frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 25 April 2024
DOI 10.3389/fpls.2024.1396183

OPEN ACCESS

EDITED BY
Ting Sun,
China Jiliang University, China

REVIEWED BY
Muhammad Ahmad Hassan,
Anhui Academy of Agricultural Sciences, China
Pan Gao,
Shihezi University, China
Haibing He,
Anhui Agricultural University, China

*CORRESPONDENCE
Jikai Liu
liujk@ahstu.edu.cn
Xinwei Li
lixw@ahstu.edu.cn

†These authors contributed equally to this work

RECEIVED 05 March 2024
ACCEPTED 11 April 2024
PUBLISHED 25 April 2024

CITATION
Nian Y, Su X, Yue H, Zhu Y, Li J, Wang W, Sheng Y, Ma Q, Liu J and Li X (2024) Estimation of the rice aboveground biomass based on the first derivative spectrum and Boruta algorithm.
Front. Plant Sci. 15:1396183.
doi: 10.3389/fpls.2024.1396183

Estimation of the rice aboveground biomass based on the first derivative spectrum and Boruta algorithm

Ying Nian[1], Xiangxiang Su[1], Hu Yue[1], Yongji Zhu[1], Jun Li[1], Weiqiang Wang[1], Yali Sheng[1], Qiang Ma[1], Jikai Liu[1,2*†] and Xinwei Li[1,2,3*†]

[1]College of Resource and Environment, Anhui Science and Technology University, Chuzhou, China, [2]Anhui Province Crop Intelligent Planting and Processing Technology Engineering Research Center, Anhui Science and Technology University, Chuzhou, Anhui, China, [3]Anhui Province Agricultural Waste Fertilizer Utilization and Cultivated Land Quality Improvement Engineering Research Center, Anhui Science and Technology University, Chuzhou, China

Aboveground biomass (AGB) is regarded as a critical variable in monitoring crop growth and yield. The use of hyperspectral remote sensing has emerged as a viable method for the rapid and precise monitoring of AGB. Due to the extensive dimensionality and volume of hyperspectral data, it is crucial to effectively reduce data dimensionality and select sensitive spectral features to enhance the accuracy of rice AGB estimation models. At present, derivative transform and feature selection algorithms have become important means to solve this problem. However, few studies have systematically evaluated the impact of derivative spectrum combined with feature selection algorithm on rice AGB estimation. To this end, at the Xiaogang Village (Chuzhou City, China) Experimental Base in 2020, this study used an ASD FieldSpec handheld 2 ground spectrometer (Analytical Spectroscopy Devices, Boulder, Colorado, USA) to obtain canopy spectral data at the critical growth stage (tillering, jointing, booting, heading, and maturity stages) of rice, and evaluated the performance of the recursive feature elimination (RFE) and Boruta feature selection algorithm through partial least squares regression (PLSR), principal component regression (PCR), support vector machine (SVM) and ridge regression (RR). Moreover, we analyzed the importance of the optimal derivative spectrum. The findings indicate that (1) as the growth stage progresses, the correlation between rice canopy spectrum and AGB shows a trend from high to low, among which the first derivative spectrum (FD) has the strongest correlation with AGB. (2) The number of feature bands selected by the Boruta algorithm is 19~35, which has a good dimensionality reduction effect. (3) The combination of FD-Boruta-PCR (FB-PCR) demonstrated the best performance in estimating rice AGB, with an increase in R^2 of approximately 10% ~ 20% and a decrease in RMSE of approximately 0.08% ~ 14%. (4) The best estimation stage is the booting stage, with R^2 values between 0.60 and 0.74 and RMSE values between 1288.23 and 1554.82 kg/hm^2. This study confirms the accuracy of hyperspectral remote sensing in estimating vegetation biomass and further explores the theoretical foundation and future direction for monitoring rice growth dynamics.

KEYWORDS

rice, AGB, remote sensing, hyperspectral data, derivative transform, machine learning algorithm

1 Introduction

Rice (*Oryza sativa*), a herbaceous plant belonging to the genus *Oryza*, has a long history of cultivation and a broad planting region (Krishnan et al., 2011). Aboveground biomass (AGB) serves as a critical indicator for describing crop growth (Wang et al., 2016). Applying scientific methods to estimate rice AGB to improve crop yields is crucial for human food security and increased economic benefits (Spiertz and Ewert, 2009; Devia et al., 2019; Cao et al., 2021b). The conventional method for acquiring AGB primarily involves field sampling, which is both time-consuming and laborious (Gnyp et al., 2014). In contrast, remote sensing technology provides a fast and nondestructive method, providing a scientific basis and technical support for accurate estimation of AGB (Schino et al., 2003). Among them, satellite remote sensing platforms is commonly used for AGB estimation in large areas, such as forests (Lucas et al., 2020), grasslands (Fang et al., 2018; Yang et al., 2018), However, it is easily affected by spatial conditions and it is difficult to obtain satisfactory results at small field scales (Yue et al., 2019). UAV remote sensing platforms can perform low-altitude operations and obtain high-resolution images of crops, but the spectral resolution is low, which is not conducive to in-depth exploration of the spectral response of crops (Bansod et al., 2017). In comparison, near-ground hyperspectral remote sensing is not only suitable for small-scale crop monitoring, but can also acquire fine spectral data with high resolution and a broad spectrum of wavelengths, encapsulating abundant crop phenotype information and thereby offering increased opportunities for estimating the physical and chemical parameters of vegetation (Galvão et al., 2005; Atzberger et al., 2015).

To enhance the AGB estimation performance via hyperspectral remote sensing, the original spectral data are typically preprocessed to reduce noise and increase the accuracy of the data (Mishra et al., 2020). The Savitzky–Golay (SG) smoothing algorithm and derivative transform are the principal methods for preprocessing. The SG smoothing algorithm helps to mitigate interference signals within spectral data and minimize random errors, thereby improving data reliability (Savitzky and Golay, 1964; Liu et al., 2014). Derivative transform spectrum emphasizes the absorption and reflection characteristics of the subject matter and augments the differentiation in spectral information (Gerretzen et al., 2015). For instance, Wang et al. (2004) demonstrated that the derivative spectrum can lessen the impact of noise, exhibits a high correlation with rice AGB, and performs effectively in estimating rice AGB. However, spectral data still face challenges such as excessive information redundancy and high dimensionality after preprocessing. There is a need for a reliable method that can reduce the dimensionality of spectral data and eliminate high correlations between bands.

In previous research, Some researchers have used a single vegetation index to estimate rice AGB and achieved good results. However, the vegetation indices is usually a combination of several sensitive wavelengths, ignoring a large volume of spectral information. In the 2020s, some scholars adopted complex dimensionality reduction methods to improve the utilization efficiency of rice spectral data. For example, Jia et al. (2022) used the continuous projection algorithm (SPA) to reduce the hyperspectral dimension and identified 12 characteristic bands to estimate the soil plant analysis development (SPAD) content of rice. Yu et al. (2020) utilized principal component analysis (PCA) to extract five principal components for the inversion of rice leaf nitrogen deficiency. The above studies used different dimensionality reduction methods to reduce spectral data redundancy, which is of positive significance for applying hyperspectral remote sensing technology to estimate rice AGB accurately. However, they present limitations. The objective of SPA is to eliminate combinations of variables that exhibit low collinearity and include distinctive information; however, this approach might omit some initial characteristics in the spectral data (Cao et al., 2021a). PCA is capable of using several independent principal components to depict the original data, yet this technique is confined to linear projection and exhibits inadequate performance in managing nonlinear data relationships (Moore, 1981; Abdi and Williams, 2010). Consequently, the effectiveness of the monitoring model in practical scenarios may be affected by different dimensionality reduction methods (Fu et al., 2020). Given these considerations, selecting a suitable dimensionality reduction method is essential for accurately identifying sensitive bands and enhancing the accuracy of the estimation model.

As a data dimensionality reduction method, the core of Boruta algorithm is to find all feature bands related to the dependent variable. Compared with other commonly used feature screening algorithms, Compared with other commonly used feature screening algorithms (Kursa and Rudnicki, 2010; Kursa et al., 2010). The algorithm was originally used in biological and medical fields and has since been used in agricultural research as well. For instance, Li et al. (2022) demonstrated that filtering spectral indices from winter wheat canopy hyperspectral data using the Boruta algorithm enhances data validity and establishes an accurate yield estimation model. The Boruta algorithm has great potential in physical and chemical parameter estimation. To further evaluate the ability of the Boruta algorithm to estimate rice AGB, this study also selected the RFE feature selection algorithm for comparison. RFE, an adaptive feature selection algorithm, iteratively removes the least important feature variables and ranks feature importance until the optimal feature subset is identified. This approach minimizes the effects of random fluctuations and interference information (Tunca et al., 2023). Yoosefzadeh-Najafabadi et al. (2021) applied RFE to determine the optimal wavelengths from hyperspectral data for soybean yield prediction. This indicates that these feature selection methods are effective in improving estimation models' accuracy. However, relatively few studies have utilized the Boruta algorithm and the RFE algorithm for feature selection of derivative spectra to construct rice AGB estimation models.

Therefore, this study used the Boruta and RFE algorithms to eliminate interference information, identify sensitive bands, and integrate partial least squares regression (PLSR), principal component regression (PCR), support vector machine (SVM) and ridge regression (RR) to develop an AGB estimation model. Additionally, this study compares the estimation effects of various combinations of dimensionality reduction algorithms and machine learning models to identify the most accurate estimation results.

The research objectives were to (1) compare the correlations between derivative spectrum of different orders and rice AGB to determine the optimal order; (2) evaluate the effectiveness of the Boruta algorithm and RFE algorithm in selecting feature bands to determine the best dimensionality reduction method; (3) evaluate the estimation capability of machine learning algorithms combined with feature selection algorithms on rice AGB; and (4) identify the optimal combination of models and the most suitable estimation stage.

2 Materials and methods

2.1 Experimental design

The study was conducted in Xiaogang village, Fengyang County, Chuzhou city, Anhui Province, China (longitude 117°46′7″E, latitude 32°48′52″N) (Figure 1A), which is characterized by a subtropical monsoon climate, an average annual temperature of 15.4°C, and an average annual precipitation of 1179.2 mm. The soil predominantly consists of clay with medium fertility on flat terrain. The experimental area included 36 plots, measuring 2 m × 8 m each. The rice varieties tested were Runzhuxiangzhan (V1), Runzhuyinzhan (V2), and Hongxiangnuo (V3), with each plot replicated three times. Four nitrogen gradient treatments were applied with nitrogen application rates of N0 (0 kg/hm^2), N1 (100 kg/hm^2), N2 (200 kg/hm^2), and N3 (300 kg/hm^2). Phosphorus (90 kg/hm^2) and potash (135 kg/hm^2) fertilizers were applied once as basal fertilizer (Figure 1B). The experiment spanned from June 2020 to October 2020. The plants experienced no drought or flooding during the seedling stage and favorable conditions of abundant sunshine and moderate temperatures during the mid-growth stage, which are conducive to rice growth and development. Field management practices followed local cultivation methods and pest control technologies.

2.2 Data acquisition

2.2.1 Ground data acquisition

Rice samples were collected on July 25 (tillering stage), August 13 (jointing stage), August 23 (booting stage), August 31 (heading stage) and September 20 (maturity stage). The sampling time was consistent with the hyperspectral data acquisition time. Rice samples were collected from three holes in each plot, and the locations of the sampling points were marked after each sampling to avoid duplicate sampling.

Rice samples were subsequently sent to the laboratory, after which stem, leaf, and ear separation was performed (Figure 1C). The sample was placed in an oven, dried for 0.5 h after the temperature increased to 105°C, and then adjusted to 75°C for more than 24 h until the mass was constant. The dry mass of the sample was obtained by summing the dry mass of the stem, leaves and ears of the plant. Finally, the aboveground biomass of the rice plants in each plot was obtained through the dry mass of the sample and the row spacing during rice planting.

2.2.2 Hyperspectral data collection

Hyperspectral data on the rice canopy spectra were collected using a handheld ground spectrometer (ASD FieldSpec® HandHeld^{TM}2) covering a wavelength range from 325 to 1075 nm with a resolution of 1 nm. Measurements were conducted between 10:00 a.m. and 2:00 p.m. for each test. Prior to each measurement session, the instrument was calibrated against a standard radiation white plate and recalibrated after every two blocks to ensure accuracy. The data were collected at three evenly distributed points along the diagonal of each plot, with the spectrometer positioned vertically 0.5 m above the rice canopy (Figure 1C). The spectral data from the same plot were processed collectively, and the average spectral reflectance of the plot was calculated.

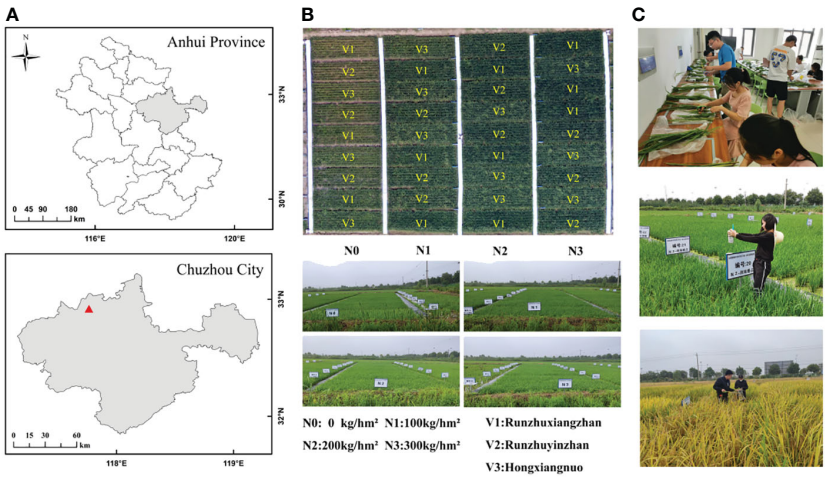

FIGURE 1
Rice experimental area **(A, B)** and data collection **(C)**.

2.3 Hyperspectral data processing

2.3.1 SG smoothing and derivative transform

The Savitzky–Golay (SG) smoothing algorithm, a digital signal processing technique proposed by Savitzky and Golay, is employed to enhance signal frequency and eliminate data noise (Savitzky and Golay, 1964). The smoothing effect of the algorithm, which varies with the window length, preserves the original shape and peak heights of the wave signal (Schafer, 2011). To minimize data fluctuations, capture the overall trend of the spectral curves, and accurately analyze the spectral characteristics of rice, this study applied SG smoothing to the canopy spectral curves from 325 to 1075 nm recorded during the growth stage (Chen et al., 2020; Shi et al., 2021). However, due to the influence of noise, only the 400 to 900 nm range was selected for analysis.

Derivative transform serves to diminish background signals, with the FD highlighting absorption peaks and shoulders in the original spectrum (OS) (Demetriades-Shah et al., 1990; Becker et al., 2005). By differentiating peaks and valleys, more precise spectral reflectance data can be obtained (Ji et al., 2022; Zhao et al., 2022). The second derivative spectrum (SD) facilitates the accurate identification of absorption peak and shoulder positions within the OS, enhancing spectral reflectance differentiation and eliminating baseline offsets in spectral reflectance data (Yang et al., 2011). This study utilized the derivative function within the Origin software's data processing menu (Origin 2018) for these transformations.

2.3.2 Feature selection

Feature selection algorithms play a pivotal role in optimizing datasets prior to model construction. In this investigation, both the Boruta algorithm and the RFE algorithm were employed for feature selection.

The Boruta algorithm (Prasad et al., 2019) is an innovative feature selection method derived from the random forest approach. It identifies all characteristic bands in spectral data that are related to dependent variables, thereby elucidating the relationship between spectral characteristics and rice AGB. The foundational concept involves shuffling the original parameters to create shadow parameters, which are then combined with the original parameters into a feature matrix for training. Based on the importance scores from the random forest, the shadow parameters are ranked by importance, with the highest value designated the Z score. Original parameters more significant than the Z score are labeled "most important", while those below are deemed "unimportant" and thus eliminated. Ultimately, the most important original parameters are selected as the optimal combination for constructing the inversion model.

Recursive feature elimination (RFE) (Guyon et al., 2002) is a wrapper-based selection method that starts with the construction of an initial model, ranking all feature bands by their importance, and iteratively removing the least significant features. By retraining the model with the remaining features and repeating this process, it gradually identifies the most critical subset of features. This algorithm excels by iteratively evaluating the contribution of each feature band to the model, thereby identifying the optimal feature subset to reveal the underlying structure and patterns within the data.

2.4 Model construction method

To thoroughly investigate the relationships between the independent and dependent variables, four machine learning models were selected for comparison: PLSR, PCR, SVM, and RR.

PLSR (Geladi and Kowalski, 1986) is a multivariate linear statistical method that focuses on finding a hyperplane that minimizes the variance between the response and independent variables. It projects predictor and observed variables into a new space to derive a linear regression model, known as a bilinear factor model, due to the presence of both data X and Y in this new space.

PCR (Kawano et al., 2018) employs PCA for predictive data mining, transforming original variables into principal components that are linear combinations of the original variables and mutually independent. Regression analysis is then performed on these principal components to derive the regression equation, which is subsequently applied to the original variables.

SVM (Suykens and Vandewalle, 1999) operates as a supervised learning model utilized for analyzing data in classification and regression analysis. The core concept involves identifying the optimal classification hyperplane for two types of samples in the original space when they are linearly separable. In instances where linear separability is not achievable, the samples are projected into a high-dimensional feature space where an optimal hyperplane is determined. This hyperplane classifies the samples, aiming to minimize the distance within similar classes and maximize the distance between distinct classes.

RR (Ivanda et al., 2021) serves as a technique for estimating the coefficients of a multiple regression model in scenarios where the independent variables exhibit high correlation. This is achieved by incorporating an L2 regularization term into the OLS loss function. The L2 regularization term is calculated as the product of the sum of the squares of the coefficients and a regularization parameter λ (lambda).

2.5 Accuracy evaluation

In this study, 70% of the sample data (n=25) were selected as the modeling set for each growth stage, and 30% of the sample data (n=11) were used as the validation set to construct the rice AGB estimation model. The coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) were used to evaluate the model accuracy. The closer R^2 is to 1, the lower the RMSE and MAE are, and the higher the accuracy of the estimation model is. R^2, RMSE and MAE are calculated using Equations 1–3, respectively:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \hat{x})^2} \quad (1)$$

10.3389/fpls.2024.1396183

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

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \quad (3)$$

Note: x_i, y_i, and $\hat{x}$ are the actual measured value, predicted value and average value of the measured data, respectively; n is the number of samples.

3 Results

SG smoothing was applied to the spectral curves within the 400 to 900 nm range, The processing results of OS, FD and SD at different growth stages are shown in Figure 2.

The processing results for the OS reveal the characteristic spectral profile of the rice canopy, marked by typical green plant reflectance features, with spectral reflectance values ranging from 0 to 0.5. Notably, a reflection peak and an absorption valley are observed near wavelengths of 550 nm and 680 nm, respectively, both of which are situated within the visible light spectrum. The spectral reflectance of rice rapidly increases within the 680 to 750 nm range, creating a "red edge" phenomenon (Gitelson et al., 1996). The first peak of the FD curve appears at 500 ~ 550 nm. The band range of 680 ~ 750 nm shows a drastic change that first rises and then falls. The SD curve exhibits continuous fluctuations across the 500 to 690 nm band range, with two notable peaks within the 700 to 800 nm band range.

3.1 Correlations between different orders of derivative spectrum and rice AGB

To fully explore the sensitivity of the different orders of rice canopy spectra and to compare the effects of spectral transformations on AGB, FD and SD transformations were carried out on the basis of OS, and their correlations with rice AGB were investigated (Figure 3).

During the full growth stage, the correlation between OS and AGB showed a consistent trend from negative to positive as the band increased, with the maximum correlation at 720 nm. The correlation curve between FD and AGB shows a continuous fluctuation trend in the 400 ~ 900 nm band, among which the 680 ~ 750 nm band has the highest correlation, with a correlation coefficient between 0.60 ~ 0.85. The correlation curve between SD and AGB was similar to that between FD and AGB, but the fluctuations were more severe.

In summary, the derivative spectrum of the different orders were compared with the maximum absolute value of the AGB correlation coefficient (Figure 4). Except for the tillering stage, the maximum values in the other four periods all occurred at FD. Comprehensive analysis revealed that the FD could better reflect the growth status of rice, so the FD was used as the input variable of the feature screening algorithm for subsequent construction of the AGB estimation model.

3.2 Feature band selection

An AGB estimation model utilizing PLSR, PCR, SVM, and RR was developed to evaluate the effectiveness of the band screening algorithms. This model incorporated characteristic bands selected by both the Boruta algorithm and the RFE algorithm, as well as the full band as input variables. The diversity and scope of the characteristic bands selected by these feature screening algorithms during different growth stages are summarized in Table 1. Detailed feature selection results are shown in Appendix 2.

The Boruta algorithm demonstrated remarkable stability, with the number of characteristic wavelengths in each growth stage not exceeding 40. In contrast, the RFE algorithm showed more significant fluctuations, with 3 characteristic wavelengths identified at both the tillering and booting stages and 439 at the maturity stage, indicating substantial variance in the number of wavelengths identified across different growth stages. Notably, the characteristic wavelengths identified by both screening algorithms predominantly fall within the "red edge" spectral range, highlighting their importance in estimating rice AGB.

3.3 Establishment and evaluation of the rice AGB estimation model based on FD

This study employs three types of model input variables: first derivative full band (FD-FB), the characteristic bands selected by

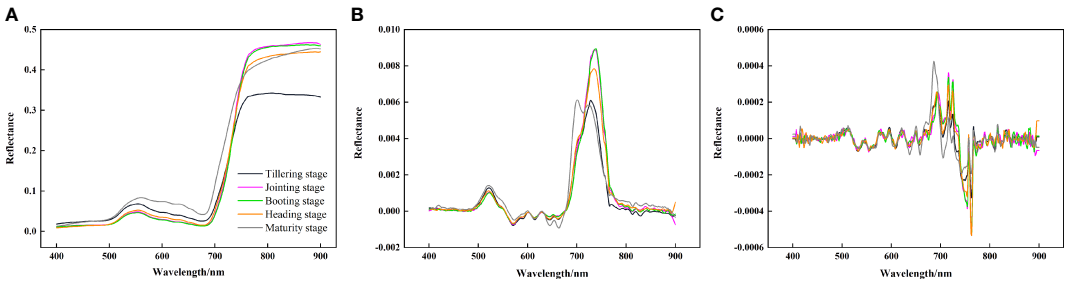

FIGURE 2
Derivative transformation spectral curves of rice canopy at different growth stages. **(A)** original spectrum, **(B)** first derivative spectrum, **(C)** second derivative spectrum.

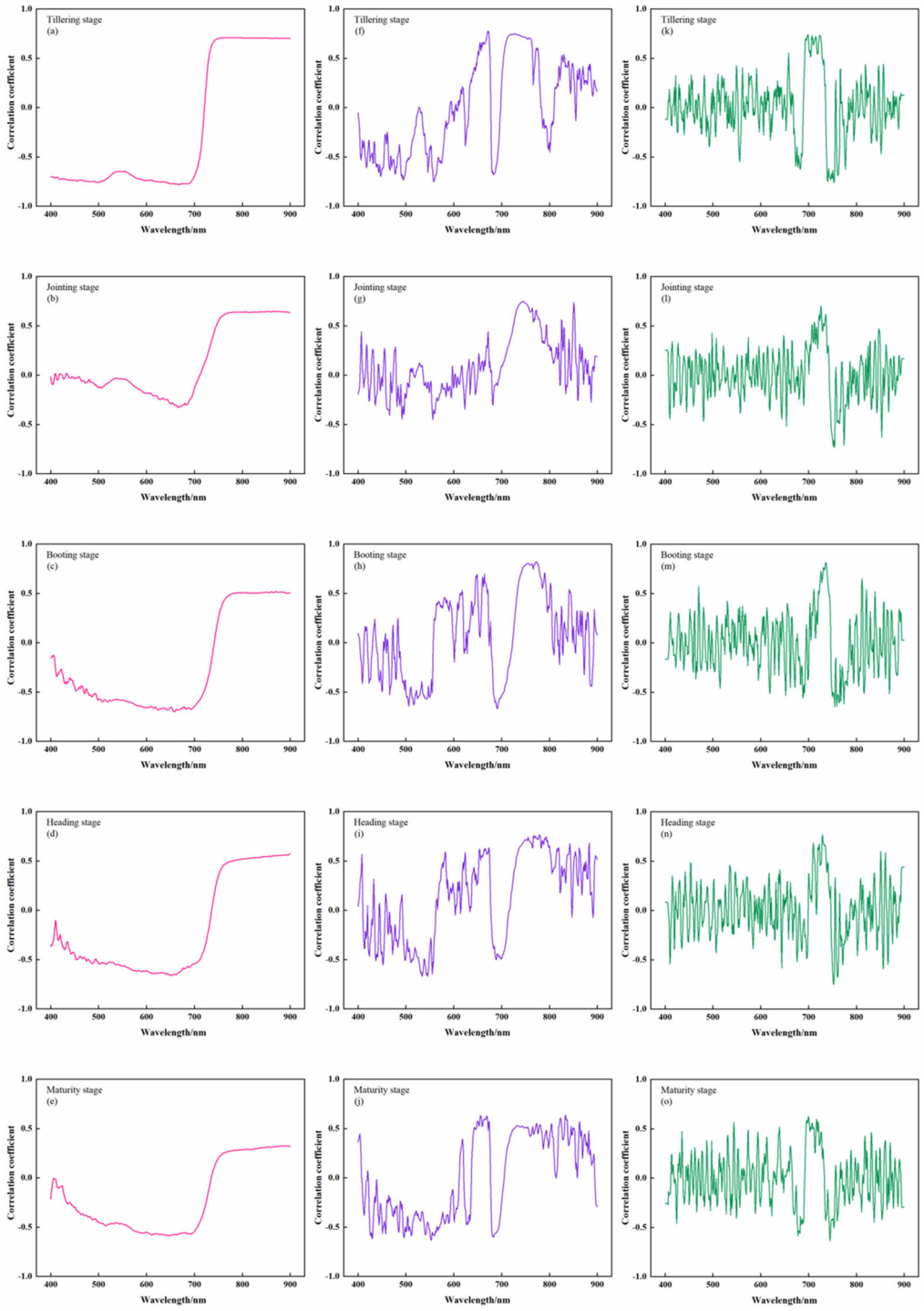

FIGURE 3
Correlation between rice canopy spectra and AGB at different growth stages. **(A–E)** The correlation between OS and AGB; **(F–J)** the correlation between FD and AGB; **(K–O)** the correlation between SD and AGB.

the Boruta algorithm (FD-Boruta), and the characteristic bands selected by the RFE algorithm (FD-RFE). Sixty AGB estimation models were developed for various growth stages using four machine learning algorithms (PLSR, PCR, SVM, RR), with R^2 (Figure 5), RMSE (Table 2) and MAE (Appendix 1) serving as the model evaluation metrics.

Initially, the AGB estimation abilities of different feature selection algorithms were compared within the same model framework. In the PLSR model, the estimation precision for FD-Boruta and FD-RFE was found to be comparable, with R^2 values ranging from 0.43 to 0.70, and RMSE values between 464.34 and 2515.77 kg/hm^2 and 493.81 and 2407.61 kg/hm^2, respectively. These

10.3389/fpls.2024.1396183

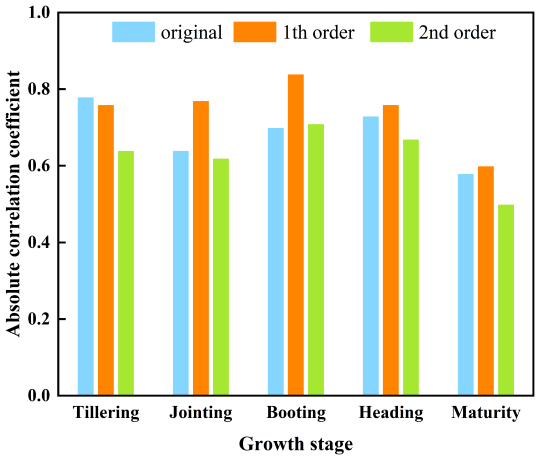

FIGURE 4
The maximum absolute value of the correlation coefficient between different order derivative spectra and AGB.

results are more accurate than those obtained using FD-FB, with R^2 values improving by 0 to 0.06 and RMSE values decreasing by 29.47 to 148.91 kg/hm^2. Among the PCR models, FD-Boruta exhibited the highest estimation accuracy, followed by FD-RFE, with FD-FB showing the least accuracy. The R^2 accuracy of FD-Boruta improved by 0.03, 0.19, 0.02, 0.01, and 0.04, respectively, compared to FD-RFE. In contrast, the R^2 accuracy of FD-RFE improved by 0.02, 0.05, 0.07, and 0.01 compared to FD-FB. The outcomes for the SVM and RR models were analogous, although the performances of the three algorithms varied significantly. R^2 ranged from 0.58 to 0.71 for FD-Boruta, 0.22 to 0.74 for FD-RFE, and 0.19 to 0.60 for FD-FB.

Second, the estimation accuracy of AGB by different models under the same feature selection algorithm was compared. For the model constructed with FD-Boruta inputs, when the estimation accuracy was similar, FD-Boruta-PCR demonstrated more stable performance. The estimation outcomes of models based on FD-RFE varied significantly. FD-RFE-PLSR and FD-RFE-PCR provided more precise estimates of rice AGB, whereas the FD-RFE-SVM and FD-RFE-RR models exhibited overfitting in certain stages. The models based on FD-RFE generally showed poor performance, with model R^2 values below 0.65, and the accuracy of the FD-FB-SVM and FD-FB-RR models varied greatly.

Finally, the performance of all model combinations was assessed for estimating AGB effects across different growth stages. In the tillering stage, FD-Boruta-SVM yielded the best estimation, with an R^2 of 0.66 and an RMSE of 34.33 kg/hm^2. During the jointing stage, FD-Boruta-SVM achieved the highest estimation accuracy, with an R^2 of 0.65 and an RMSE of 985.55 kg/hm^2, while FD-FB-SVM showed the lowest accuracy, with an R^2 of 0.32 and an RMSE of 1412.37 kg/hm2. In the booting stage, the estimation accuracies R^2 of models using FD-FB and FD-RFE inputs were both above 0.7, with FD-RFE-SVM and FD-Boruta-PCR performing the best, achieving R^2 values of 0.74 and 0.72, respectively, and RMSE values of 1311.04 kg/hm^2 and 1290.24 kg/hm^2, respectively. At the heading stage, FD-Boruta-PCR had the highest estimation accuracy, with an R^2 of 0.71, while FD-FB-SVM had the lowest accuracy, with an R^2 of 0.30. In the maturity stage, the AGB estimation accuracy of all the models decreased, with R^2 values ranging from 0.19 to 0.53.

In conclusion, models built using the Boruta algorithm are more reliable and exhibit greater stability, among which FD-Boruta-PCR is the most stable, followed by FD-Boruta-PLSR. Considering the model estimation accuracy, the booting period is the most accurately estimated, with R^2 values exceeding 0.7.

TABLE 1 Comparison of feature selection results between the Boruta algorithm and RFE algorithm.

Feature selection algorithm	Growth Stage	Number	Feature band range (nm)
Boruta	Tillering	19	724 ~ 761
	Jointing	35	644, 721 ~ 771, 816, 817
	Booting	27	664 ~ 694, 717 ~ 774
	Heading	24	414, 416, 699 ~ 777,817, 873
	Maturity	30	420, 498, 686 ~ 759, 842
RFE	Tillering	3	742, 751, 761
	Jointing	392	402 ~ 900
	Booting	3	753, 754, 764
	Heading	127	402 ~ 495, 507 ~ 594, 614 ~ 699, 702 ~ 900
	Maturity	439	400 ~ 900

4 Discussion

4.1 Correlations between OS, FD, SD and AGB

An in-depth analysis of the correlation between spectral information and AGB will facilitate a comprehensive understanding of the growth status of rice. This study performed a correlation analysis between different orders of derivative spectrum and AGB at different growth stages of rice. Generally, the correlation between the derivative spectrum of various orders and AGB tended to increase and then decrease throughout the growth stage, potentially due to the influence of the rice growth cycle (Xu et al., 2023). From the tillering stage to heading stage, as the chlorophyll content in rice leaves increases and the vegetation canopy structure becomes fully developed, the canopy becomes richer in spectral information, which can more accurately reflect crop characteristics (Ren et al., 2018). In the mature stage, as stems and leaves progressively wither and yellow, the chlorophyll content drops sharply, the spectral characteristics obtained are less able to accurately represent crop growth, leading to a decrease in the correlation between the canopy spectrum and AGB (Yang and Chen, 2004; He et al., 2019; Zhang et al., 2019).

Within the 680~750 nm band range, the correlation between the canopy spectrum and AGB reached its maximum value in different

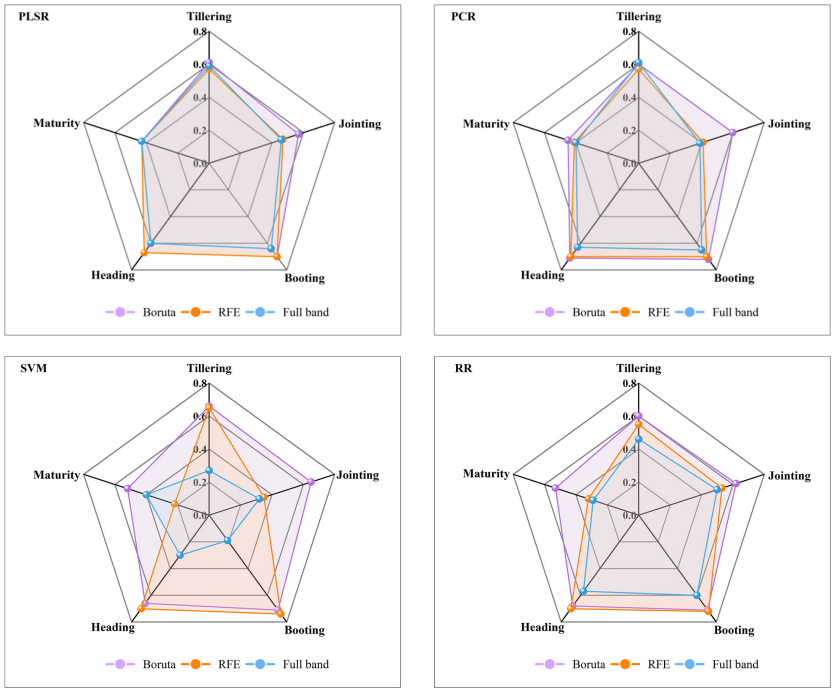

FIGURE 5
Comparison of the R^2 values of the different estimation models for the validation sets. PLSR, partial least squares regression; PCR, principal component regression; SVM, support vector machine; RR, ridge regression.

stages (Figure 3). This occurs because this band serves as the transition region between the infrared and near-infrared bands, where spectral reflectance transitions rapidly from a low negative correlation to a high positive correlation. This shift is attributed to strong absorption and reflection (Kanke et al., 2016; Xu et al., 2022). The correlation between FD and AGB was generally greater than that between OS and SD at the different fertility stages (Figure 4). This observation aligns with the findings of Liang et al. (2018) and Tong et al. (2023). As a method for analyzing spectral information, derivative transform can diminish noise and enhance data accuracy (Li and Xie, 2015). FD reflects the slope of the spectrum, while SD represents the change in the slope of the reflection spectrum. While SD identifies more absorption peaks, it also introduces noise and may result in errors (Shen et al., 2022). Therefore, FD is strongly correlated with AGB and can be effectively utilized for estimating rice AGB.

4.2 Rice AGB estimation based on different feature selection algorithms

In the realm of feature selection, prior research has demonstrated that employing screened feature variables for model construction enhances the estimation power, inversion accuracy, and utility of the original models (Gao et al., 2019; Sun et al., 2021). For example, Yu et al. (2023) used the SPA algorithm to extract sensitive bands in rice canopy spectral data, and combined with PROSAIL to establish a rice AGB estimation model, achieving high accuracy (R^2=0.69). This study yielded similar findings, the characteristic band had a greater estimation ability than the original band. The difference is that previous research is usually based on OS and does not involve derivative transform. Therefore, whether the estimation model constructed by derivative transformed spectrum is better than OS remains to be determined. This study performs SG smoothing and derivative transform on OS and selects characteristic bands based on the optimal derivative spectrum. The results show that the Boruta algorithm is more robust than the RFE algorithm, and the selected feature bands are more sensitive, which is conducive to constructing subsequent rice AGB estimation model.

Under identical conditions, the most accurate AGB estimates were achieved using feature bands selected by the Boruta algorithm. This superiority of the Boruta algorithm is attributed to its ability to identify bands of features that are genuinely relevant to the dependent variable, thereby enhancing the prediction accuracy of the model (Kursa and Rudnicki, 2010), a conclusion that aligns with the findings of Degenhardt et al. (2019). However, the RFE algorithm exhibited suboptimal performance in the jointing, heading, and maturity stages. This could be due to the RFE algorithm generates feature subsets with corresponding accuracy by continuously building models. This process may result in retaining a large number of features, leading to significant collinearity between bands and diminishing the model's estimation accuracy (Paul et al., 2015; Chen et al., 2018), echoing the research results of Lin et al. (2023). The poorest performance was observed when the entire band was utilized as an input variable, attributed to the presence of redundant information and increased collinearity among bands, which impaired the model's estimation capability (Hansen and Schjoerring, 2003; Pan et al., 2017).

TABLE 2 Rice AGB estimation model under different machine learning algorithms (R^2, RMSE).

machine learning algorithm	Bandselection algorithm	Tillering				Jointing				Booting				Heading				Maturity			
		Modeling Set		Validation Set		Modeling Set		Validation Set		Modeling Set		Validation Set		Modeling Set		ValidationSet		Modeling Set		Validation Set	
		R^2	RMSE (kg/hm^2)	R^2	RMSE (kg/hm^2)	R^2	RMSE (kg/hm^2)	R^2	RMSE (kg/hm^2)	R^2	RMSE (kg/hm^2)	R^2	RMSE (kg/hm^2)	R^2	RMSE (kg/hm^2)	R^2	RMSE (kg/hm^2)	R^2	RMSE (kg/hm^2)	R^2	RMSE (kg/hm^2)
PLSR	Boruta	0.54	427.85	0.61	464.34	0.68	801.67	0.57	1072.49	0.72	1096.66	0.70	1351.62	0.72	1125.80	0.67	1405.25	0.49	2017.00	0.43	2515.77
	RFE	0.53	434.66	0.57	493.81	0.78	635.91	0.47	1226.82	0.70	1148.63	0.70	1336.75	0.80	908.01	0.67	1398.94	0.63	1736.85	0.43	2407.61
	FB	0.72	324.48	0.59	473.29	0.77	644.07	0.46	1245.08	0.81	912.34	0.64	1474.44	0.79	948.73	0.60	1547.85	0.63	1744.52	0.43	2406.05
PCR	Boruta	0.54	427.33	0.60	469.15	0.64	853.54	0.60	1057.88	0.72	1101.43	0.72	1290.24	0.70	1171.49	0.71	1307.01	0.45	2120.01	0.45	2392.65
	RFE	0.53	435.68	0.57	493.77	0.43	1067.73	0.41	1311.00	0.70	1150.33	0.70	1338.99	0.70	1180.64	0.70	1321.14	0.43	2178.91	0.41	2412.86
	FB	0.55	423.94	0.61	477.47	0.41	1087.71	0.39	1332.82	0.68	1186.97	0.65	1465.01	0.59	1377.74	0.63	1541.66	0.41	2205.87	0.40	2430.66
SVM	Boruta	0.67	368.29	0.66	434.33	0.80	663.56	0.65	985.55	0.86	800.08	0.71	1349.87	0.82	902.99	0.66	1411.43	0.85	1135.73	0.52	2189.95
	RFE	0.67	367.30	0.65	445.05	0.97	329.32	0.35	1370.60	0.75	1073.13	0.74	1311.04	0.99	216.76	0.70	1430.80	0.96	716.15	0.22	2964.74
	FB	0.80	481.27	0.27	701.52	0.98	276.86	0.32	1412.37	0.97	486.02	0.19	2316.48	0.98	386.45	0.30	2266.78	0.95	852.86	0.19	2980.00
RR	Boruta	0.58	421.84	0.60	504.74	0.72	764.35	0.62	1009.78	0.76	1036.43	0.71	1301.00	0.81	955.61	0.68	1370.52	0.63	1827.13	0.53	2217.62
	RFE	0.53	435.50	0.55	513.04	0.90	513.16	0.53	1120.75	0.70	1144.30	0.72	1288.23	0.94	579.73	0.70	1318.70	0.91	1102.65	0.32	2628.54
	FB	0.95	192.63	0.46	514.86	0.90	518.87	0.50	1160.32	0.94	627.07	0.60	1554.82	0.95	612.77	0.57	1588.79	0.92	1095.07	0.29	2648.03

PLSR, partial least squares regression; PCR, principal component regression; SVM, support vector machine; RR, ridge regression.

10.3389/fpls.2024.1396183

Compared with the RFE algorithm, the Boruta algorithm identified fewer feature bands (Table 1); however, the resulting estimation model was more stable. This stability may be due to the characteristic bands determined by the Boruta algorithm being more effective and independent (Poona et al., 2016). Additionally, when comparing feature bands screened by both the Boruta and RFE algorithms, it was noted that their intersection predominantly occurred in the "red edge" region. This could be due to the band encapsulates rich crop information and is highly correlated with AGB (Horler et al., 1983; Filella and Penuelas, 1994), consistent with the findings of Cheng et al. (2017), who assessed the estimation effects of eight vegetation indices on rice canopy composition AGB and found the red edge index to be particularly sensitive to leaf AGB, achieving the highest estimation accuracy.

Another aspect of interest is that the characteristic bands identified by the Boruta algorithm, which are relatively evenly distributed across the spectrum, are predominantly found at wavelengths of 420 nm, 644 nm, 750 nm, and 842 nm. These characteristic bands are mainly located in areas sensitive to crop AGB response and are relatively evenly distributed. During the tillering stage, the characteristic bands are primarily located in the "red edge" region, which may be due to the low vegetation coverage at this time and the spectrum response is not obvious (Kokaly and Clark, 1999; Clevers et al., 2001). In the jointing stage, characteristic wavelengths are observed in the near-infrared region, likely due to the increase in rice canopy biomass, making the wavelength response in this region more pronounced and easier to detect during the selection process (Datt, 1998). From the heading to the maturity stage, a few characteristic wavelengths in the blue light region are identified, possibly because the chlorophyll content in the leaves gradually decreases, leading to increased spectral reflectance in the blue light region (Yan-lin et al., 2004). Thus, the Boruta algorithm effectively mines the spectral information of the rice canopy, and the selected characteristic bands align with the spectral response changes in the rice canopy, thereby enhancing the accuracy of AGB estimation.

4.3 Rice AGB estimation based on different machine learning algorithms

The performances of various models were meticulously compared, with the PCR model emerging as the most reliable and stable for estimating AGB. The differences between the R^2 values for the modeling set and the validation set are minimal, ranging from 0 to 0.1. Both the RMSE and MAE are lower for the PCR model than for the other models. This superior performance of the PCR model can be attributed to its foundation in predictive data mining technology through principal component analysis (PCA), which efficiently reduces the dimensionality of independent variables, mitigates the effects of multicollinearity among variables, and enhances the model's adaptability (Suryakala and Prince, 2019). In the PLSR model. Models constructed with three different input variables all demonstrated the ability to estimate rice AGB, particularly at the booting stage, when the R^2 of the validation set exceeded 0.6. This performance is likely due to the robust adaptability of the PLSR algorithm, its ability to diminish the dimensionality of spectral data, its ability to effectively handle complex collinear relationships between independent variables, and its overall enhancement of the model's estimation accuracy (Helland, 2001). In the SVM model, estimation models built on FD-RFE and FD-FB inputs exhibited overfitting at the jointing, heading, and maturity stages. Due to the excessive number of FD-RFE and FD-FB, when fitting as sample data, the noise interference in the model is large, resulting in an increase in estimation error. Even some noise may be recognized as characteristic frequency bands by machine learning algorithms, destroying the preset classification rules and significantly affecting the accuracy of model estimation (Cherkassky, 1997; Shafaei and Kisi, 2017; Raja et al., 2022). FD-Boruta has a small number but high estimation potential and can solve collinearity problems well when combined with the SVM algorithm. Similar outcomes were also observed in the RR model. This may be because the RR algorithm cannot automatically select important feature variables when fitting sample data. Therefore, when the amount of input sample data is large, it is easily affected by outliers, thereby reducing the model's predictive ability. It may also be that because determining optimal ridge parameters is critical when building an estimation model, poor parameter selection can lead to overfitting of the model, thereby reducing its interpretability (Smith and Campbell, 1980; Forrester and Kalivas, 2004). Moreover, this study demonstrates that the model achieves the highest estimation accuracy during the booting stage. This can be ascribed to the booting stage, which is the primary period for increases in rice chlorophyll and is characterized by high vegetation coverage and the absence of panicles. At this juncture, noise interference is minimized, and the spectral information collected is purer and more precise, providing advantageous conditions for estimating rice AGB (Chang et al., 2005; Takai et al., 2006; Kawamura et al., 2018).

5 Conclusion

This study implements SG smoothing and derivative transform for the preprocessing of spectral data, aiming to diminish fluctuations and noise interference. Subsequently, the Boruta and RFE algorithms are applied to select features from the first-order spectrum, thereby reducing the data dimensions and information redundancy and enhancing the data accuracy. A rice AGB estimation model was developed from the tillering stage to the maturity stage by integrating PLSR, PCR, SVM, and ridge regression machine learning algorithms. The key findings of the research are as follows:

(1) Except during the booting stage, FD exhibited the strongest correlation with AGB across the other four growth stages, followed by OS and then SD. As the rice growth stage progresses, the correlation shows a trend of first increasing and then decreasing.

(2) The performance of the Boruta algorithm is more stable, the selected characteristic bands are more sensitive, and effective dimensionality reduction of spectrum data is

achieved. The number of feature strips provided by the RFE algorithm ranges from 3 to 439, and the number of feature strips selected by the Boruta algorithm is within 40.

(3) Among all model combinations, the characteristic band selected by the Boruta algorithm performs best, and FD-Boruta-PCR emerged as the superior model for estimating rice AGB, with R^2 values between 0.45 and 0.72 and RMSE values between 469.15 and 2392.65 kg/hm^2. The worst performing model is the FD-FB-SVM model, with R^2 values between 0.19 and 0.32 and RMSE values between 701.52 and 2980.00 kg/hm^2.

(4) As the progression from the tillering stage to the mature stage occurs, the accuracy of the AGB estimation model decreases. Among them, the booting stage was determined to be the most accurate prediction stage, as evidenced by an R^2 of 0.72 and an RMSE of 1290.24 kg/hm^2.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

YN: Conceptualization, Writing – original draft, Writing – review & editing, Data curation, Formal Analysis. XS: Writing – review & editing. HY: Investigation, Writing – review & editing. YZ: Methodology, Supervision, Writing – review & editing. JL: Formal Analysis, Writing – review & editing. WW: Project administration, Writing – review & editing. YS: Validation, Writing – review & editing. QM: Visualization, Writing – review & editing. JKL: Resources, Writing – review & editing. XL: Funding acquisition, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by Scientific research projects in higher education institutions of Anhui Province (no. 2023AH051855; 2022AH051623); Anhui Province Crop Intelligent Planting and Processing Technology Engineering Research Center Open Research Project (no. ZHKF202303); Natural Science Foundation of Hebei Province (no. C2023408010), and College Students ' Innovation and Entrepreneur ship Training Project (no. 202210879043; 202310879114; 202310879115).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1396183/full#supplementary-material

References

Abdi, H., and Williams, L. J. (2010). Principal component analysis. *WIREs Comput. Stat* 2, 433–459. doi: 10.1002/wics.101

Atzberger, C., Darvishzadeh, R., Immitzer, M., Schlerf, M., Skidmore, A., and le Maire, G. (2015). Comparative analysis of different retrieval methods for mapping grassland leaf area index using airborne imaging spectroscopy. *Int. J. Appl. Earth Observation Geoinformation* 43, 19–31. doi: 10.1016/j.jag.2015.01.009

Bansod, B., Singh, R., Thakur, R., and Singhal, G. (2017). A comparision between satellite based and drone based remote sensing technology to achieve sustainable development: a review. *J. Agric. Environ. Int. Dev. (JAEID)* 111, 383–407. doi: 10.12895/jaeid.20172.690

Becker, B. L., Lusch, D. P., and Qi, J. (2005). Identifying optimal spectral bands from in *situ* measurements of Great Lakes coastal wetlands using second-derivative analysis. *Remote Sens. Environ.* 97, 238–248. doi: 10.1016/j.rse.2005.04.020

Cao, C., Wang, T., Gao, M., Li, Y., Li, D., and Zhang, H. (2021a). Hyperspectral inversion of nitrogen content in maize leaves based on different dimensionality reduction algorithms. *Comput. Electron. Agric.* 190, 106461. doi: 10.1016/j.compag.2021.106461

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., et al. (2021b). Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches. *Agric. For. Meteorol.* 297, 108275. doi: 10.1016/j.agrformet.2020.108275

Chang, K.-W., Shen, Y., and Lo, J.-C. (2005). Predicting rice yield using canopy reflectance measured at booting stage. *Agron. J.* 97, 872–878. doi: 10.2134/agronj2004.0162

Chen, Q., Meng, Z., Liu, X., Jin, Q., and Su, R. (2018). Decision variants for the automatic determination of optimal feature subset in RF-RFE. *Genes* 9, 301. doi: 10.3390/genes9060301

Chen, S., Hu, T., Luo, L., He, Q., Zhang, S., Li, M., et al. (2020). Rapid estimation of leaf nitrogen content in apple-trees based on canopy hyperspectral reflectance using multivariate methods. *Infrared Phys. Technol.* 111, 103542. doi: 10.1016/j.infrared.2020.103542

Cheng, T., Song, R., Li, D., Zhou, K., Zheng, H., Yao, X., et al. (2017). Spectroscopic estimation of biomass in canopy components of paddy rice using dry matter and chlorophyll indices. *Remote Sens.* 9, 319. doi: 10.3390/rs9040319

Cherkassky, V. (1997). The nature of statistical learning theory~. *IEEE Trans. Neural Netw.* 8, 1564–1564. doi: 10.1109/TNN.72

Clevers, J. G. P. W., de Jong, S. M., Epema, G. F., van der Meer, F., Bakker, W. H., Skidmore, A. K., et al. (2001). MERIS and the red-edge position. *Int. J. Appl. Earth Observation Geoinformation* 3, 313–320. doi: 10.1016/S0303-2434(01)85038-8

Datt, B. (1998). Remote Sensing of Chlorophyll a, Chlorophyll b, Chlorophyll a+b, and Total Carotenoid Content in Eucalyptus Leaves. *Remote Sens. Environ.* 66, 111–121. doi: 10.1016/S0034-4257(98)00046-7

Degenhardt, F., Seifert, S., and Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings Bioinf.* 20, 492–503. doi: 10.1093/bib/bbx124

Demetriades-Shah, T. H., Steven, M. D., and Clark, J. A. (1990). High resolution derivative spectra in remote sensing. *Remote Sens. Environ.* 33, 55–64. doi: 10.1016/0034-4257(90)90055-Q

Devia, C. A., Rojas, J. P., Petro, E., Martinez, C., Mondragon, I. F., Patino, D., et al. (2019). High-throughput biomass estimation in rice crops using UAV multispectral imagery. *J. Intell. Robot Syst.* 96, 573–589. doi: 10.1007/s10846-019-01001-5

Fang, Q., Wang, G., Liu, T., Xue, B.-L., and Yinglan, A. (2018). Controls of carbon flux in a semi-arid grassland ecosystem experiencing wetland loss: Vegetation patterns and environmental variables. *Agric. For. Meteorol.* 259, 196–210. doi: 10.1016/j.agrformet.2018.05.002

Filella, I., and Penuelas, J. (1994). The red edge position and shape as indicators of plant chlorophyll content, biomass and hydric status. *Int. J. Remote Sens.* 15, 1459–1470. doi: 10.1080/01431169408954177

Forrester, J. B., and Kalivas, J. H. (2004). Ridge regression optimization using a harmonious approach. *J. Chemometrics* 18, 372–384. doi: 10.1002/cem.883

Fu, Y., Yang, G., Li, Z., Li, H., Li, Z., Xu, X., et al. (2020). Progress of hyperspectral data processing and modelling for cereal crop nitrogen monitoring. *Comput. Electron. Agric.* 172, 105321. doi: 10.1016/j.compag.2020.105321

Galvão, L. S., Formaggio, A. R., and Tisot, D. A. (2005). Discrimination of sugarcane varieties in Southeastern Brazil with EO-1 Hyperion data. *Remote Sens. Environ.* 94, 523–534. doi: 10.1016/j.rse.2004.11.012

Gao, J., Meng, B., Liang, T., Feng, Q., Ge, J., Yin, J., et al. (2019). Modeling alpine grassland forage phosphorus based on hyperspectral remote sensing and a multi-factor machine learning algorithm in the east of Tibetan Plateau, China. *ISPRS J. Photogrammetry Remote Sens.* 147, 104–117. doi: 10.1016/j.isprsjprs.2018.11.015

Geladi, P., and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta* 185, 1–17. doi: 10.1016/0003-2670(86)80028-9

Gerretzen, J., Szymańska, E., Jansen, J. J., Bart, J., Van Manen, H.-J., Van Den Heuvel, E. R., et al. (2015). Simple and effective way for data preprocessing selection based on design of experiments. *Anal. Chem.* 87, 12096–12103. doi: 10.1021/acs.analchem.5b02832

Gitelson, A. A., Merzlyak, M. N., and Lichtenthaler, H. K. (1996). Detection of red edge position and chlorophyll content by reflectance measurements near 700 nm. *J. Plant Physiol.* 148, 501–508. doi: 10.1016/S0176-1617(96)80285-9

Gnyp, M. L., Miao, Y., Yuan, F., Ustin, S. L., Yu, K., Yao, Y., et al. (2014). Hyperspectral canopy sensing of paddy rice aboveground biomass at different growth stages. *Field Crops Res.* 155, 42–55. doi: 10.1016/j.fcr.2013.09.023

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–422. doi: 10.1023/A:1012487302797

Hansen, P. M., and Schjoerring, J. K. (2003). Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Remote Sens. Environ.* 86, 542–553. doi: 10.1016/S0034-4257(03)00131-7

He, J., Zhang, N., Su, X., Lu, J., Yao, X., Cheng, T., et al. (2019). Estimating leaf area index with a new vegetation index considering the influence of rice panicles. *Remote Sens.* 11, 1809. doi: 10.3390/rs11151809

Helland, I. S. (2001). Some theoretical aspects of partial least squares regression. *Chemometrics Intelligent Lab. Syst.* 58, 97–107. doi: 10.1016/S0169-7439(01)00154-X

Horler, D. N. H., Dockray, M., and Barber, J. (1983). The red edge of plant leaf reflectance. *Int. J. Remote Sens.* 4, 273–288. doi: 10.1080/01431168308948546

Ivanda, A., Šerić, L., Bugarić, M., and Braović, M. (2021). Mapping chlorophyll-a concentrations in the kaštela bay and brač Channel using ridge regression and sentinel-2 satellite images. *Electronics* 10, 3004. doi: 10.3390/electronics10233004

Ji, S., Gu, C., Xi, X., Zhang, Z., Hong, Q., Huo, Z., et al. (2022). Quantitative monitoring of leaf area index in rice based on hyperspectral feature bands and ridge regression algorithm. *Remote Sens.* 14, 2777. doi: 10.3390/rs14122777

Jia, Y., Zhang, H., Zhang, X., and Su, Z. (2022). Quantitative analysis and hyperspectral remote sensing inversion of rice canopy spad in A cold region. *Eng. Agric.* 42, e20220030. doi: 10.1590/1809-4430-eng.agric.v42n4e20220030/2022

Kanke, Y., Tubaña, B., Dalen, M., and Harrell, D. (2016). Evaluation of red and red-edge reflectance-based vegetation indices for rice biomass and grain yield prediction models in paddy fields. *Precis. Agric.* 17, 507–530. doi: 10.1007/s11119-016-9433-1

Kawamura, K., Ikeura, H., Phongchanmaixay, S., and Khanthavong, P. (2018). Canopy Hyperspectral Sensing of Paddy Fields at the Booting Stage and PLS Regression can Assess Grain Yield. *Remote Sens.* 10, 1249. doi: 10.3390/rs10081249

Kawano, S., Fujisawa, H., Takada, T., and Shiroishi, T. (2018). Sparse principal component regression for generalized linear models. *Comput. Stat Data Anal.* 124, 180–196. doi: 10.1016/j.csda.2018.03.008

Kokaly, R. F., and Clark, R. N. (1999). Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression. *Remote Sens. Environ.* 67, 267–287. doi: 10.1016/S0034-4257(98)00084-4

Krishnan, P., Ramakrishnan, B., Reddy, K. R., and Reddy, V. R. (2011). Chapter three - high-temperature effects on rice growth, yield, and grain quality. *Adv. Agron.* 111, 87–206. doi: 10.1016/B978-0-12-387689-8.00004-7

Kursa, M. B., Jankowski, A., and Rudnicki, W. R. (2010). Boruta – A system for feature selection. *Fundamenta Informaticae* 101, 271–285. doi: 10.3233/FI-2010-288

Kursa, M. B., and Rudnicki, W. R. (2010). Feature selection with the boruta package. *J. Stat. Software* 36, 1–13. doi: 10.18637/jss.v036.i11

Li, Z., Chen, Z., Cheng, Q., Duan, F., Sui, R., Huang, X., et al. (2022). UAV-based hyperspectral and ensemble machine learning for predicting yield in winter wheat. *Agronomy* 12, 202. doi: 10.3390/agronomy12010202

Li, B., and Xie, W. (2015). Adaptive fractional differential approach and its application to medical image enhancement. *Comput. Electrical Eng.* 45, 324–335. doi: 10.1016/j.compeleceng.2015.02.013

Liang, L., Di, L., Huang, T., Wang, J., Lin, L., Wang, L., et al. (2018). Estimation of leaf nitrogen content in wheat using new hyperspectral indices and a random forest regression algorithm. *Remote Sens.* 10, 1940. doi: 10.3390/rs10121940

Lin, X., Ren, C., Li, Y., Yue, W., Liang, J., and Yin, A. (2023). Eucalyptus plantation area extraction based on SLPSO-RFE feature selection and multi-temporal sentinel-1/2 data. *Forests* 14, 1864. doi: 10.3390/f14091864

Liu, G.-S., Guo, H.-S., Pan, T., Wang, J.-H., and Cao, G. (2014). Vis-NIR spectroscopic pattern recognition combined with SG smoothing applied to breed screening of transgenic sugarcane. *Guang Pu Xue Yu Guang Pu Fen Xi* 34, 2701–2706.

Lucas, R., Van De Kerchove, R., Otero, V., Lagomasino, D., Fatoyinbo, L., Omar, H., et al. (2020). Structural characterisation of mangrove forests achieved through combining multiple sources of remote sensing data. *Remote Sens. Environ.* 237, 111543. doi: 10.1016/j.rse.2019.111543

Mishra, P., Biancolillo, A., Roger, J. M., Marini, F., and Rutledge, D. N. (2020). New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC Trends Analytical Chem.* 132, 116045. doi: 10.1016/j.trac.2020.116045

Moore, B. (1981). Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Trans. Automatic Control* 26, 17–32. doi: 10.1109/TAC.1981.1102568

Pan, L., Li, H.-C., Deng, Y.-J., Zhang, F., Chen, X.-D., and Du, Q. (2017). Hyperspectral dimensionality reduction by tensor sparse and low-rank graph-based discriminant analysis. *Remote Sens.* 9, 452. doi: 10.3390/rs9050452

Paul, J., D'Ambrosio, R., and Dupont, P. (2015). Kernel methods for heterogeneous feature selection. *Neurocomputing* 169, 187–195. doi: 10.1016/j.neucom.2014.12.098

Poona, N. K., Niekerk, A., Nadel, R. L., and Ismail, R. (2016). Random forest (RF) wrappers for waveband selection and classification of hyperspectral data. *Appl. Spectrosc. AS* 70, 322–333. doi: 10.1177/0003702815620545

Prasad, R., Deo, R. C., Li, Y., and Maraseni, T. (2019). Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach. *CATENA* 177, 149–166. doi: 10.1016/j.catena.2019.02.012

Raja, S. P., Sawicka, B., Stamenkovic, Z., and Mariammal, G. (2022). Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers. *IEEE Access* 10, 23625–23641. doi: 10.1109/ACCESS.2022.3154350

Ren, H., Zhou, G., and Zhang, F. (2018). Using negative soil adjustment factor in soil-adjusted vegetation index (SAVI) for aboveground living biomass estimation in arid grasslands. *Remote Sens. Environ.* 209, 439–445. doi: 10.1016/j.rse.2018.02.068

Savitzky, A., and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36, 1627–1639. doi: 10.1021/ac60214a047

Schafer, R. W. (2011). What is a savitzky-golay filter? [Lecture notes]. *IEEE Signal Process. Magazine* 28, 111–117. doi: 10.1109/MSP.2011.941097

Schino, G., Borfecchia, F., De Cecco, L., Dibari, C., Iannetta, M., Martini, S., et al. (2003). Satellite estimate of grass biomass in a mountainous range in central Italy. *Agroforestry Syst.* 59, 157–162. doi: 10.1023/A:1026308928874

Shafaei, M., and Kisi, O. (2017). Predicting river daily flow using wavelet-artificial neural networks based on regression analyses in comparison with artificial neural networks and support vector machine models. *Neural Comput. Applic* 28, 15–28. doi: 10.1007/s00521-016-2293-9

Shen, L., Gao, M., Yan, J., Wang, Q., and Shen, H. (2022). Winter wheat SPAD value inversion based on multiple pretreatment methods. *Remote Sens.* 14, 4660. doi: 10.3390/rs14184660

Shi, X., Yao, L., and Pan, T. (2021). Visible and near-infrared spectroscopy with multi-parameters optimization of savitzky-golay smoothing applied to rapid analysis of soil cr content of pearl river delta. *J. Geosci. Environ. Prot.* 09, 75. doi: 10.4236/gep.2021.93006

Smith, G., and Campbell, F. (1980). A critique of some ridge regression methods. *J. Am. Stat. Assoc.* 75, 74–81. doi: 10.1080/01621459.1980.10477428

Spiertz, J. H. J., and Ewert, F. (2009). Crop production and resource use to meet the growing demand for food, feed and fuel: opportunities and constraints. *NJAS: Wageningen J. Life Sci.* 56, 281–300. doi: 10.1016/S1573-5214(09)80001-8

Sun, H., Feng, M., Xiao, L., Yang, W., Ding, G., Wang, C., et al. (2021). Potential of multivariate statistical technique based on the effective spectra bands to estimate the

plant water content of wheat under different irrigation regimes. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.631573

Suryakala, S. V., and Prince, S. (2019). Investigation of goodness of model data fit using PLSR and PCR regression models to determine informative wavelength band in NIR region for non-invasive blood glucose prediction. *Opt Quant Electron* 51, 271. doi: 10.1007/s11082-019-1985-7

Suykens, J. A. K., and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Process. Lett.* 9, 293–300. doi: 10.1023/A:1018628609742

Takai, T., Matsuura, S., Nishio, T., Ohsumi, A., Shiraiwa, T., and Horie, T. (2006). Rice yield potential is closely related to crop growth rate during late reproductive period. *Field Crops Res.* 96, 328–335. doi: 10.1016/j.fcr.2005.08.001

Tong, X., Duan, L., Liu, T., Yang, Z., Wang, Y., and Singh, V. P. (2023). Estimation of grassland aboveground biomass combining optimal derivative and raw reflectance vegetation indices at peak productive growth stage. *Geocarto Int.* 38, 2186497. doi: 10.1080/10106049.2023.2186497

Tunca, E., Köksal, E. S., Öztürk, E., Akay, H., and Çetin Taner, S. (2023). Accurate estimation of sorghum crop water content under different water stress levels using machine learning and hyperspectral data. *Environ. Monit Assess.* 195, 877. doi: 10.1007/s10661-023-11536-8

Wang, X., Huang, J., Li, Y., and Wang, R. (2004). Study on hyperspectral remote sensing estimation models about aboveground fresh biomass of rice. *Remote Sens. Modeling Ecosyst. Sustainability (SPIE)*, 336–345. doi: 10.1117/12.557403

Wang, L., Zhou, X., Zhu, X., Dong, Z., and Guo, W. (2016). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *Crop J.* 4, 212–219. doi: 10.1016/j.cj.2016.01.008

Xu, T., Wang, F., Shi, Z., Xie, L., and Yao, X. (2023). Dynamic estimation of rice aboveground biomass based on spectral and spatial information extracted from hyperspectral remote sensing images at different combinations of growth stages. *ISPRS J. Photogrammetry Remote Sens.* 202, 169–183. doi: 10.1016/j.isprsjprs.2023.05.021

Xu, T., Wang, F., Xie, L., Yao, X., Zheng, J., Li, J., et al. (2022). Integrating the textural and spectral information of UAV hyperspectral images for the improved estimation of rice aboveground biomass. *Remote Sens.* 14, 2534. doi: 10.3390/rs14112534

Yang, C.-M., and Chen, R.-K. (2004). Modeling rice growth with hyperspectral reflectance data. *Crop Sci.* 44, 1283–1290. doi: 10.2135/cropsci2004.1283

Yang, S., Feng, Q., Liang, T., Liu, B., Zhang, W., and Xie, H. (2018). Modeling grassland above-ground biomass based on artificial neural network and remote sensing in the Three-River Headwaters Region. *Remote Sens. Environ.* 204, 448–455. doi: 10.1016/j.rse.2017.10.011

Yang, X., Huang, J., Wu, Y., Wang, J., Wang, P., Wang, X., et al. (2011). Estimating biophysical parameters of rice with remote sensing data using support vector machines. *Sci. China Life Sci.* 54, 272–281. doi: 10.1007/s11427-011-4135-4

Yan-lin, T., Jing-feng, H., and Ren-chao, W. (2004). Change law of hyperspectral data in related with chlorophyll and carotenoid in rice at different developmental stages. *Rice Sci.* 11, 274–282.

Yoosefzadeh-Najafabadi, M., Earl, H. J., Tulpan, D., Sulik, J., and Eskandari, M. (2021). Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.624273

Yu, F., Bai, J., Jin, Z., Zhang, H., and Xu, T. (2023). Inversion of rice leaf biomass based on PROSAIL model optimization. *J. Huazhong Agric. Univ. (Natural Sci. Edition) (English Edition)* 42, 187–194. doi: 10.13300/j.cnki.hnlkxb.2023.03.022

Yu, F., Feng, S., Du, W., Wang, D., Guo, Z., Xing, S., et al. (2020). A study of nitrogen deficiency inversion in rice leaves based on the hyperspectral reflectance differential. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.573272

Yue, J., Yang, G., Tian, Q., Feng, H., Xu, K., and Zhou, C. (2019). Estimate of winter-wheat above-ground biomass based on UAV ultrahigh-ground-resolution image textures and vegetation indices. *ISPRS J. Photogrammetry Remote Sens.* 150, 226–244. doi: 10.1016/j.isprsjprs.2019.02.022

Zhang, K., Ge, X., Shen, P., Li, W., Liu, X., Cao, Q., et al. (2019). Predicting rice grain yield based on dynamic changes in vegetation indexes during early to mid-growth stages. *Remote Sens.* 11, 387. doi: 10.3390/rs11040387

Zhao, M., Gao, Y., Lu, Y., and Wang, S. (2022). Hyperspectral modeling of soil organic matter based on characteristic wavelength in east China. *Sustainability* 14, 8455. doi: 10.3390/su14148455