (Fig. 1) was used for VIS/NIR spectral data acquisition. The sensor is capable of measuring the reflectance spectrum between 400 and 950 nm with an optical resolution of 2.0 nm. The diameter of the measuring chamber is 7.6 mm. A blue LED and an incandescent lamp were used as the radiation sources, with a leaf clipping mechanism to protect the measurement area from the environmental light interference. Before spectral data collection, the sensor was calibrated using a whiteboard. During data acquisition, measurements were taken at 4 randomly selected positions of the leaves. The measurements were repeated 10 times at each position, and the average vector was taken as the reflection spectrum of the leaf.

## Chlorophyll content determination

After completing the leaf spectral measurement, the chlorophyll content of the leaf was determined by laboratory chemical analysis. A standard protocol was followed (Lichtenthaler & Wellburn, 1983): firstly, the surface of the sampled leaf was cleaned by wiping the dirt. The leaf sample (with midrib removed) was then sheared and placed in a 25 mL capacity flask. Then 20 mL of 80% acetone was added to the flask. The samples were placed in a dark environment and soaked until the tissue was white, and then the volume was brought to 25 mL with dip reagent and shaken thoroughly. The dip reagent was used as a reference for the absorbance of the blank, then the absorbance value at 663 nm and 645 nm ( $OD_{663}$ ,  $OD_{645}$ ) were taken by a UV–Vis spectrophotometer (UV-1800, Shimadzu Corporation, Kyoto, Japan) for colorimetric analysis. The chlorophyll content was calculated as:

$$\text{Chlorophyll content (mg/cm}^2\text{)} = (8.02 \times OD_{663} + 20.21 \times OD_{645}) \times V / (S \times 1000), \quad (1)$$

where  $V$  is the liquid volume, and  $S$  is the area of the sampled leaf.

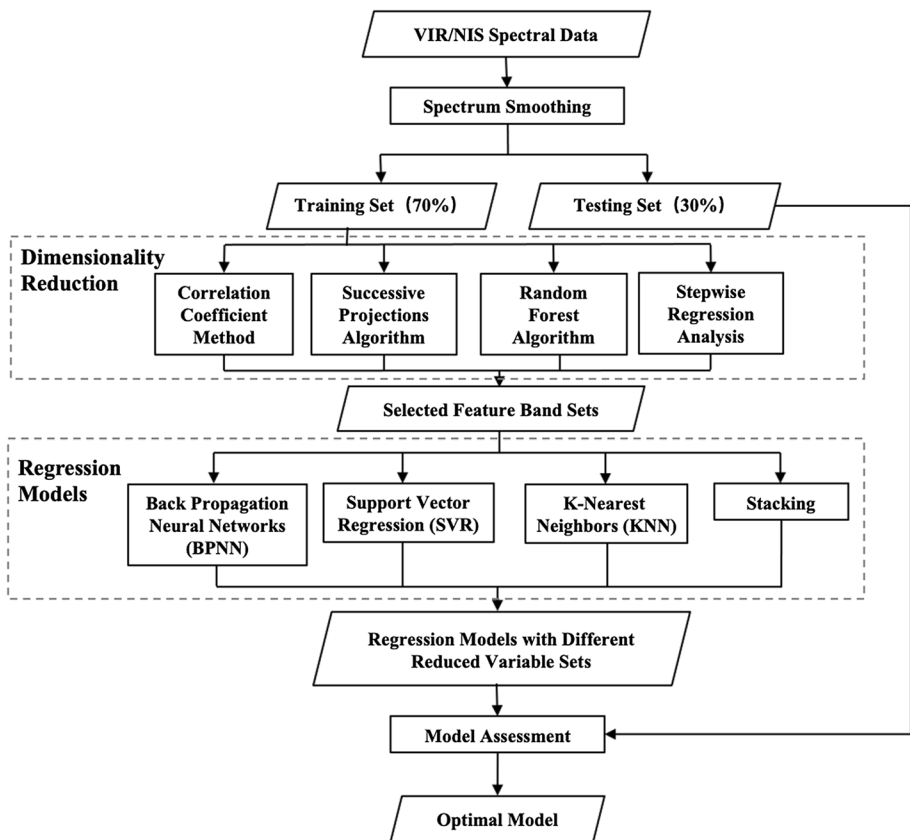
**Fig. 1** The portable CI-710 fiber spectrometer to measure the reflectance spectrum of a sugarcane leaf



## Data processing

In this study, an algorithmic framework consisting of pre-processing, dimensionality reduction, and regression steps was used to process the collected spectrums. Firstly, as a pre-processing step, a smoothing algorithm was applied to the obtained spectra to remove the noise caused by human operations, environmental and instrumental factors. Secondly, the dimensionality of the spectra data was reduced using feature extraction techniques. Dimensionality reduction is necessary because each measured reflectance spectrum contains hundreds of values and exhibits a high degree of redundancy. Incorporating all the data into a prediction model leads to severe overfitting problems. Finally, a regression model was established using machine learning techniques for the estimation of chlorophyll content.

The technical procedure for constructing the optimal model for estimation of sugarcane leaf chlorophyll content is shown in Fig. 2, in which different combinations of dimensionality reduction methods and regression models were constructed. Then these combinations were systematically evaluated and compared before the optimal chlorophyll estimation model was finally selected.



**Fig. 2** The technical procedure for constructing the optimal model for estimation of sugarcane leaf chlorophyll content

Microsoft Excel 2019 was used to organize the data, Origin 2019 was used for statistical analysis and plotting, and Python with “Scikit-learn” library (Pedregosa et al., 2011) was used to implement the data processing algorithms.

## Spectral data preprocessing

The measurement noise, which appears as glitches on the collected spectrum (Fig. 3a), can negatively affect the accuracy of the estimation. Smoothing algorithm is one of the most commonly used methods to eliminate noise. Savitzky–Golay (SG) is a convolutional smoothing method that performs polynomial least squares fit to the spectral data in a moving window manner to eliminate glitches (Fig. 3b). The SG convolutional smoothing algorithm is:

$$x_{a\_smooth} = \frac{1}{H} \sum_{i=-w}^{+w} x_{a+i} h_i, \quad (2)$$

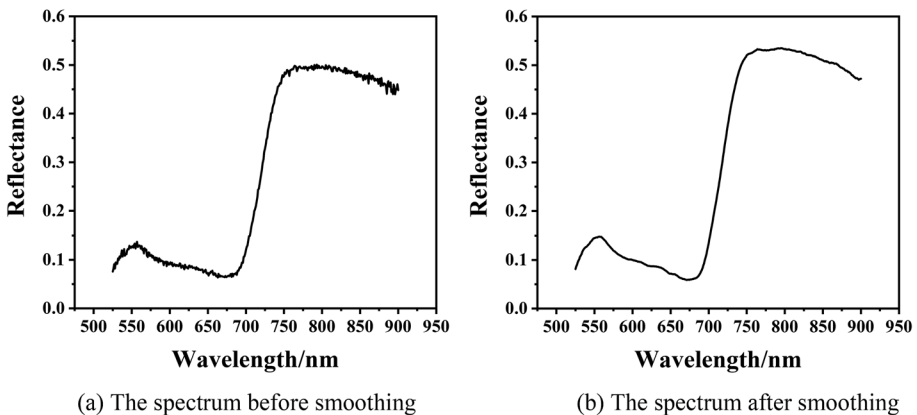
where  $h_i$  is the smoothing coefficient and obtained by least squares fitting.  $H$  is the normalization coefficient.

## Dimensionality reduction and feature band extraction

In this study, multiple characteristic spectral bands extraction algorithms was applied to the pre-processed spectral data in order to reduce the data dimension and reduce the data redundancy.

Pearson correlation coefficient method (Wang et al., 2017): The correlation coefficient between the reflectance of each band in the spectrum and the content of a specific substance, e.g. chlorophyll content, was calculated as follows:

$$R_j = \frac{\sum_{i=1}^N (x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^N (y_i - \bar{y})^2}}, \quad (3)$$



**Fig. 3** Comparison before and after spectrum smoothing

where  $x_{ij}$ ,  $y_i$  are the reflectance value of the  $j$ th band and the chlorophyll content in the  $i$ th spectrum data sample, respectively, and  $\bar{x}_j$ ,  $\bar{y}$  are the mean of the  $j$ th spectral reflectance values and the mean of chlorophyll contents of all the spectrum data samples, respectively. A higher absolute value of correlation coefficient indicates a higher explanatory power of a certain spectral band in the estimation of chlorophyll content. Therefore the spectral band values with significant correlations (i.e.  $|R| \geq 0.8$ ) with chlorophyll content were selected as the characteristic bands (Källén, 2016).

Successive projections algorithm (SPA) (Araújo et al., 2001): SPA is a forward feature selection method for solving collinearity issues. In characteristic bands selection, starting from a randomly selected band, a new band with minimum information redundancy was iteratively selected until certain number of bands were selected. Projection analysis was used to evaluate the information redundancy of a certain band with respect to the selected bands. By comparing the magnitude of the projected vectors onto the orthogonal complement space of the selected bands, the band with the largest projected vector magnitude was selected as a new characteristic band. With different SPA-selected band sets started with different initial bands, a partial least square regression model was applied to the training set to compare these sets and select the best combination. The flow of the SPA algorithm is expressed as follows:

Algorithm 1: successive projections algorithm.

**Input:** target characteristic bands number  $N$ , number of spectrum samples  $M$ , number of bands  $J$ , spectra data matrix  $X$ ,

**Output:** the indices list of the selected characteristic bands  $k$ .

Step 1: Let  $n = 1$ , randomly select a column of the spectral matrix  $X$  (e.g., the  $j$ th column,  $1 \leq j \leq J$ ) into the selected band set, assign the  $j$ th column to  $\mathbf{x}_{k(n-1)}$ ,  $\mathbf{k}(n-1) = j$ .

Step 2: The set of unselected column vector positions is denoted as  $S$ ,

$$S = \{i, 1 \leq i \leq J, i \notin \{\mathbf{k}(0), \dots, \mathbf{k}(n-1)\}\}$$

Step 3: Calculate  $\mathbf{p}_{x_i}$ , the projection of  $\mathbf{x}_{k(n-1)}$  to the remaining column vectors.

$$\mathbf{p}_{x_i} = \mathbf{x}_i - (\mathbf{x}_i^T \mathbf{x}_{k(n-1)}) \mathbf{x}_{k(n-1)} (\mathbf{x}_{k(n-1)}^T \mathbf{x}_{k(n-1)})^{-1}, i \in S$$

Step 4: Extract the spectral band index with the maximum projection vector length, and insert into selected bands set:

$$\mathbf{k}(n+1) = \operatorname{argmax}_{i \in S} (\|\mathbf{p}_i\|)$$

Step 5: Let  $n = n + 1$  and  $\mathbf{x}_j = \mathbf{p}_{x_i}$ , if  $n < N$ , then return to the second step.

Random Forest algorithm (RF) (Bylander, 2002): RF is a model capable of identifying the relative importance of each characteristic band in the prediction. If there is correlation between high-dimensional characteristic bands, the contribution to the regression is not significant and it reduces the computational efficiency. The Gini coefficient can be used as an evaluation criterion to determine the contribution of each input feature to the random forest-based regression. A higher coefficient indicates a higher feature importance.

Stepwise regression analysis (SR) (Greenland, 1989): SR is an iterative characteristic band selection method. Firstly, arbitrary bands were selected to establish an initial regression model. Then, an  $F$  test was performed on each band in the regression equation to test their correlation significance to the chlorophyll content. Bands with insignificant correlation were eliminated. After elimination,  $F$  tests were performed on a new set of candidate bands. Bands tested for significantly correlation with chlorophyll content were selected for



inclusion in the regression model. The iterations of elimination and selection continued until no new bands were selected.

## Regression models, evaluation criterions and dataset partitioning

After dimensionality reduction of the spectrum data, machine learning-based regression models were used to estimate the chlorophyll content of sugarcane leaves. In this study, the performance of several commonly-used regression models was evaluated and compared. The regression models include Partial Least Squares Regression (PLSR) (Rosipal & Krämer, 2006), Back Propagation Neural Network (BPNN) (Khoshhal & Mokarram, 2012), Support Vector Regression (SVR) (Fan et al., 2018), K-Nearest Neighbors (KNN) (Cover & Hart, 1967) and Stacking (Wu et al., 2021). PLSR is capable of determining a linear relationship between the input variables and the response variables by simultaneously decomposing and screening data matrices with a few independent principal components (PCs). BPNN typically uses a forward neural network with input, hidden and output layers to build non-linear predictive models from a large dataset. The nonlinear mapping between inputs and outputs was approximated by layers of simple operation collections (neurons). SVR is the generalization of support vector machines for nonlinear regression problems, the algorithm inherits the advantages of SVM including the low complexity and non-linearity. KNN is an intuitive data classification method and also applicable to regression prediction problems. Stacking, with several weak learners based on different algorithms trained from the initial training dataset, the prediction results of these regressors are then used as a new training set to train a meta learner for the best combination of the weak learners into a strong learner. In this study, BPNN, SVR and KNN were used as weak learners of the Stacking algorithm, and a linear regressor was used as the meta-learner.

As the performance metrics, the coefficient of determination ( $R^2$ ), the root mean square error ( $RMSE$ ) were used for the performance assessment of different regression models, the equations are:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}, \quad (4)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}. \quad (5)$$

In which  $N$  is the number of spectrum samples,  $y_i$  is the measured chlorophyll content corresponding to the  $i$ th sample,  $\hat{y}_i$  is the model predicted value of the  $i$ th sample, and  $\bar{y}$  is the mean of the measured chlorophyll content of all samples.  $R_t^2$ ,  $RMSE_t$  and  $R_p^2$ ,  $RMSE_p$  were used as the matrices for performance evaluation, within which the subscript letters 't' and 'p' represent the prediction performance of using the training and testing datasets, respectively. Larger  $R^2$  and smaller  $RMSE$  indicate better the prediction performance of the model. The unit of  $RMSE$  is ( $\text{mg}/\text{cm}^2$ ).

The data samples was randomly divided into a training set (70%) and a testing set (30%). The training dataset was used for selecting the feature bands, evaluate the dimensionality reduction models, and train different regression models. The testing set was used for the assessment of the constructed regression models.

**Table 1** Several VIs reported effective in the estimation of leaf chlorophyll content

Vegetation index	Formula	References
GI	$R_{554}/R_{667}$	Zarco-Tejada et al. (2005)
Chl <sub>green</sub>	$R_{800}/R_{550} - 1.0$	Gitelson et al. (2006)
RGVI	$R_{550}/R_{670}$	Jordan (1969)
CARI	$CAR \times R_{700}/R_{670}$ $CARI_a = (R_{790} - R_{720})/150$ $CARI_b = R_{500} - CARI_a \times 500$ $CAR =  CARI_a \times R_{670} + R_{670} + CARI_b /\sqrt{CARI_a^2 + 1.0}$	Kim et al. (1994)
MTCI	$(R_{754} - R_{709})/(R_{709} + R_{681})$	Dash & Curran (2004)
GNDVI	$(R_{801} - R_{550})/(R_{801} + R_{550})$	Daughtry et al. (2000)
CI <sub>red edge</sub>	$(R_{750} - R_{700})/R_{700}$	Yu et al. (2014)
CI <sub>green</sub>	$(R_{750} - R_{550})/R_{550}$	Yu et al. (2014)

**Table 2** The empirical VIs and the corresponding formulas (Sun et al., 2019),  $R_i$  and  $R_j$  indicate the reflectance at bands  $i$  and  $j$  nm respectively

Vegetation index	Formula
DI	$R_i - R_j$
RI	$R_i/R_j$
NDVI	$(R_i - R_j)/(R_i + R_j)$
RDVI	$(R_i - R_j)/(R_i + R_j)^{0.5}$
VI <sub>opt</sub>	$(1 + 0.45)(R_i^2 + 1)/(R_j + 0.45)$
MSR	$((R_i/R_j) - 1)/(R_i/R_j + 1)^{0.5}$

Five-fold cross-validation technique was applied during training the regression models, in order to search for the optimal hyper-parameters of the regression models and avoid the impact of data bias during model assessment. The training set was further divided into five mutually exclusive subsets, in which four subsets were selected as the cross-training set and one subset as the cross-validation set. The validation was repeated until all subsets have been cross-validated, and the average of the five results were used as the final validation result. The optimal hyper-parameters were determined based on the validation results.

### Vegetation index-based models

A VI is a parameter constituted by a linear or non-linear combination of specific spectral bands, which effectively enhances certain details or features of vegetation. Many VIs have been reported to be effective in estimating biochemical parameters of plants including leaf chlorophyll content (Table 1). In this study, 8 reported effective VIs were selected, and a linear regression models was fitted for each VI using the training dataset. The prediction performance of these models were then evaluated using the testing dataset.

In addition, science the characteristic sensitive wavelengths of leaf chlorophyll content are different among different plant species, also varies upon the growing environment. Therefore, in this study, we reselected variable wavelengths from 6 empirical VIs (Table 2) in order to investigate a series of optimized VIs for chlorophyll content estimation in

sugarcane leaves. The optimized Vis were constructed by searching over all available band combinations in the VI formulas. The correlation coefficients  $R$  to the chlorophyll content were calculated for each formed VI. The combination with the maximum  $|R|$  was selected to form the optimal VI. A linear regression models was then fitted for each VI using training data, for estimation of chlorophyll content of sugarcane leaves.

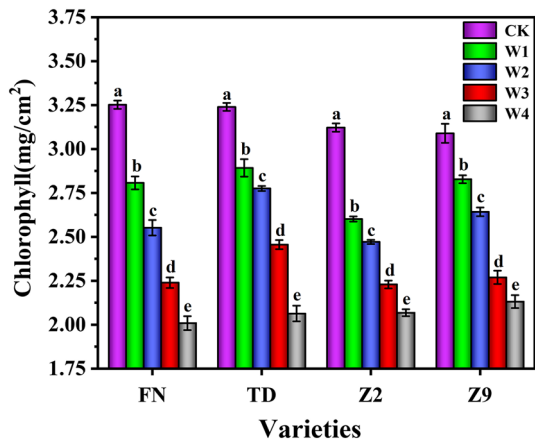
## Results

### Analysis of the effects of drought stress on sugarcane leaves

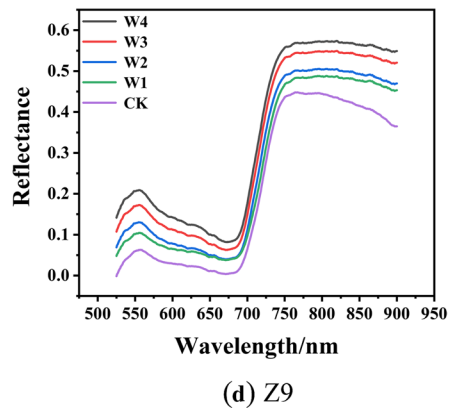
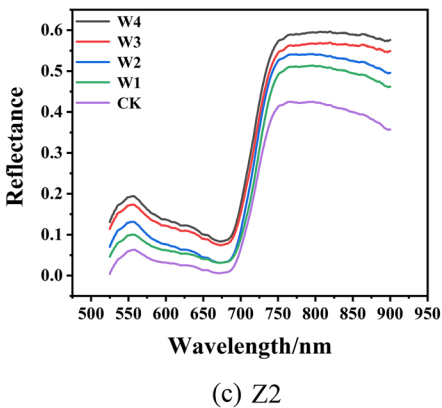
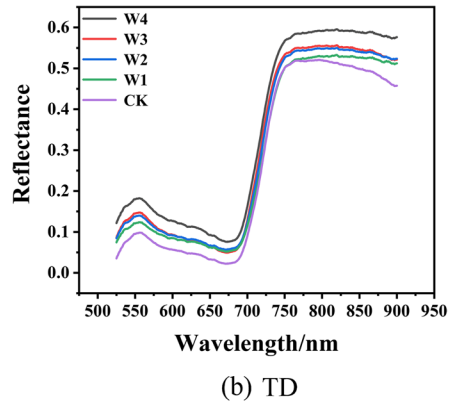
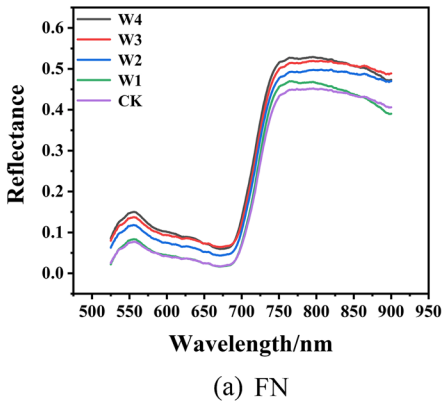
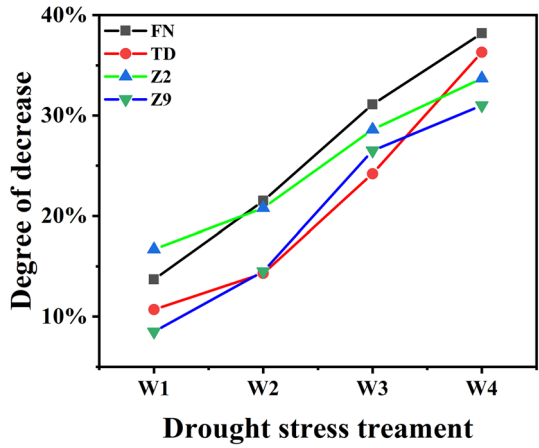
One-way ANOVA test and Turkey's HSD method was applied to the chlorophyll content data collected from plants with different treatments for each cultivar. The results of the analysis showed that the leaf chlorophyll content of different sugarcane varieties tended to decrease with the increase of drought stress, and the differences of chlorophyll content were test significant ( $p$ -value  $< 0.05$ ) between different treatments (Fig. 4). The average leaf chlorophyll contents under different drought stress treatments were compared with CK group. Under W1, W2, and W3 treatments, the decrease in chlorophyll content was greater in FN and Z2 than in TD and Z9 with increasing drought stress (Fig. 5). That is due to the fact that FN and Z2 are drought-sensitive cultivars, while TD and Z9 are more drought-tolerant.

In addition, for each sugarcane cultivar, sugarcane leaves under different stress treatments showed different spectral characteristics, but the overall trend of the spectral curves was the same (Fig. 6). There is a peak in reflectance at 560 nm in the visible green light region (525–600 nm), which is due to the photosynthetic effect absorb less light in this region. In the visible red region (630–690 nm), a reflectance valley is formed at 685 nm (known as the red valley), chlorophyll in plants has strong absorption and photosynthetic activity for light in this waveband. In the near infrared region (700–900 nm), the spectrum curve formed a reflectance peak at 755 nm, which was due to the reflection from the internal foliar cells inside the crop. The reflectance spectrum at 525–700 nm and 750–900 nm increases significant with increasing drought stress.

**Fig. 4** Statistical analysis and Tukey's HSD compact letters for chlorophyll content in leaves of different sugarcane cultivars under different drought stress treatments. Different letters indicate significant group-wise differences



**Fig. 5** The degrees of decreasing leaf chlorophyll content of different sugarcane cultivars compared with the control group (CK) as drought stress level increases



**Fig. 6** Mean spectral profiles (525–900 nm) of different sugarcane cultivars under different drought stress treatments

Given the aforementioned points, it is clear that drought stress causes changes in chlorophyll content of sugarcane leaves, which further leads to changes in the spectral profile. This result supports the feasibility of using spectral reflectance to predict the chlorophyll content of sugarcane leaves.

## Results of different spectral data dimensionality reduction methods

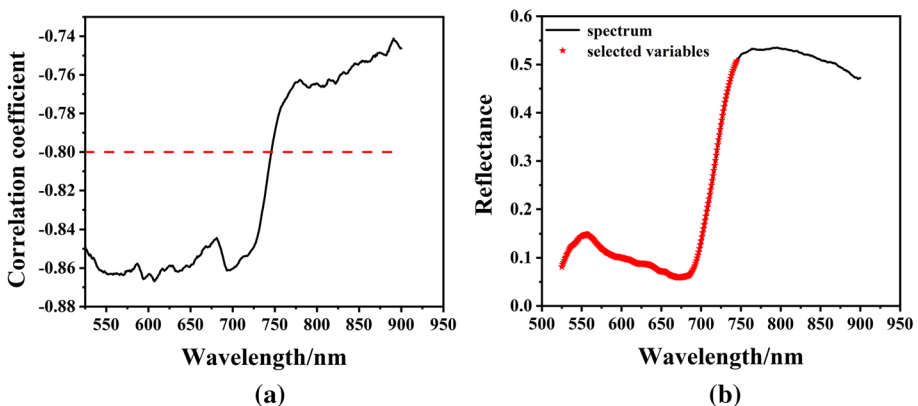
In order to reduce redundant spectral information and the computational cost of training machine learning models, dimensionality reduction techniques including the correlation coefficient method, SPA, RF and SR were applied to the spectra in the dataset. The results of these dimensionality reduction methods were analyzed, and their performance was evaluated and compared.

### Correlation coefficient method

By calculating the correlation coefficients between reflectance and chlorophyll content in each band, it was observed that the chlorophyll content in sugarcane leaves were negatively correlated with the spectral reflectance. The absolute values of the correlation coefficients between reflectance of most bands and the chlorophyll content were greater than 0.7, indicating strong correlations (Fig. 7a). The bands with absolute correlation coefficients greater than 0.8 were selected in this study, which were mainly concentrated in 525–745 nm (Fig. 7b). A total of 221 bands were selected, accounting for 58.78% of total bands (Table 3).

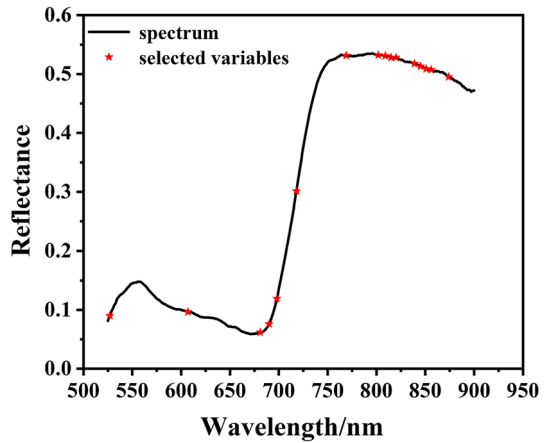
### Successive projections algorithm (SPA)

In this study, 16 characteristic bands were selected using SPA (Fig. 8), accounting for 4.3% of the total bands (Table 3). Of these, 7 filtered bands were in the visible region and 11



**Fig. 7** Results of the correlation coefficient method. **a** Correlation coefficients between each band reflectance and chlorophyll content; **b** bands selected by the correlation coefficient method

**Fig. 8** Bands screened by successive projections algorithm



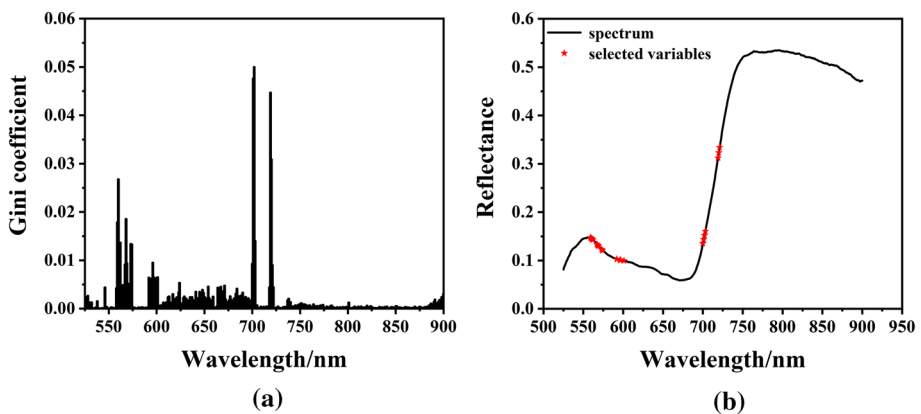
bands were in the NIR region. Figure 8 shows the bands selected by the successive projection algorithm.

### Random forest (RF) algorithm

The importance score of each band for chlorophyll prediction was obtained by RF (Fig. 9a). The bands with high importance were mainly concentrated in the 560–725 nm range, which implies the spectral reflectance within this range was obviously influenced by drought stress. The top 20 bands with the highest characteristic importance were selected (Fig. 9b), accounting for 5.32% of the total bands (Table 3).

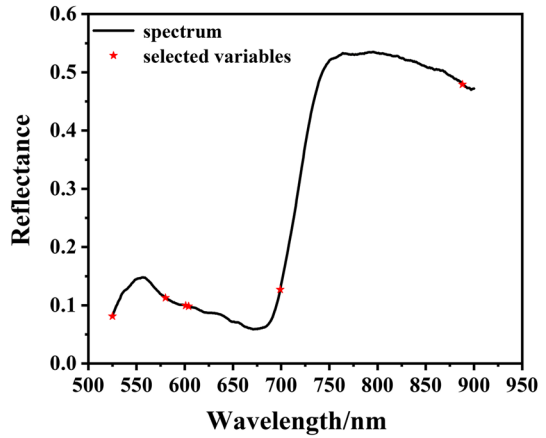
### Stepwise regression (SR) analysis

Six characteristic bands were selected using stepwise regression analysis, accounting for 2.5% of the total bands. Figure 10 shows the bands screened using the stepwise regression



**Fig. 9** Results of the RF method. **a** Importance of each band obtained by the RF algorithm; **b** selected bands with the RF algorithm

**Fig. 10** Bands screened by SR analysis



**Table 3** Comparative analysis of using input variable sets selected by different dimensionality reduction methods (M2–M6) with directly using the original spectrum (M1)

Variable set	Input variables (% of original)	$R^2$	$RMSE_{\cdot}$	$R_n^z$	$RMSE_{\cdot}$
M1	372 (100%)	0.8713	0.1371	<b>0.8836</b>	0.1371
M2	221 (59.4%)	0.8724	0.1366	0.8599	0.1579
M3	16 (4.3%)	0.8463	0.1499	<b>0.8662</b>	0.1543
M4	20 (5.4%)	0.7755	0.1811	0.8305	0.1737
M5	6 (1.6%)	0.7547	0.1894	0.7816	0.1972

The performance of using M1 and the best performance achieved after applying dimensionality reduction were labeled in bold

**Table 4** Grouping of chlorophyll content prediction model input variables

Variable set name	Dimensionality reduction method	Selected wavelengths/indices
M1	N/A (Full spectrum)	525–900 nm
M2	Correlation coefficient method	525–745 nm
M3	SPA	690 nm, 815 nm, 820 nm, 718 nm, 607 nm, 851 nm, 681 nm, 527 nm, 874 nm, 769 nm, 856 nm, 809 nm, 698 nm, 845 nm, 802 nm, 839 nm
M4	RF	702 nm, 701 nm, 719 nm, 720 nm, 560 nm, 568 nm, 559 nm, 703 nm, 562 nm, 561 nm, 573 nm, 574 nm, 596 nm, 569 nm, 700 nm, 567 nm, 721 nm, 597 nm, 601 nm, 592 nm
M5	SR	604 nm, 601 nm, 888 nm, 525 nm, 580 nm, 699 nm

analysis. The bands screened using the stepwise regression analysis were mainly between 400 and 750 nm, with fewer bands selected after 750 nm. This indicates the bands after 750 nm contained less important information.

Based on the previous results, six sets of variables were established as the input variables for the regression model (Table 4). Among them, M1 is the full spectrum, M2 is the set of the 221 characteristic bands selected by the correlation coefficient method, M3 is the set of the 16 characteristic bands selected by SPA, M4 is the set of the 20 characteristic bands selected by RF, M5 is the set of the 6 characteristic bands selected by SR.

Applying dimensionality reduction methods greatly reduced the complexity of the chlorophyll content prediction models comparing with the full spectrum (M1), at the cost of a slight decrease in prediction accuracy. Among the dimensionality reduction methods, SPA (M3) achieved the largest  $R_p^2$  and the smallest  $RMSE_p$  values when using a PLSR as a baseline regression model for chlorophyll content estimation, in which the  $R_p^2$  was only 0.0174 lower than using the full spectrum, but with 95.7% fewer inputs variables. In summary, dimensionality reduction is effective in eliminating redundant data in the spectra for sugarcane leaf chlorophyll content estimation, and SPA was the optimal method in both reducing computational effort and preserving the prediction accuracy.

### Comparative analysis between different regression models

For each input variable set, four regression methods, including BPNN, SVR, KNN and Stacking, were used to build the prediction models of sugarcane leaves chlorophyll content. The hyper-parameters of each model were optimized through grid searches with 5-fold

**Table 5** Performance of different combinations of input variable sets and regression models for predicting chlorophyll content in sugarcane leaves

Input variable set	Regression model	$R^2$	$R_M$	$R_{cv}^2$	$RMSE_{cv}$	$R_n^2$	$RMSE$
M1	BPNN	0.9871	0.0434	0.9655	0.0641	<b>0.9854</b>	<b>0.0509</b>
	SVR	0.9715	0.0644	0.9598	0.0734	0.9696	0.0736
	KNN	0.9952	0.0260	0.9734	0.0553	0.9742	0.0678
	Stacking	0.9950	0.0269	0.9782	0.0501	0.9832	0.0547
M2	BPNN	0.9840	0.0474	0.9697	0.0633	0.9813	0.0576
	SVR	0.9717	0.0641	0.9515	0.0775	0.9768	0.0643
	KNN	0.9961	0.0232	0.9449	0.0716	0.9736	0.0686
	Stacking	0.9937	0.0291	0.9670	0.0590	<b>0.9841</b>	<b>0.0533</b>
M3	BPNN	0.9786	0.0557	0.9616	0.0713	0.9778	0.0629
	SVR	0.9723	0.0635	0.9606	0.0722	0.9767	0.0644
	KNN	0.9948	0.0271	0.9690	0.0599	0.9755	0.0661
	Stacking	0.9935	0.0306	0.9750	0.0558	<b>0.9834</b>	<b>0.0544</b>
M4	BPNN	0.9601	0.0757	0.8952	0.1061	0.9521	0.0924
	SVR	0.968	0.0682	0.9270	0.0974	0.9647	0.0793
	KNN	0.9887	0.0404	0.9509	0.0740	0.9667	0.0770
	Stacking	0.9893	0.0393	0.9532	0.0726	<b>0.9725</b>	<b>0.0699</b>
M5	BPNN	0.9546	0.0807	0.9360	0.1781	0.9401	0.1033
	SVR	0.9714	0.0645	0.9516	0.0783	0.9567	0.0878
	KNN	0.9930	0.0317	0.9702	0.0620	0.9601	0.0843
	Stacking	0.9925	0.0329	0.9719	0.0603	<b>0.9661</b>	<b>0.0777</b>

The best performance achieved for each input variable set was label in bold, respectively



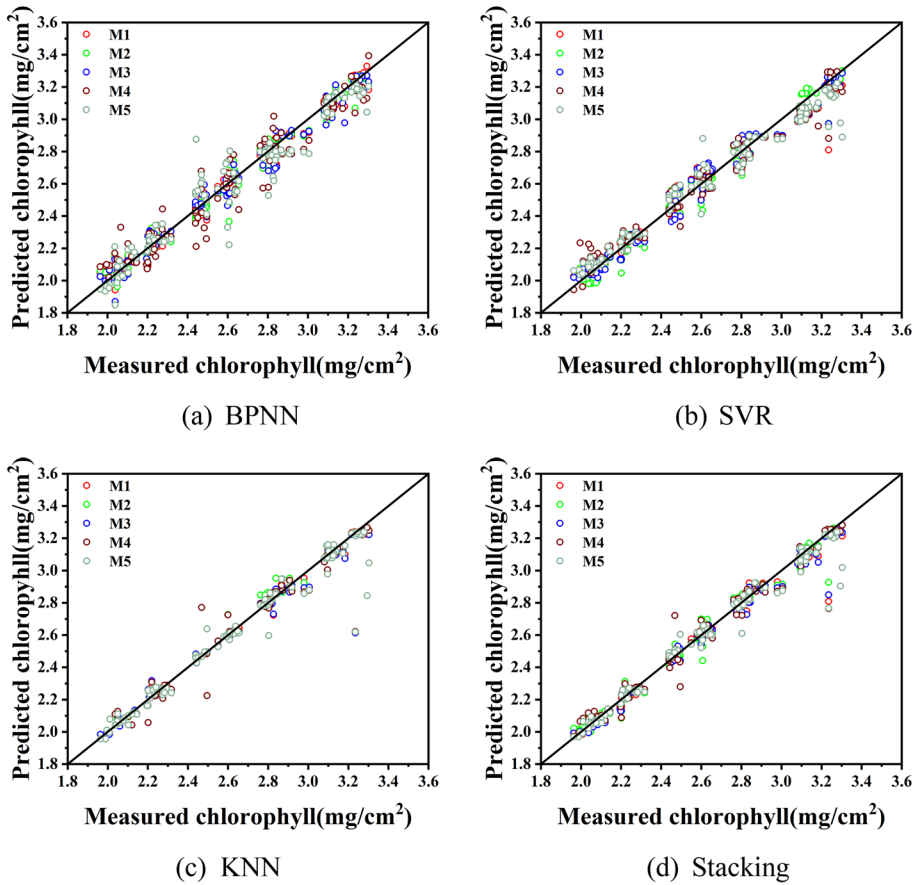
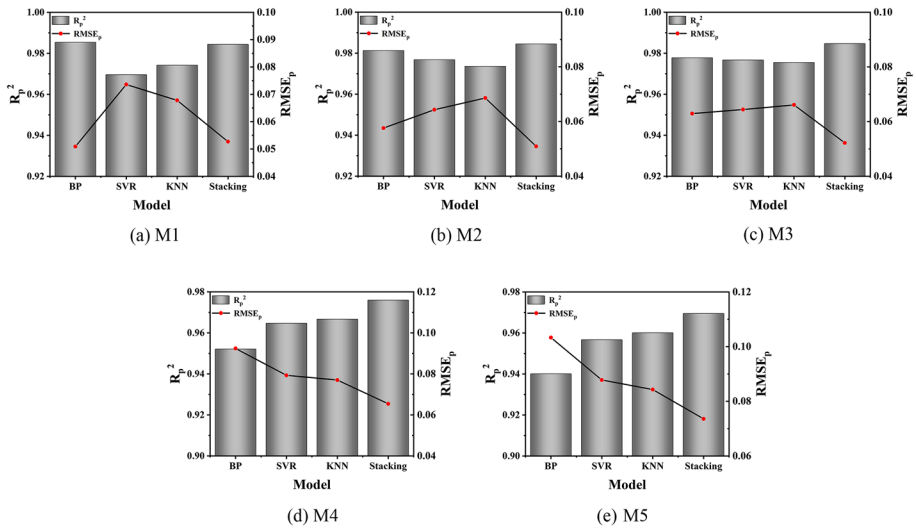


Fig. 11 Performance of different chlorophyll content prediction models using testing sets

cross-validation. The performance of these prediction models with different combinations of input variables are shown in Table 5. It can be observed that all fitted models had  $R_t^2$ ,  $R_{cv}^2$  and  $R_p^2$  values greater than 0.7, and  $RMSE_t$ ,  $RMSE_{cv}$  and  $RMSE_p$  values less than 0.2, indicating the models were trained properly, without overfitting issues.

The prediction performance of the four models with different input variable sets were evaluated and compared. The prediction results of different models were plotted in Fig. 11. the fitted lines were close to the reference line ( $y = x$ ), indicating the models were trained properly, without overfitting issues.

As observed in Table 5; Fig. 12, there is no method outperforms others. BPNN has the highest prediction accuracy when using input variable sets M1 (full spectrum), with  $R_p^2$  equals to 0.9854, and  $RMSE_p$  equals to 0.0641. For models using reduced input variable sets (M2–M5), Stacking-based models have the highest prediction accuracy while using each reduced input variable sets. And the prediction performance is comparable to using full spectrum and BPNN while using M2 (59.4% of original spectrum)



**Fig. 12** Performance comparison of different chlorophyll content prediction models composed of different input variable set and different regression methods

and M3 (4.3% of original spectrum) as input variable sets, with  $R_p^2$  equals to 0.9841 and 0.9834, respectively, and  $RMSE_p$  equals to 0.0533 and 0.0544, respectively.

Considering both prediction performance and model complexity, the model combining Stacking method with M3 (reduced by SPA) was selected as the optimal model for our future experiments. The optimal  $R_p^2$  is 0.9834 and the  $RMSE_p$  is 0.0544.

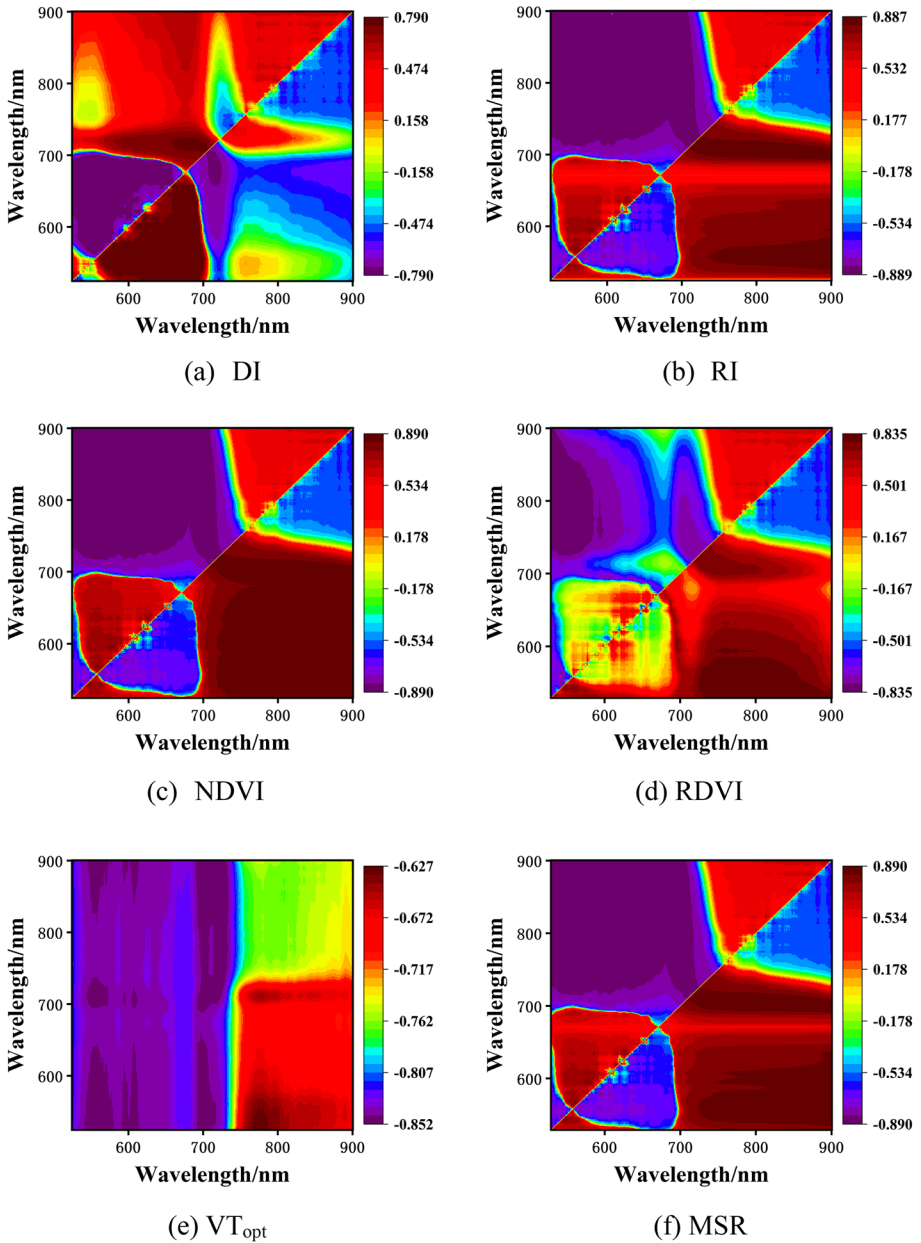
### Prediction models based on vegetation indices

A series of VIs were optimized by investigating the combination of bands variables in the formulas of each defined empirical VIs (Table 2). For each VI, the formed indices were evaluated by their absolute correlation coefficient  $|R|$  to the chlorophyll content (Fig. 13). The maximum  $|R|$  for each index definition with their corresponding variable wavelength combination are shown in Table 6.

It can be observed that the  $|R|_{max}$  of the eight optimized VIs were all higher than 0.75, indicating a good correlation with the leaf chlorophyll content. The highest  $|R|_{max} = 0.8888$  was obtained using the MSR index definition, with band combinations of 607 and 896 nm.

Figure 13. Plots of correlation coefficients between different VIs and the chlorophyll content. For each VI definition, the x-axis and y-axis are the variable wavelength combination used in the formula, range from 525 to 900 nm. The color corresponds to the correlation coefficient  $R$  between the VI formed by certain wavelength combination and the leaf chlorophyll content.

The training data was used to fit linear regression models describing how both existing and optimized VIs are related to the chlorophyll contents, respectively (Table 7). The results revealed that the linear models based on optimized VIs generally outperforms existing VIs in terms of chlorophyll content prediction performance. And model with optimized NDVI (570, 894) had the best performance with the highest  $R_p^2$  equals to 0.8208 and the lowest  $RMSE_p$  equal to 0.1786.



**Fig. 13** Plots of correlation coefficients between different VIs and the chlorophyll content. For each VI definition, the x-axis and y-axis are the variable wavelength combination used in the formula, range from 525 to 900nm. The color corresponds to the correlation coefficient R between the VI formed by certain wavelength combination and the leaf chlorophyll content

**Table 6** The Maximum correlation coefficient and the corresponding band position for different VIs.

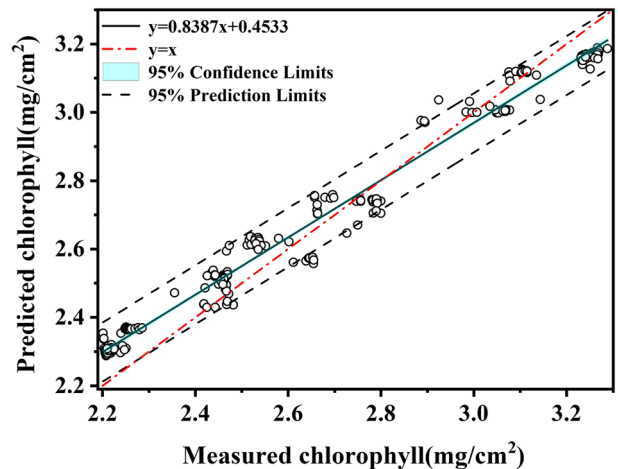
Vegetation index	Maximum correlation coefficient	
	$ R _{\max}$	Band position (i, j)/nm
DI	0.7893	(644, 689)
RI	0.8887	(607, 886)
NDVI	0.8886	(570, 894)
RDVI	0.8333	(529, 872)
VI <sub>opt</sub>	0.8511	(872, 727)
MSR	0.8888	(607, 896)

**Table 7** The regression models for different existing and optimized VIs.

VI	Regression equations	$R_t^2$	$RMSE_t$	$R_n^2$	$RMSE_p$
GI	$y = 0.0348x + 2.4348$	0.1563	0.3512	0.1045	0.3993
Chl <sub>green</sub>	$y = 0.2574x + 1.6924$	0.6828	0.2153	0.7328	0.2181
RGVI	$y = 0.0174x + 2.5043$	0.1154	0.3695	0.0705	0.3877
CARI	$y = -0.0554x + 2.6168$	0.0022	0.3924	0.0009	0.4020
MTCI	$y = 0.7175x + 0.9101$	0.4966	0.2712	0.5422	0.2855
GNDVI	$y = 4.2565x - 0.0404$	0.7408	0.1947	<b>0.7881</b>	0.1942
CI <sub>red edg</sub>	$y = 0.3036x + 1.4754$	0.6951	0.2111	0.7501	0.2109
CI <sub>green</sub>	$y = 0.2560x + 1.7229$	0.6672	0.2206	0.7160	0.2249
DI (644, 689)	$y = -118.2206x + 3.0982$	0.6025	0.2411	0.6545	0.2480
RI (607, 886)	$y = 7.1677x + 3.7258$	0.7736	0.1819	0.8176	0.1802
NDVI (570, 894)	$y = -4.7827x - 0.4539$	0.7715	0.1828	<b>0.8208</b>	0.1786
RDVI (529, 872)	$y = -9.9735x - 2.8317$	0.6972	0.2104	0.6820	0.2379
VI <sub>opt</sub> (872, 727)	$y = 5.1661x - 8.9117$	0.4993	0.2705	0.3594	0.2801
MSR (607, 896)	$y = -5.6620x + 1.8447$	0.7736	0.1819	0.8179	0.1801

The best performance achieved for using existing VIs and using optimized Vis were labeled in bold, respectively

**Fig. 14** Performance of the proposed model for chlorophyll content estimation on field samples



## Generalizability assessment of the chlorophyll content prediction model

To further validate the generalizability of the optimal chlorophyll content prediction model, the established optimal model proposed was applied to the 240 samples collected from sugarcane plants grown in the field. The predicted chlorophyll content values were found to be highly correlated ( $R^2=0.9577$ ) and accurate ( $RMSE=0.0775$ ) with respect to the ground truth measurements (Fig. 14). The fitted linear model between measured and predicted values has a slope of 0.8387, and an intercept of 0.4533 mg/cm<sup>2</sup>. The 95% confidence limit was about  $\pm 1$  mg/cm<sup>2</sup> for chlorophyll content prediction.

It was shown that the model established by SPA-based dimensionality reduction with Stacking achieved certain level of stability and generalizability in predicting chlorophyll content of sugarcane leaves.

## Discussions

### Effects of drought stress on chlorophyll content and spectrum of sugarcane leaves

The results showed that the chlorophyll content of sugarcane leaves decreases with increasing drought stress, and the reflectance spectrum of sugarcane leaves of all cultivars showed a significant trend with increasing drought stress. From the biological point of view, this is because drought stress accelerates the decomposition of chlorophyll in sugarcane leaves (Bin Zhang et al., 2020). As a consequence, the light absorption was hindered and therefore the reflectance of the spectrum at 525–700 nm, which is associated with respiratory pigments. Water stress also causes damage to the cellular structure of sugarcane leaves, which affects light absorption and increase reflectance to the spectrum of 750–950 nm (Zarco-Tejada et al., 2009). Therefore, the apparent changes in spectra caused by water stress made it possible to quantify the chlorophyll content using NIR/VIS reflectance spectroscopy, providing a foundation for predicting the degree of drought stress.

### Dimensionality reduction and regression methods

A significant reduction in the number of input variables was obtained through dimensionality reduction methods (correlation coefficient method, SPA, RF and SR) without a significant loss in prediction accuracy. This demonstrates the effectiveness of dimensionality reduction in reducing the redundancy of spectrum data for chlorophyll content estimation in sugarcane leaves.

In the aspect of selected bands, the characteristic bands selected by SR and SPA were mainly around 500 and 700 nm in the visible light region and 850–900 nm in the NIR region. This is consistent with previous studies on the important spectrum for chlorophyll content prediction (Carter & Knapp, 2001; Zhou et al., 2021). In contrast, the correlation coefficient method and the RF method select the band mainly in the visible light and fail to extract the important spectra in the near-infrared region.

In the aspect of performance in dimensionality reduction, as the experimental results, the overall performance of SR and SPA was better than the correlation coefficient and RF. That is because correlation coefficient method and RF only extracts variables that are highly correlated to the measured chlorophyll content without intercorrelation checked. While SPA and SR are algorithms capable of selecting features with less intercorrelation,

which further reduced the redundancy of the data. This also revealed the necessity of NIR spectrum in the predication of chlorophyll content prediction in sugarcane leaves.

With machine learning-based regression models, as shown in Fig. 13, the predicted chlorophyll contents were found highly correlated and accurate with respect to measured chlorophyll contents under most situations. However, in all cases, none of the methods outperforms others in terms of correlation ( $R^2$ ) and accuracy ( $RMSE$ ). The optimal regression model varied with the input variable set selected. Considering both prediction performance and model complexity, the optimal prediction model was constructed by combining Stacking regression algorithm with 16 characteristic bands selected by SPA. The model could yield a prediction accuracy of  $R_p^2$  equals to 0.9834 and  $RMSE_p$  equals to 0.0544.

### Comparison of our proposed model with VI-based models

A series of linear regression models based on existing VIs and optimized VIs were constructed in this study, as these were the common approaches for the estimation of biochemical parameters of sugarcane plants (Amarasingam et al., 2022; Li et al., 2015). Performance of linear models based on optimized VIs is generally higher than those based on existing VIs reported effective in Chlorophyll content estimation, which is consistent with the findings in Yang et al. (2021). The models with optimized NDVI (570, 894), optimized RI (607, 886) and optimized MSR (607, 896) yield a  $R_p^2$  greater than 0.8. The models with optimized NDVI (570, 894) has the best performance ( $R_p^2 = 0.8208$  and  $RMSE_p = 0.1786$ ) in VI-based models. The characteristic bands selected were around 600 and 890 nm, which were in red zone related to the absorbance band of chlorophyll, and in NIR zone with high reflectance related to the leaf cellular structure. Therefore, optimized VI-based models can achieved a decent performance in leaf chlorophyll content prediction with very few variable bands, and results of selected bands could be useful in designing optical sensors for leaf chlorophyll content prediction.

The prediction performance of the proposed optimal model (i.e. SPA + Stacking) were compared with VI-based models. The performance of our best optimized NDVI-based models ( $R_p^2 = 0.8208$  and  $RMSE_p = 0.1786$ ) is still lower than the proposed model based on feature band extraction and machine learning ( $R_p^2 = 0.9834$  and  $RMSE_p = 0.0544$ ). Our results demonstrated that our proposed model outperforms of commonly used VI-based methods in estimation of chlorophyll content of sugarcane leaves. That is because the VIs' nature of dimensionality reduction may lead to the loss of significant information reflecting the physiological effects corresponding to chlorophyll content changes, thus decrease the estimation accuracy, thus not suitable for an accurate estimation of leaf chlorophyll content.

### Generalizability of the chlorophyll content prediction model

When applying the trained optimal regression model onto field-collected data for generalizability evaluation, the predicted chlorophyll content values were found to be highly correlated ( $R^2 = 0.9577$ ) and accurate ( $RMSE = 0.0775$ ) with respect to the ground truth values. A certain level of generalizability was obtained in monitoring sugarcane drought stress. However, the prediction accuracy was lower compared to data collected in the greenhouse. A slope below one was observed, indicating the relationship between the spectrum and chlorophyll content varies for sugarcane plants cultivated in the greenhouse and in the

field. The main reason was that the field environment was less controllable and different from that of the greenhouse. This affected the growth of the sugarcane plants, including the cellular structures.

Therefore, the question regarding the optimal dimensionality reduction method and regression model has been answered in this paper. A series of spectral data and models for chlorophyll content estimation of sugarcane leaves at elongation stage were proposed and could be treated as priority indices to be tested in any follow-up work. Further research will be conducted regarding the optimal strategy for drought stress assessment of the sugarcane plants using the chlorophyll content as an intermedium variable. At the same time, environmental variables affecting plant growth should be investigated and incorporated into the estimation model to improve the generalizability of the predication models.

## Conclusions

In this study, we proposed models for the estimation of leaf chlorophyll content of drought-affected sugarcane using VIS/NIR reflectance spectroscopy. By comparing the combinations of different dimensionality reduction techniques and different regression models, a chlorophyll content estimation model with optimal performance was constructed. The results show that the selection of characteristic bands is effective in reducing the redundancy in the spectrum, while improving the prediction accuracy of certain models. Considering both prediction performance and model complexity, the optimal prediction model was constructed by combining successive projection algorithm (SPA) and Stacking regression algorithm, with  $R^2=0.9834$ ,  $RMSE = 0.0544$ . The proposed model outperforms the commonly used VI-based approaches. By applying the trained optimal model onto field-collected data, the predicted chlorophyll content values were found correlated ( $R^2=0.9577$ ) and accurate ( $RMSE = 0.0775$ ) with respect to the measured ground truth, which demonstrated a certain level of generalizability of the model.

The overall results manifests that the proposed SPA + Stacking model could provide an effective, accurate and non-destructive way for monitoring the leaf chlorophyll content of sugarcane plants. As chlorophyll content is closely related to the degree of drought, applying the proposed chlorophyll content prediction method to sugarcane cultivation can provide general guidance for irrigation management and a scientific basis for water resources allocation and drought preparation, thus improving sugarcane yield and reduce drought risk.

However, the developed prediction model developed has the major limitation that only a limited range of environmental factors affecting the sugarcane growth were considered in this study. Factors including meteorological conditions, water diversion and other uncontrollable factors were not considered, which may reduce the performance of the prediction model. Therefore, our future research plan is to further investigate and incorporate environmental factors into the model proposed in this study.

**Acknowledgements** This work was supported by National Natural Science Foundation of China (31901466).

## References

Abdel-Rahman, E. M., Ahmed, F. B., & Ismail, R. (2013). Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 hyperion hyperspectral data.

- International Journal of Remote Sensing*, 34(2), 712–728. <https://doi.org/10.1080/01431161.2012.713142>
- Amarasingam, N., Salgadoe, A. S. A., Powell, K., Gonzalez, L. F., & Natarajan, S. (2022). A review of UAV platforms, sensors, and applications for monitoring of sugarcane crops. *Remote Sensing Applications: Society and Environment*, 26, 100712. <https://doi.org/10.1016/j.rsase.2022.100712>
- Araújo, M. C. U., Saldanha, T. C. B., Galvão, R. K. H., Yoneyama, T., Chame, H. C., & Visani, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*. [https://doi.org/10.1016/S0169-7439\(01\)00119-8](https://doi.org/10.1016/S0169-7439(01)00119-8)
- Bhadra, S., Sagan, V., Maimaitijiang, M., Maimaitiyiming, M., Newcomb, M., Shakoor, N., & Mockler, T. C. (2020). Quantifying leaf chlorophyll concentration of sorghum from hyperspectral data using derivative calculus and machine learning. *Remote Sensing*. <https://doi.org/10.3390/rs12132082>
- Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates. *Machine Learning*, 48, 1–3. <https://doi.org/10.1023/A:1013964023376>.
- Carter, G. A., & Knapp, A. K. (2001). Leaf optical properties in higher plants: Linking spectral characteristics to stress and chlorophyll concentration. *American Journal of Botany*, 88(4), 677–684. <https://doi.org/10.2307/2657068>
- Chaoyang, W. U., Wang, L. I., Niu, Z., Gao, S., & Mingquan, W. U. (2010). Nondestructive estimation of canopy chlorophyll content using hyperion and landsat/TM images. *International Journal of Remote Sensing*, 31(7–8), 2159–2167.
- Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. <https://doi.org/10.1109/TIT.1967.1053964>
- DAIE, J. (1992). Annual review of plant physiology and plant molecular biology. *Soil Science*, 154(6), 508. <https://doi.org/10.1097/00010694-199212000-00011>
- Dash, J., & Curran, P. J. (2004). MTCI: The meris terrestrial chlorophyll index. *International Journal of Remote Sensing*, 25(549), 151–161.
- Daughtry, C. S. T., Walthall, C. L., Kim, M. S., de Colstoun, E. B., & McMurtrey, J. E. (2000). Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. *Remote Sensing of Environment*, 74(2), 229–239. [https://doi.org/10.1016/S0034-4257\(00\)00113-9](https://doi.org/10.1016/S0034-4257(00)00113-9)
- El-Hendawy, S., Elsayed, S., Al-Suhaibani, N., Alotaibi, M., Tahir, M. U., Mubushar, M., Attia, A., & Hassan, W. M. (2021). Use of hyperspectral reflectance sensing for assessing growth and chlorophyll content of spring wheat grown under simulated saline field conditions. *Plants*. <https://doi.org/10.3390/plants10010101>
- Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H., Wang, X., Lu, X., & Xiang, Y. (2018). Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. *Agricultural and Forest Meteorology*. <https://doi.org/10.1016/j.agrformet.2018.08.019>
- Gitelson, A. A., & Merzlyak, M. N. (1996). Signature analysis of leaf reflectance spectra: Algorithm development for remote sensing of chlorophyll. *Journal of Plant Physiology*, 148, 3–4. [https://doi.org/10.1016/S0176-1617\(96\)80284-7](https://doi.org/10.1016/S0176-1617(96)80284-7)
- Gitelson, A. A., Keydan, G. P., & Merzlyak, M. N. (2006). Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves. *Geophysical Research Letters*. <https://doi.org/10.1029/2006GL026457>
- Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health*. <https://doi.org/10.2105/AJPH.79.3.340>
- Hoepfner, J. M., Skidmore, A. K., Darvishzadeh, R., Heurich, M., Chang, H. C., & Gara, T. W. (2020). Mapping canopy chlorophyll content in a temperate forest using airborne hyperspectral data. *Remote Sensing*. <https://doi.org/10.3390/rs12213573>
- Inman-Bamber, N. G., & Smith, D. M. (2005). Water relations in sugarcane and response to water deficits. *Field Crops Research*, 92(2–3), 185–202. <https://doi.org/10.1016/j.fcr.2005.01.023>
- Inoue, Y., Guérif, M., Baret, F., Skidmore, A., Gitelson, A., Schlerf, M., Darvishzadeh, R., & Olioso, A. (2016). Simple and robust methods for remote sensing of canopy chlorophyll content: A comparative analysis of hyperspectral data for different types of vegetation. *Plant Cell and Environment*. <https://doi.org/10.1111/pce.12815>
- Jacquemoud, S., & Baret, F. (1990). PROSPECT: A model of leaf optical properties spectra. *Remote Sensing of Environment*, 34, 75–91. [https://doi.org/10.1016/0034-4257\(90\)90100-Z](https://doi.org/10.1016/0034-4257(90)90100-Z)
- Jordan, C. F. (1969). Derivation of leaf-area index from quality of light on the forest floor. *Ecology*, 50(4), 663–666. <https://doi.org/10.2307/1936256>
- Källén, B. (2016). Statistics for dummies. In C. Schaefer, P. Peters, & R. R. Miller (Eds.), *Drugs during pregnancy*. Elsevier.



- Khoshhal, J., & Mokarram, M. (2012). Model for prediction of evapotranspiration using MLP neural network. *International Journal of Environmental Sciences*, 3, 1000–1009.
- Kim, M. S., Daughtry, C., Chappelle, E. W., McMurtrey, J. E., & Walthall, C. L. (1994). The use of high spectral resolution bands for estimating absorbed photosynthetically active radiation (A par). *CNES, Proceedings of 6th International Symposium on Physical Measurements and Signatures in Remote Sensing*.
- Krishna, G., Sahoo, R. N., Singh, P., Bajpai, V., Patra, H., Kumar, S., Dandapani, R., Gupta, V. K., Viswanathan, C., Ahmad, T., & Sahoo, P. M. (2019). Comparison of various modelling approaches for water deficit stress monitoring in rice crop through hyperspectral remote sensing. *Agricultural Water Management*, 213, 231–244. <https://doi.org/10.1016/j.agwat.2018.08.029>
- Li, X., Chen, X., Zhou, Y., Nong, M., Liu, X., & Ai, J. (2015). Spectral characteristics analysis and chlorophyll content detection of sugarcane leaves under different fertilizer treatments. *Transactions of the Chinese Society of Agricultural Engineering*, 31, 118–123. <https://doi.org/10.11975/j.issn.1002-6819.2015.z2.018>
- Liang, L., Di, L., Huang, T., Wang, J., Lin, L., Wang, L., & Yang, M. (2018). Estimation of leaf nitrogen content in wheat using new hyperspectral indices and a random forest regression algorithm. *Remote Sensing*. <https://doi.org/10.3390/rs10121940>
- Lichtenthaler, H. K., & Wellburn, A. R. (1983). Determinations of total carotenoids and chlorophylls a and b of leaf extracts in different solvents. *Biochemical Society Transactions*, 11(5), 591–592. <https://doi.org/10.1042/bst0110591>
- Liu, N., Qiao, L., Xing, Z., Li, M., Sun, H., Zhang, J., & Zhang, Y. (2020). Detection of chlorophyll content in growth potato based on spectral variable analysis. *Spectroscopy Letters*. <https://doi.org/10.1080/00387010.2020.1772827>
- Osco, L. P., Junior, J. M., Ramos, A. P. M., Furuya, D. E. G., Santana, D. C., Teodoro, L. P. R., Gonçalves, W. N., Baio, F. H. R., Pistori, H., da Junior, C. A. S., & Teodoro, P. E. (2020). Leaf nitrogen concentration and plant height prediction for maize using UAV-based multispectral imagery and machine learning techniques. *Remote Sensing*, 12(19), 1–17. <https://doi.org/10.3390/rs12193237>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Peng, Z., Guan, L., Liao, Y., & Lian, S. (2019). Estimating total leaf chlorophyll content of gannan navel orange leaves using hyperspectral data based on partial least squares regression. *IEEE Access : Practical Innovations, Open Solutions*, 7, 155540–155551. <https://doi.org/10.1109/access.2019.2949866>
- Qiao, L., Tang, W., Gao, D., Zhao, R., An, L., Li, M., Sun, H., & Song, D. (2022). UAV-based chlorophyll content estimation by evaluating vegetation index responses under different crop coverages. *Computers and Electronics in Agriculture*, 196, 106775. <https://doi.org/10.1016/j.compag.2022.106775>
- Rospal, R., & Krämer, N. (2006). Overview and recent advances in partial least squares. *Lecture Notes in Computer Science(Including Subseries Lecture Notes in Artificial Intelligence And Lecture Notes in Bioinformatics)*, 3940 LNCS, (pp. 34–51). [https://doi.org/10.1007/11752790\\_2](https://doi.org/10.1007/11752790_2)
- Sims, D. A., & Gamon, J. A. (2002). Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sensing of Environment*, 81(2), 337–354. [https://doi.org/10.1016/S0034-4257\(02\)00010-X](https://doi.org/10.1016/S0034-4257(02)00010-X)
- Sonobe, R., Yamashita, H., Mihara, H., Morita, A., & Ikka, T. (2021). Hyperspectral reflectance sensing for quantifying leaf chlorophyll content in wasabi leaves using spectral pre-processing techniques and machine learning algorithms. *International Journal of Remote Sensing*, 42(4), 1311–1329. <https://doi.org/10.1080/01431161.2020.1826065>
- Sun, H., Feng, M., Xiao, L., Yang, W., Wang, C., Jia, X., Zhao, Y., Zhao, C., Muhammad, S. K., & Li, D. (2019). Assessment of plant water status in winter wheat (*Triticum aestivum* L.) based on canopy spectral indices. *PLoS One*. <https://doi.org/10.1371/journal.pone.0216890>
- Verhoef, W. (1984). Light scattering by leaf layers with application to canopy reflectance modeling: The SAIL model. *Remote Sensing of Environment*. [https://doi.org/10.1016/0034-4257\(84\)90057-9](https://doi.org/10.1016/0034-4257(84)90057-9)
- Wang, H. L., Yang, G. G., Zhang, Y., Bao, Y. D., & He, Y. (2017). Detection of fungal disease on tomato leaves with competitive adaptive reweighted sampling and correlation analysis methods. *Guang Pu Xue Yu Guang Pu Fen Xi/Spectroscopy and Spectral Analysis*. [https://doi.org/10.3964/j.issn.1000-0593\(2017\)07-2115-05](https://doi.org/10.3964/j.issn.1000-0593(2017)07-2115-05)
- Wang, K., Li, W., Deng, L., Lyu, Q., Zheng, Y., Yi, S., Xie, R., Ma, Y., & He, S. (2018). Rapid detection of chlorophyll content and distribution in citrus orchards based on low-altitude remote sensing and bio-sensors. *International Journal of Agricultural and Biological Engineering*. <https://doi.org/10.25165/j.ijabe.20181102.3189>

- Wang, J. J., Li, Z., Jin, X., Liang, G., Struik, P. C., Gu, J., & Zhou, Y. (2019). Phenotyping flag leaf nitrogen content in rice using a three-band spectral index. *Computers and Electronics in Agriculture*. <https://doi.org/10.1016/j.compag.2019.04.042>
- Wang, L., Liao, Q., Xu, X., Li, Z., & Zhu, H. (2020). Estimating the vertical distribution of chlorophyll in winter wheat based on multi-angle hyperspectral data. *Remote Sensing Letters*. <https://doi.org/10.1080/2150704X.2020.1817605>
- Wu, T., Zhang, W., Jiao, X., Guo, W., & Alhaj Hamoud, Y. (2021). Evaluation of stacking and blending ensemble learning methods for estimating daily reference evapotranspiration. *Computers and Electronics in Agriculture*. <https://doi.org/10.1016/j.compag.2021.106039>
- Xuan, G., Gao, C., Shao, Y., Wang, X., Wang, Y., & Wang, K. (2021). Maturity determination at harvest and spatial assessment of moisture content in okra using Vis–NIR hyperspectral imaging. *Postharvest Biology and Technology*. <https://doi.org/10.1016/j.postharvbio.2021.111597>
- Yadav, S., Jackson, P., Wei, X., Ross, E. M., Aitken, K., Deomano, E., Atkin, F., Hayes, B. J., & Voss-Fels, K. P. (2020). Accelerating genetic gain in sugarcane breeding using genomic selection. *Agronomy*. <https://doi.org/10.3390/agronomy10040585>
- Yamashita, H., Sonobe, R., Hirono, Y., Morita, A., & Ikka, T. (2020). Dissection of hyperspectral reflectance to estimate nitrogen and chlorophyll contents in tea leaves based on machine learning algorithms. *Scientific Reports*, 10(1), 1–11. <https://doi.org/10.1038/s41598-020-73745-2>
- Yang, J., Zhang, Y., Du, L., Liu, X., Shi, S., & Chen, B. (2021). Improving the selection of vegetation index characteristic wavelengths by using the prospect model for leaf water content estimation. *Remote Sensing*, 13(4), 1–21. <https://doi.org/10.3390/RS13040821>
- Yoosofzadeh-Najafabadi, M., Earl, H. J., Tulpan, D., Sulik, J., & Eskandari, M. (2021). Application of machine learning algorithms in plant breeding: Predicting yield from hyperspectral reflectance in soybean. *Frontiers in Plant Science*, 11(January), 1–14. <https://doi.org/10.3389/fpls.2020.624273>
- Yu, K., Lenz-Wiedemann, V., Chen, X., & Bareth, G. (2014). Estimating leaf chlorophyll of barley at different growth stages using spectral indices to reduce soil background and canopy structure effects. *ISPRS Journal of Photogrammetry and Remote Sensing*, 97, 58–77. <https://doi.org/10.1016/j.isprsjprs.2014.08.005>
- Zarco-Tejada, P. J., Miller, J. R., Mohammed, G. H., & Noland, T. L. (2000). Chlorophyll fluorescence effects on vegetation apparent reflectance: I. Leaf-level measurements and model simulation. *Remote Sensing of Environment*. [https://doi.org/10.1016/S0034-4257\(00\)00148-6](https://doi.org/10.1016/S0034-4257(00)00148-6)
- Zarco-Tejada, P. J., Berjón, A., López-Lozano, R., Miller, J. R., Martín, P., Cachorro, V., González, M. R., & de Frutos, A. (2005). Assessing vineyard condition with hyperspectral indices: Leaf and canopy reflectance simulation in a row-structured discontinuous canopy. *Remote Sensing of Environment*, 99(3), 271–287. <https://doi.org/10.1016/j.rse.2005.09.002>
- Zarco-Tejada, J., Asner, G. P., François, C., Ustin, S. L., Jacquemoud, S., Verhoef, W., Baret, F., & Bacour, C. (2009). PROSPECT + SAIL models: A review of use for vegetation characterization. *Remote Sensing of Environment*. <https://doi.org/10.1016/j.rse.2008.01.026>
- Zhang, Y., Bin, Yang, S. L., Dao, J. M., Deng, J., Shahzad, A. N., Fan, X., Li, R. D., Quan, Y. J., Bukhari, S. A. H., & Zeng, Z. H. (2020). Drought-induced alterations in photosynthetic, ultrastructural and biochemical traits of contrasting sugarcane genotypes. *PLoS One*. <https://doi.org/10.1371/journal.pone.0235845>
- Zhang, H., Ge, Y., Xie, X., Atefi, A., Wijewardane, N. K., & Thapa, S. (2022). High throughput analysis of leaf chlorophyll content in sorghum using RGB, hyperspectral, and fluorescence imaging and sensor fusion. *Plant Methods*, 18(1), 1–17. <https://doi.org/10.1186/s13007-022-00892-0>
- Zhao, D., & Li, Y. R. (2015). Climate change and sugarcane production: potential impact and mitigation strategies. *International Journal of Agronomy*. <https://doi.org/10.1155/2015/547386>
- Zhou, J. J., Zhang, Y. H., Han, Z. M., Liu, X. Y., Jian, Y. F., Hu, C. G., & Dian, Y. Y. (2021). Evaluating the performance of hyperspectral leaf reflectance to detect water stress and estimation of photosynthetic capacities. *Remote Sensing*, 13(11), 1–21. <https://doi.org/10.3390/rs13112160>
- Zhu, W., Sun, Z., Yang, T., Li, J., Peng, J., Zhu, K., Li, S., Gong, H., Lyu, Y., Li, B., & Liao, X. (2020). Estimating leaf chlorophyll content of crops via optimal unmanned aerial vehicle hyperspectral data at multi-scales. *Computers and Electronics in Agriculture*. <https://doi.org/10.1016/j.compag.2020.105786>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

Jingyao Gai<sup>1</sup>  · Jingyong Wang<sup>1</sup> · Sasa Xie<sup>2</sup> · Lirong Xiang<sup>3</sup> · Ziting Wang<sup>1,4</sup>

✉ Jingyao Gai  
jygai@gxu.edu.cn

✉ Ziting Wang  
zitingwang@gxu.edu.cn

<sup>1</sup> College of Mechanical Engineering, Guangxi University, Nanning 530004, Guangxi, People's Republic of China

<sup>2</sup> College of Agronomy, Guangxi University, Nanning 530004, Guangxi, People's Republic of China

<sup>3</sup> Department of Biological & Agricultural Engineering, North Carolina State University, Raleigh, NC 27695, USA

<sup>4</sup> State Key Laboratory for Conservation & Utilization of Subtropical Agro-bioresources, Guangxi University, Nanning 530004, Guangxi, People's Republic of China