SPAD PREDICTION MODEL FOR TEA LEAVES BASED ON THE IRIV ALGORITHM
/
基于 IRIV 算法的茶叶叶片 SPAD 预测模型

Gong CHENG [1], Tengxiang YANG[1], Chengqian JIN [1*], Zeyu CAI[1], Man CHEN [1,2*], Xiaoqiang SUN [3]
[1] Nanjing Institute of Agricultural Mechanization, Ministry of Agriculture and Rural Affairs, Nanjing, Jiangsu / China;
[2] National Digital Agriculture Equipment (South China Intelligence Agricultural Machine) Innovation Sub-center, Nanjing, Jiangsu / China;
[3] Wuxi Xintiandi Agricultural Development Co., Ltd, Wuxi, Jiangsu / China;
Tel: +86 025-84346113; E-mail: chenman@caas.cn; jinchengqian@caas.cn
DOI: https://doi.org/10.35633/inmateh-78-30

Keywords: *Spectral Analysis; Tea Leaves; Chlorophyll Content; Iteratively Retained Informative Variables Algorithm; SPAD*

ABSTRACT
This study focused on three tea cultivars from the Jianghan Plain to construct an inversion model between multispectral features and chlorophyll content in tea leaves. Based on 120 samples across two growth stages, indoor multispectral imaging technology was used to simultaneously acquire leaf multispectral data and SPAD values. Through the analysis of the spectral-chlorophyll response mechanism and the evaluation of feature wavelength autocorrelation, the Iteratively Retained Informative Variables (IRIV) algorithm was integrated for feature selection. An evaluation system consisting of seven machine learning models, including Partial Least Squares Regression (PLSR) and Support Vector Regression (SVR), was established. The results showed that the model combining the adjacent band change rate features selected by IRIV with Multiple Linear Regression (MLR) achieved the optimal inversion accuracy (R^2=0.785, RMSE=4.241). Additionally, the vegetation index-MLR combination (R^2=0.791, RMSE=4.222) and the mixed feature-LASSO combination (R^2=0.773, RMSE=4.403) performed prominently under different feature dimensions. This study provides a feature engineering scheme with strong interpretability and a model optimization path for hyperspectral non-destructive detection of tea physiological parameters.

摘要
本研究以江汉平原的三个茶树品种作为研究对象，构建多光谱特征与茶叶叶片叶绿素含量的反演模型。基于两个生长阶段的 120 组样本，采用室内多光谱成像技术同步获取叶片多光谱数据及 SPAD 值，通过光谱-叶绿素响应机制解析与特征波长自相关性评估，融合迭代保留信息变量（IRIV）进行特征筛选，构建包含偏最小二乘回归（PLSR）、支持向量回归（SVR）等七种机器学习模型的评估体系。研究表明：基于 IRIV 筛选的相邻波段变化率特征结合多元线性回归（MLR）模型反演精度最优（R^2=0.785，RMSE=4.241），而植被指数-MLR 组合（R^2=0.791，RMSE=4.222）及混合特征-LASSO 组合（R^2=0.773，RMSE=4.403）在不同特征维度下表现突出。本研究为高光谱无损检测茶叶生理参数提供了可解释性强的特征工程方案与模型优化路径。

INTRODUCTION
Chlorophyll in tea leaves is a crucial pigment for photosynthesis *(Duan Dan-dan et al., 2024),* and the content and distribution of chlorophyll in leaves are closely related to tea quality *(Xu et al., 2024)*. Traditional methods for estimating chlorophyll content mainly rely on physical and chemical approaches. Although these methods can yield relatively accurate results, they are time-consuming and labor-intensive during sample collection, damage the tissue structure *(Zhang and Kovacs, 2012)*, and make it difficult to monitor the entire life cycle of the same sample *(Yang et al., 2021)*. In contrast, spectral technology has significant potential and advantages in non-destructive and efficient crop monitoring.

In recent years, studies on the relationship between spectral data and chlorophyll content using spectral technology have emerged continuously. Under normal outdoor light conditions, researchers have conducted extensive studies on chlorophyll content in crops such as rice *(Chen et al., 2024)*, wheat *(X. Chen et al., 2025; Li et al., 2025; Yin et al., 2023)*, cotton *(Wang et al., 2024)* and tea *(Qi et al., 2025)*, confirming the feasibility of spectral inversion models. Under indoor artificial light sources, researchers have further improved the accuracy of chlorophyll prediction for crops like rice *(Chen Y. et al., 2025*), soybeans *(Mao et al., 2020)*, Chinese cabbage *(Zhang et al., 2023)*, and lemons *(Li et al., 2022)* by standardizing lighting conditions.

Notably, although some scholars have initially explored the spectral response law of tea leaves *(Qi et al., 2026; Tsuchiya et al., 2025; Wang et al., 2026; Wu et al., 2025; Xie et al., 2025)*, the complex influence of surface characteristics of different cultivars on spectral scattering remains unclear, which limits the universality of the models.

This study focused on multiple tea leaf cultivars. A high-precision multispectral imaging system was used to simultaneously collect spectral information of tea leaves under indoor controlled light sources, and a dataset was constructed by combining the measured SPAD values. By comparing the spectral feature selection capabilities and non-linear fitting performance of different modeling methods, the optimization path of each model in non-destructive chlorophyll detection of tea leaves was systematically evaluated, and the model performance was compared and analyzed. The results of this study establish a technical paradigm for rapid diagnosis and dynamic monitoring of tea chlorophyll, and the revealed spectral response mechanism lays a theoretical and methodological foundation for the design of multispectral sensing equipment, promoting the engineering transformation of tea chlorophyll detection technology from laboratory analysis to real-time field monitoring.

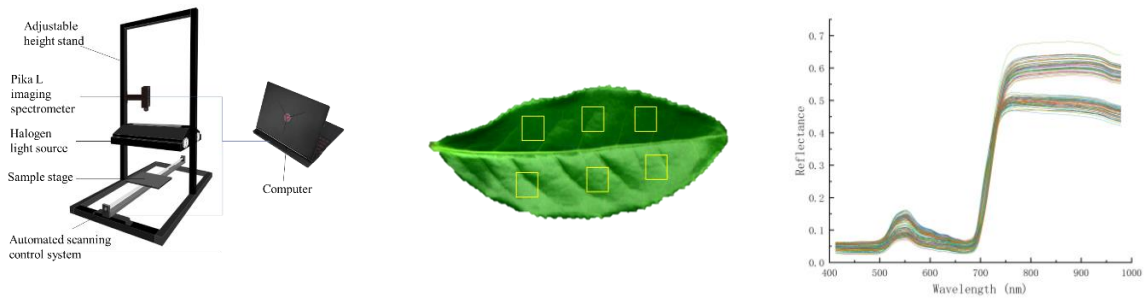
MATERIALS AND METHODS

Collection of Tea Leaf Samples

In this study, leaves at different growth stages were selected as research objects. The tea samples used in the experiment were collected from the Yangfan Agricultural High-Tech R&D and Demonstration Park in Xian'an District, Xianning City, Hubei Province (longitude: 114.44°E, latitude: 29.88°N). The experimental cultivars included Zhongcha 108, Zhongming 7, and Echa 1. Among them, 30 groups were selected for Zhongcha 108, and 15 groups each for Zhongming 7 and Echa 1. Each group contained 4 leaves. Tea leaf collection was conducted on April 23, 2024, and June 13, 2024, respectively. The size of each sampling area was 1m×1m, and a total of 480 leaves were collected in the two samplings. The collected leaves were placed in self-sealing bags, with air exhausted as much as possible, and stored in an ice box temporarily to ensure the accuracy of leaf spectral measurement.

Acquisition of Multispectral Data of Tea Leaves

The spectral imaging system was composed of a Pika L spectrometer (Resonon Inc., USA), a halogen light source, a stage, a sample scanning motion control system, and a computer, as shown in Figure 1(a). The spectral acquisition software was Spectronon Pro (Version 3.4.10), and the stage movement speed was set to 3.426 cm/s to prevent image distortion during line scanning. After preheating the halogen light source for 20 minutes, a whiteboard and a black background were used for calibration before imaging. After setting the exposure time and focusing the objective lens, the spectral images of the tea leaf samples were collected. The camera parameters were set as follows: exposure time of 29.5 ms, spectral scanning range from 377.2 nm to 1019.79 nm, spectral interval of 2.27 nm, number of scanning bands of 300, and objective lens distance of 30 cm. Using the above equipment and parameters, 6 sampling points were selected for each tea leaf sample, and the average value of these 6 regions of interest was taken as the spectral value of the entire leaf, as shown in the collection area in Figure 1(b). A spectral matrix of 120×300 (number of samples × wavelength) was obtained, and the original spectral data are shown in Figure 1(c).



(a) Spectral Imaging System (b) Sample Collection Area (c) Original Spectrum

Fig. 1 - Hyperspectral Imaging System

Determination of SPAD Values of Tea Leaves

The SPAD value of chlorophyll in tea leaves can effectively reflect the chlorophyll content in tea leaves. In this study, a SPAD-502Plus chlorophyll meter (Konica Minolta, Japan) was used to measure the SPAD values of the collected tea leaves. After measurement, the average value of the SPAD values at 4 sampling points of each leaf was calculated. The dataset was divided into a training set (96 groups) and a test set (24 groups) at a ratio of 4:1. The specific data distribution characteristics are shown in the statistical results in Table 1.

Table 1

Statistics of tea leaf SPAD values

Sample Category	Number	Minimum Value	Maximum Value	Average Value	Standard Deviation	Coefficient of Variation
Training Set	96	45.76	79.83	60.92	9.47	0.16
Test Set	24	48.43	77.19	62.45	9.14	0.15
Total	120	45.76	79.83	61.23	9.34	0.15

Selection of Vegetation Indices

Based on the spectral response law of vegetation, this study used multispectral imaging technology to collect leaf reflectance spectral data. Combined with previous research foundations, 23 vegetation indices related to key parameters such as chlorophyll sensitivity, moisture status, and canopy structure were selected, including classical parameters such as NDVI, PRI, and SR (Table 2). This index system has been verified by the spectral response mechanism and can accurately analyze the non-linear response mechanism between leaf photosynthetic pigment content and physiological status.

Table 2

Calculation formulas of vegetation indices

Vegetation Index	Formula
Normalized Difference Vegetation Index (NDVI)	$NDVI = (\rho_{nir} - \rho_r) / (\rho_{nir} + \rho_r)$
Green Normalized Difference Vegetation Index (GNDVI)	$GNDVI = (\rho_{nir} - \rho_g) / (\rho_{nir} + \rho_g)$
Normalized Difference Red Edge Index (NDRE)	$NDRE = (\rho_{nir} - \rho_{re}) / (\rho_{nir} + \rho_{re})$
Leaf Chlorophyll Index (LCI)	$LCI = (\rho_{nir} - \rho_r) / (\rho_{nir} + \rho_r)$
Difference Vegetation Index (DVI)	$DVI = \rho_{nir} - \rho_r$
Ratio Vegetation Index (RVI)	$RVI = \rho_{nir} / \rho_r$
Enhanced Vegetation Index (EVI)	$EVI = 2.5 \times (\rho_{nir} - \rho_r) / (\rho_{nir} + 6 \times \rho_r - 7.5 \times \rho_b + 1)$
Triangular Vegetation Index (TVI)	$TVI = 60 \times (\rho_{nir} - \rho_g) - 100 \times (\rho_r - \rho_g)$
Chlorophyll Green Index (CGI)	$CGI = \rho_{nir} / \rho_g - 1$
Green Difference Vegetation Index (GDVI)	$GDVI = \rho_{nir} - \rho_g$
Modified soil-adjusted vegetation index (MSAVI)	$MSAVI = 2\rho_{nir} + 1 - \sqrt{((2\rho_{nir} + 1)^2 - 8 \times (\rho_{nir} - \rho_r))} / 2$
Atmospherically Resistant Vegetation Index (ARVI)	$ARVI = (\rho_{nir} - [\rho_r - 2 \times (\rho_\beta - \rho_r)]) / (\rho_{nir} + [\rho_r - 2 \times (\rho_\beta - \rho_r)])$
Structure Insensitive Pigment Index (SIPI)	$SIPI = (\rho_{nir} - \rho_\beta) / (\rho_{nir} - \rho_r)$
Optimized Soil-Adjusted Vegetation Index (OSAVI)	$OSAVI = 1.16 \times (\rho_{nir} - \rho_r) / (\rho_{nir} + \rho_r + 0.16)$
Green Optimized Soil-Adjusted Vegetation Index (GOSAVI)	$GOSAVI = (\rho_{nir} - \rho_g) / (\rho_{nir} + \rho_g + 0.16)$
Excess Green Index (ExG)	$ExG = 2 \times \rho_g - \rho_r - \rho_b$
Excess Red Index (ExR)	$ExR = (1.4 \times \rho_r) - \rho_g - \rho_b$
Excess Green-Red Index (ExGR)	$ExGR = ExG - ExR$
Green-Red Vegetation Index (GRVI)	$GRVI = (\rho_r - \rho_g) / (\rho_g + \rho_r)$
Normalized Difference Index (NDI)	$NDI = (\rho_{nir} - \rho_g) / (\rho_{nir} + \rho_g)$
Red-Green Ratio Index (RGI)	$RGI = \rho_r / \rho_g$
Enhanced Normalized Difference Vegetation Index (ENDVI)	$ENDVI = (\rho_{nir} + \rho_g - 2 \times \rho_r) / (\rho_{nir} + \rho_g + 2 \times \rho_r)$
Simple Ratio Index (SRI)	$SRI = \rho_{nir} / \rho_r$

Note: ρb: Blue band reflectance; ρg: Green band reflectance; ρr: Red band reflectance; ρre: Red edge band reflectance; ρnir: Near-infrared band reflectance.

Data Preprocessing and Outlier Handling

To address the low Signal-to-Noise Ratio (SNR) of spectral data and significant noise interference in the 377-410 nm and 980-1019 nm spectral bands, this study excluded data in the above spectral ranges and retained 265 valid bands to improve the stability of the analysis. To enhance the anti-interference ability of the model, outlier detection and correction of the spectral data were performed based on the boxplot principle (Figure 2).

First, the Interquartile Range (IQR) method was used to set the threshold range from Q_1-1.5IQR to Q_3+1.5IQR (IQR=Q_3-Q_1) to identify discrete outliers, with outliers marked in red and the mean value marked in yellow. Second, the 3σ criterion was applied for secondary screening of non-normally distributed regions to remove observation points beyond the range of μ±3σ. Verification via the Kolmogorov-Smirnov test (p>0.05) confirmed that the processed data met the requirements of normal distribution. The analysis of spectral data dispersion showed that outliers only existed in specific wavelength ranges, and the distribution characteristics of data in other bands were consistent with theoretical expectations.

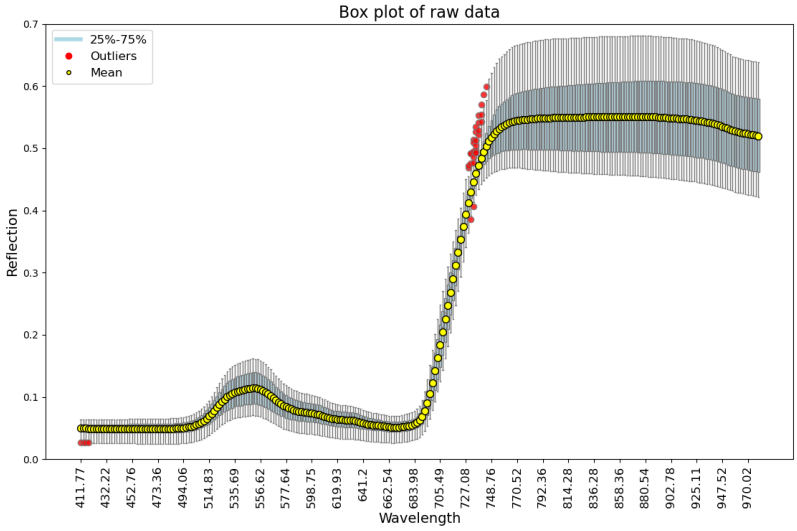


Fig. 2 - Boxplot of spectral reflectance

Detection Indicators

After outlier handling (Section 2.3.2), the standardized dataset was divided into a training set and a test set at a ratio of 4:1 for model construction and accuracy verification, respectively. A feature set was constructed based on the selected vegetation indices and their combinations. Two indicators, the coefficient of determination (R^2) and the Root Mean Square Error (RMSE), were used to evaluate the model performance. An R^2 value close to 1 indicates stronger model stability and goodness of fit, while an RMSE value close to 0 reflects better prediction ability.

The R^2 and RMSE are respectively expressed by Equations (1) and (2):

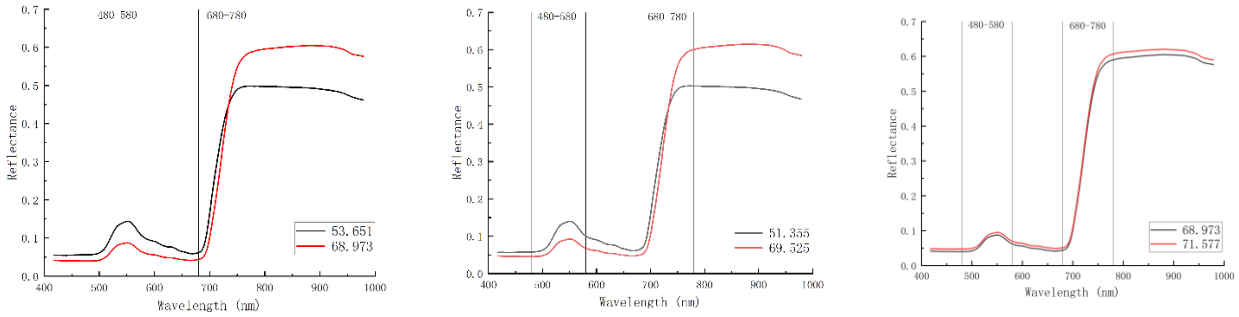
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

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

RESULTS AND DISCUSSION

Correlation Analysis between Adjacent Band Change Rate and SPAD Value

Figure 3 shows the average reflectance curves of leaves of the three tea cultivars across two growth stages. The spectral reflectance of the three tea cultivars showed a similar change trend in the same bands, but there were differences in reflectance due to variations in the content of biochemical substances among different cultivars. In the visible light region of 480-580 nm (the dominant response region of chlorophyll), the spectral curves all formed obvious peaks; the reflectance showed a rapid upward trend in the 680-780 nm range; after entering the near-infrared region of 720-1000 nm, the optical heterogeneity of leaf cell walls and intercellular spaces resulted in high reflectance. According to reference [7], there is a certain relationship between the change rate of adjacent channels and chlorophyll content. The spectral reflectance curves of each sample showed consistency in spectral shape characteristics, but there were differences in the specific reflectance values. In this study, first-order derivative processing was performed on the spectral data in two bands with large changes, and these values were used as the sample feature values, resulting in a total of 92 groups of features.



a) Zhongcha 108 *b) Zhongming 7* *c) Echa 1*

Fig. 3 - Average reflectance profiles of three tea cultivars (Zhongcha 108, Zhongming 7, and Echa 1) across two stages

The experimental results showed that the feature extraction method based on the spectral change rate of adjacent bands exhibited significant advantages in the prediction of SPAD values of tea leaves. The modeling results showed that the coefficient of determination R^2 obtained by this method reached 0.762, and the Root Mean Square Error (RMSE) was 4.3633, confirming the effectiveness of the feature engineering. In the near-infrared characteristic band range of 711.96-755.28 nm, the reflectance change rate of adjacent bands showed a strong negative correlation with the SPAD value (Pearson correlation coefficient r=-0.885~-0.895). This phenomenon is highly consistent with the spectral response characteristics of chlorophyll in plant leaves in the red edge region (680-750 nm). This band range covers the sensitive region where the strong absorption of chlorophyll transitions to high reflectance in the near-infrared region, and the first-order derivative change of adjacent bands can sensitively capture the characteristics of spectral curves under different chlorophyll concentrations. Compared with the traditional full-band modeling method, the feature selection strategy based on band gradient changes not only achieves data dimensionality reduction but also, more importantly, improves the model's ability to analyze subtle differences in chlorophyll content by extracting local differential information of the spectral curve, confirming that the adjacent spectral change rate in the infrared band can be used as an optical indicator for characterizing the SPAD value of tea leaves.

Correlation Analysis between Vegetation Indices and SPAD Value

In this study, Pearson correlation analysis was used to systematically analyze the correlation between chlorophyll content and spectral coefficients, and then sensitive vegetation indices with statistical significance were selected. The analysis results are presented in Figure 4.

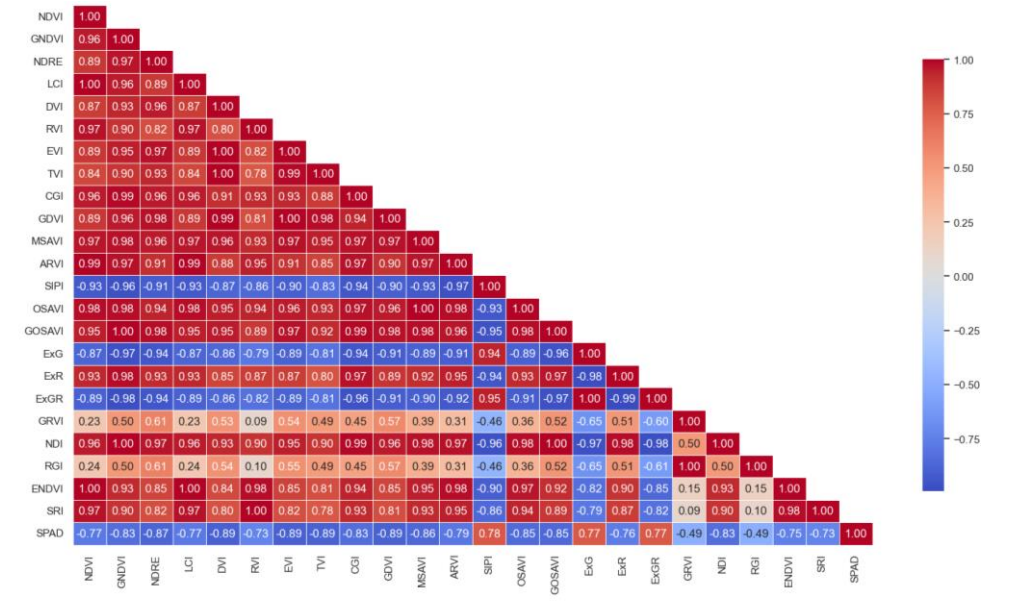


Fig. 4 - Pearson correlation coefficient matrix

The study revealed a significant multi-dimensional response mechanism between the SPAD value of tea leaves and multispectral vegetation indices. Significance tests confirmed that the correlation levels between the 23 spectral parameters and the SPAD value all reached the extremely significant threshold ($P<0.01$), among which parameters such as DVI, EVI, TVI, and GDVI were particularly prominent ($r=0.89$). This phenomenon is consistent with the mechanistic advantages of these indices in leaf area dynamic monitoring and photosynthetic pigment quantitative characterization. However, due to the multiple collinearity existing among multispectral parameters, feature selection is required to construct the optimal feature set of the model.

Model Establishment and Comparison

Hyperspectral data have significant multiple collinearity characteristics, and their high-dimensional data nature is prone to information redundancy. Therefore, it is necessary to perform feature band selection on the spectral data to improve the generalization ability of the model by eliminating band redundancy.

In this study, three feature selection methods were used to extract feature values: Competitive Adaptive Reweighted Sampling (CARS), Least Angle Regression (LARS), and Iteratively Retained Informative Variables (IRIV). The CARS algorithm is based on the Monte Carlo sampling framework, which dynamically weights the regression coefficients of iterative Partial Least Squares (PLS) to adaptively eliminate redundant variables. Through weighted sampling and exponential decay, it simultaneously realizes global search and local optimization of the feature space. The LARS algorithm adopts a piecewise linear regression strategy and achieves progressive feature selection of high-dimensional data through the projection of the least angle vector. This method maintains computational efficiency while analyzing the marginal contribution of feature variables through the least squares constraint path. The IRIV algorithm is based on the model ensemble analysis framework. It evaluates the information entropy of variables through iterative random feature combinations and uses permutation tests and variable importance in projection indicators for dual verification, effectively identifying the synergy and antagonism between spectral features. It is particularly suitable for solving the multiple collinearity problem of high-dimensional small-sample data. In this study, the CARS, LARS, and IRIV algorithms were used to select spectral sensitive factors, and the optimal variables were determined through competitive weight evaluation. Prediction models were constructed, and the performance differences were quantified based on the two indicators (R^2 and RMSE).

Seven models were used to establish a multi-spectral SPAD quantitative inversion model, covering three categories of machine learning algorithms: linear models (Partial Least Squares Regression, PLSR; Ridge Regression, RR; Multiple Linear Regression, MLR), ensemble learning (eXtreme Gradient Boosting, XGBoost; Random Forest Regression, RFR), and non-linear models (Support Vector Regression, SVR; Least Absolute Shrinkage and Selection Operator Regression, Lasso). The hyperparameters of each model were systematically optimized using the Grid Search algorithm, and the results are shown in Table 3.

Table 3

Key parameters of each model

Model	Parameter	Value
PLSR	Number of Components	5
SVR	Regularization Coefficient	1.0
Ridge Regression	Regularization Coefficient	1.0
XGBoost	Learning Rate	0.01
Lasso	Regularization Coefficient	0.05
Random Forest Regression	Maximum Depth	5

Based on the prediction results of the seven models (PLSR, SVR, RR, MLR, XGBoost, Lasso, and RFR), this study compared the effects of different spectral data collection parameters on the estimation of SPAD values of tea leaves. As shown in Table 4, the left side lists the average values of the five groups of verification results of each model under the optimal spectral data collection parameters, and the right side shows the results of each model on the verification set.

Table 4

Optimal evaluation results of different models

Algorithm Model		Training Set		Validation Set	
		R^2	RMSE	R^2	RMSE
CK	PLSR	0.797	4.197	0.748	4.483
	SVR	0.790	4.270	0.689	4.980
	RR	0.824	3.906	**0.767**	**4.313**
	MLR	0.548	6.265	0.337	7.27
	XGBoost	0.595	5.934	0.350	8.771
	Lasso	0.808	6.798	0.735	6.558
	RFR	0.969	3.772	0.698	7.277
CARS	PLSR	0.804	4.128	**0.787**	**4.118**
	SVR	0.810	4.059	0.719	4.734
	RR	0.813	4.032	0.752	4.443
	MLR	0.813	4.030	0.751	4.455
	XGBoost	0.815	3.998	0.606	5.608
	Lasso	0.813	4.031	0.750	4.462
	RFR	0.963	1.783	0.728	4.655
LARS	PLSR	0.805	4.116	0.685	5.010
	SVR	0.804	4.128	0.734	4.608
	RR	0.806	4.102	**0.743**	**4.544**
	MLR	0.618	5.905	0.426	8.608
	XGBoost	0.798	4.195	0.583	5.765
	Lasso	0.806	4.107	0.741	4.547
	RFR	0.970	1.623	0.721	4.713
IRIV	PLSR	0.804	4.123	0.685	5.016
	SVR	0.833	3.807	0.779	4.194
	RR	0.862	3.466	**0.853**	**3.428**
	MLR	0.727	4.497	0.707	4.608
	XGBoost	0.714	4.697	0.607	5.595
	Lasso	0.811	4.047	0.745	4.512
	RFR	0.969	1.633	0.704	4.858

Note: CK indicates no feature selection.

The experimental results showed that IRIV, CARS, and LARS could reduce the number of bands while maintaining accuracy. Among them, IRIV-RR achieved the best performance, with an R^2 of 0.853 and an RMSE of 3.428 on the validation set, indicating that the IRIV-RR model can better estimate the SPAD values of tea leaves.

The IRIV algorithm showed the best performance on the validation set (R^2=0.607–0.853, RMSE=3.428–5.595), with an average R^2 of 0.726 on the validation set. The CARS algorithm performed slightly lower than the IRIV algorithm on the training set and validation set, with an average validation R^2 of 0.721 and an RMSE of 4.689. Although the LARS algorithm performed well on the training set (R^2=0.801), its performance on the validation set significantly degraded (R^2=0.662, RMSE=5.399).

Comprehensive analysis showed that the model constructed based on the optimal features selected by IRIV had the best performance. Therefore, IRIV was determined as the subsequent feature selection method. By comparing the modeling performance of different feature sets, the vegetation indices and adjacent band derivative features selected by IRIV were used for combination analysis.

Comparative Analysis of Models

The modeling results showed that in the feature modeling of spectral change rate, the MLR model achieved the optimal prediction performance with 6 feature bands (R^2=0.785, RMSE=4.241).

When modeling based on vegetation indices, the MLR model reached the highest accuracy with 7 feature bands (R^2=0.791, RMSE=4.222); although the LASSO model integrating the two types of features required 13 feature bands (R^2=0.773, RMSE=4.403), it revealed the trade-off relationship between feature dimensions and prediction accuracy. The scatter regression models of measured and predicted values are shown in Figure 5.

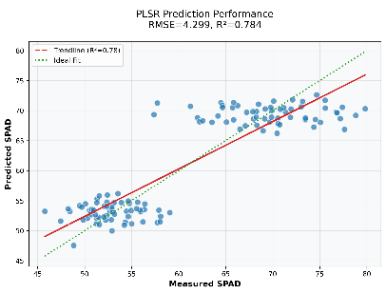
Table 5

Model comparison using different data sources

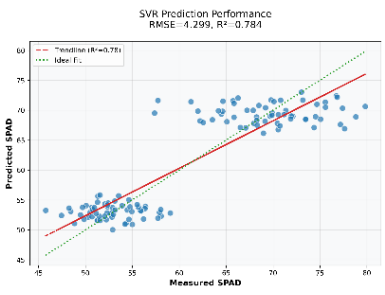
	Adjacent Band Change Rate			Vegetation Index			Adjacent Band Change Rate & Vegetation Index		
Model	Count	R^2	RMSE	Count	R^2	RMSE	Count	R^2	RMSE
PLSR	8	0.751	4.589	8	0.786	4.277	6	0.755	4.553
SVR	92	0.767	4.448	1	0.578	5.994	115	0.766	4.453
RR	6	0.742	4.674	6	0.788	4.258	9	0.754	4.560
MLR	6	0.785	4.271	7	0.791	4.222	5	0.745	4.664
XGBoost	5	0.578	5.914	6	0.619	5.566	3	0.578	5.923
Lasso	92	0.762	4.499	23	0.785	4.285	115	0.773	4.403
RFR	5	0.782	4.298	7	0.771	4.384	5	0.751	4.572

The PLSR model showed significant sensitivity in the green band (550.34 nm) and the near-infrared region (755.28-770.52 nm), and had high load coefficients for DVI, EVI, and TVI. This is closely related to the chlorophyll absorption characteristics in the green band and the spectral reflectance transition effect in the red edge region. The RR algorithm showed a specific response to the visible-near-infrared band combination and GRVI, which may be attributed to the optimization and selection mechanism of collinear features by its regularization constraint. The MLR model showed a tendency of intensive feature selection in the red edge transition zone (696.88-714.12 nm), which corresponds to the red edge shift phenomenon caused by changes in chlorophyll concentration. In contrast, the XGBoost ensemble learning model focused on two key wavelengths, 521.08 nm (the secondary absorption peak of chlorophyll) and 714.12 nm (the starting point of the red edge), and combined with ENDVI, highlighting the ability of tree models to capture non-linear interactions of features. The feature importance distribution of RFR further verified that the synergistic effect of 759.62 nm (near-infrared platform region) and 755.28 nm (red edge inflection point) is crucial for chlorophyll inversion.

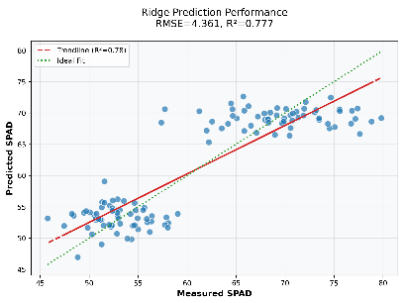
As shown in the 1:1 line diagram of the model validation set, the measured and predicted values of the model with optimal features selected by IRIV were densely distributed along the 1:1 reference line, confirming that the inversion accuracy of the SPAD value was significantly improved. This result confirms that the IRIV method can effectively select core feature bands.



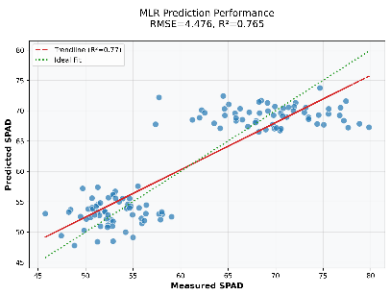
a) PLSR



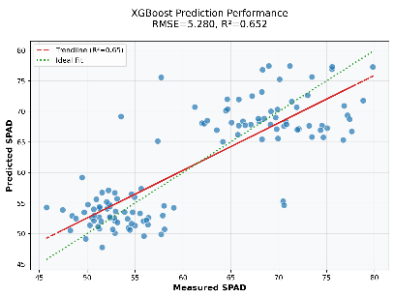
b) SVR



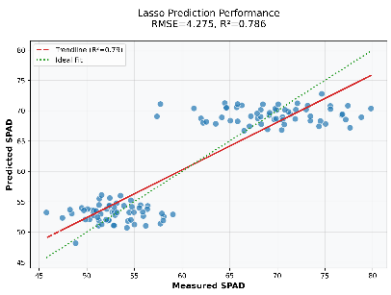
c) RR



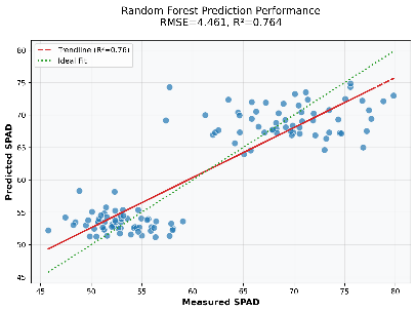
d) MLR



e) XGBoost



f) Lasso



g) RFR

Fig. 5 - Comparison between measured and predicted SPAD values

The PLSR model showed significant sensitivity in the green band (550.34 nm) and the near-infrared region (755.28-770.52 nm), and had high load coefficients for DVI, EVI, and TVI. This is closely related to the chlorophyll absorption characteristics in the green band and the spectral reflectance transition effect in the red edge region. The RR algorithm showed a specific response to the visible-near-infrared band combination and GRVI, which may be attributed to the optimization and selection mechanism of collinear features by its regularization constraint. The MLR model showed a tendency of intensive feature selection in the red edge transition zone (696.88-714.12 nm), which corresponds to the red edge shift phenomenon caused by changes in chlorophyll concentration. In contrast, the XGBoost ensemble learning model focused on two key wavelengths, 521.08 nm (the secondary absorption peak of chlorophyll) and 714.12 nm (the starting point of the red edge), and combined with ENDVI, highlighting the ability of tree models to capture non-linear interactions of features. The feature importance distribution of RFR further verified that the synergistic effect of 759.62 nm (near-infrared platform region) and 755.28 nm (red edge inflection point) is crucial for chlorophyll inversion.

As shown in the 1:1 line diagram of the model validation set, the measured and predicted values of the model with optimal features selected by IRIV were densely distributed along the 1:1 reference line, confirming that the inversion accuracy of the SPAD value was significantly improved. This result confirms that the IRIV method can effectively select core feature bands.

CONCLUSIONS

This study integrated hyperspectral imaging technology (wavelength range: 400-1000 nm) with multi-dimensional feature selection algorithms. Based on the spectral response mechanism of three tea leaf cultivars across two growth stages, a spectral feature wavelength selection model system was constructed. The main conclusions are as follows:

(1) Based on the constructed spectral dataset (sample reflectance and corresponding wavelengths), spectral curve visualization analysis revealed that there was consistency in spectral shape among different samples. The adjacent spectral change rates in two bands (480-580 nm and 680-780 nm) were extracted as feature variables, and an inversion model for the SPAD value of tea leaves was constructed (R^2=0.762, RMSE=4.36). The first-order derivative features in this band showed a significant negative correlation with the SPAD value (r=-0.885~-0.895), confirming the response sensitivity of spectral red edge parameters to chlorophyll metabolism.

(2) The vegetation indices of the samples and the reflectance of each channel of the samples were calculated as the feature values of the samples. The boxplot criterion was used to eliminate abnormal samples, and the Pearson correlation analysis framework was used to analyze the correlation characteristics between vegetation indices and SPAD values. The experiment showed that DVI, EVI, TVI, and GDVI indices had a significant negative correlation with the SPAD value of tea leaves.

(3) Three algorithms (CARS, LARS, and IRIV) were used for selection, and 21 groups of prediction models were constructed by combining seven types of machine learning algorithms (PLSR/SVR/RR/MLR/XGBoost/Lasso/RFR). Experimental comparison showed that in the pure spectral data modeling under the IRIV feature selection system, the MLR model performed the best, with R^2 and RMSE of 0.785 and 4.241, respectively; in the model using vegetation indices as sample features, the MLR model also performed the best, with R^2 and RMSE of 0.791 and 4.222, respectively; by integrating band gradient parameters (adjacent band change rate) and vegetation indices to construct a multi-dimensional feature space, experimental verification showed that the LASSO model achieved the highest accuracy in spectral-vegetation index synergistic inversion (R^2=0.773, RMSE=4.403).

REFERENCES

[1] Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R., & Sousa, J. J. (2017). Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry. *Remote Sensing*, 9(11), 1110. https://doi.org/10.3390/rs9111110

[2] Chen, X., Li, F., Chang, Q., Miao, Y., & Yu, K. (2025). Improving winter wheat plant nitrogen concentration prediction by combining proximal hyperspectral sensing and weather information with machine learning. *Computers and Electronics in Agriculture*, 232, 110072. https://doi.org/10.1016/j.compag.2025.110072

[3] Chen, Y., Wang, X., Zhang, X., Wang, D., Xu, X., & Huang, X. (2025). Parameter optimization for spectral data collection in dark environments for rice leaf chlorophyll content estimation. *Computers and Electronics in Agriculture*, 230, 109828. https://doi.org/10.1016/j.compag.2024.109828

[4] Chen, Y., Wang, X., Zhang, X., Xu, X., Huang, X., Wang, D., & Amin, A. (2024). Spectral-based estimation of chlorophyll content and determination of background interference mechanisms in low-coverage rice. *Computers and Electronics in Agriculture*, 226, 109442. https://doi.org/10.1016/j.compag.2024.109442

[5] Duan, D.-D., Liu, Z.-H., Zhao, C.-J., Zhao, Y., & Wang, F. (2024). Estimation of leaf and canopy scale tea polyphenol content based on characteristic spectral parameters. *Spectroscopy and Spectral Analysis*, 44(3), 814–820.

[6] Li, X., Wei, Z., Peng, F., Liu, J., & Han, G. (2022). Estimating the distribution of chlorophyll content in CYVCV infected lemon leaf using hyperspectral imaging. *Computers and Electronics in Agriculture*, 198, 107036. https://doi.org/10.1016/j.compag.2022.107036

[7] Li, Z., Cheng, Q., Chen, L., Yang, J., Zhai, W., Mao, B., Li, Y., Zhou, X., & Chen, Z. (2025). Enhancing winter wheat plant nitrogen content prediction across different regions: Integration of UAV spectral data and transfer learning strategies. *Computers and Electronics in Agriculture*, 234, 110322. https://doi.org/10.1016/j.compag.2025.110322

[8] Mao, Z.-H., Deng, L., Duan, F.-Z., Li, X.-J., & Qiao, D.-Y. (2020). Angle effects of vegetation indices and the influence on prediction of SPAD values in soybean and maize. *International Journal of Applied Earth Observation and Geoinformation*, 93, 102198. https://doi.org/10.1016/j.jag.2020.102198

[9] Qi, N., Yang, H., Qi, J., Li, W., Cheng, J., Yang, X., Xu, B., Xu, Z., Yang, G., & Zhao, C. (2026). Decoupling tea-bud heap structure from non-imaging hyperspectral spectra for accurate single-bud trace biochemistry retrieval. *Artificial Intelligence in Agriculture*, 16, 397–411. https://doi.org/10.1016/j.aiia.2025.11.003

[10] Qi, Q., Lu, J., Zhang, J., Zheng, G., Zhang, Q., Zhang, F., Chen, F., Fang, W., Chen, S., & Guan, Z. (2025). Enhanced UAV-based SPAD values estimation in tea chrysanthemum: An optimized and interpretable machine learning approach integrating spectral and textural information. *Smart Agricultural Technology*, 12, 101449. https://doi.org/10.1016/j.atech.2025.101449

[11] Tsuchiya, Y., Yoshida, K., Ishiguro, Y., Kawaki, J., Yamashita, H., Ikka, T., & Sonobe, R. (2025). Optimizing chlorophyll content prediction in tea leaves via spectral transformations and deep learning. *BMC Plant Biology*, 26, 26. https://doi.org/10.1186/s12870-025-07863-2

[12] Wang, X., Li, J., Zhang, J., Yang, L., Cui, W., Han, X., Qin, D., Han, G., Zhou, Q., Wang, Z., Zhao, J., & Lan, Y. (2024). Estimation of cotton SPAD based on multi-source feature fusion and voting regression ensemble learning in intercropping pattern of cotton and soybean. *Agronomy*, 14(10), 2245. https://doi.org/10.3390/agronomy14102245

[13] Wang, Y., Zhang, X., Ye, S., Yu, G., Gouda, M., Li, X., & He, Y. (2026). Dual-branch CNN-based fusion of computer vision and near-infrared spectroscopy for quantitative prediction: A case of black tea processing. *Future Foods*, 13, 100928. https://doi.org/10.1016/j.fufo.2026.100928

[14] Wu, W., Pei, G., Lu, Z., Zhou, B., Qian, X., Wang, B., & Yang, L. (2025). Lightweight multi-view fusion network for non-destructive chlorophyll and nitrogen content estimation in tea leaves using front and back RGB images. *Agronomy*, 15(10), 2355. https://doi.org/10.3390/agronomy15102355

[15] Xie, J., Chen, L., Wu, J., Li, Z., Chen, Y., Lu, M., Zou, Y., Gao, P., Shen, Z., Sun, D., Wang, W., & Li, J. (2025). Multispectral remote sensing-driven evaluation of chlorophyll in tea plant canopies. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 1–18. https://doi.org/10.1109/TGRS.2025.3607512

[16] Xu, P., Yu, J., Ma, R., Ji, Y., Hu, Q., Mao, Y., Ding, C., Li, Z., Ge, S., Deng, W.-W., & Li, X. (2024). Chlorophyll and carotenoid metabolism varies with growth temperatures among tea genotypes with different leaf colors in Camellia sinensis. *International Journal of Molecular Sciences*, 25(19), 10772. https://doi.org/10.3390/ijms251910772

[17] Yang, Z., Tian, J., Feng, K., Gong, X., & Liu, J. (2021). Application of a hyperspectral imaging system to quantify leaf-scale chlorophyll, nitrogen and chlorophyll fluorescence parameters in grapevine. *Plant Physiology and Biochemistry*, 166, 723–737. https://doi.org/10.1016/j.plaphy.2021.06.015

[18] Yin, Q., Zhang, Y., Li, W., Wang, J., Wang, W., Ahmad, I., Zhou, G., & Huo, Z. (2023). Estimation of winter wheat SPAD values based on UAV multispectral remote sensing. *Remote Sensing*, 15(14), 3595. https://doi.org/10.3390/rs15143595

[19] Zhang, C., & Kovacs, J. M. (2012). The application of small unmanned aerial systems for precision agriculture: a review. *Precision Agriculture*, 13, 693–712. https://doi.org/10.1007/s11119-012-9274-5

[20] Zhang, D., Zhang, J., Peng, B., Wu, T., Jiao, Z., Lu, Y., Li, G., Fan, X., Shen, S., Gu, A., & Zhao, J. (2023). Hyperspectral model based on genetic algorithm and SA-1DCNN for predicting Chinese cabbage chlorophyll content. *Scientia Horticulturae*, 321, 112334. https://doi.org/10.1016/j.scienta.2023.112334