Early estimation of faba bean yield based on unmanned aerial systems hyperspectral images and stacking ensemble

Yuxing Cui
Chinese Academy of Agricultural Sciences
Yishan Ji
Chinese Academy of Agricultural Sciences
Shuaipeng Fei
China Agricultural University
Zehao Liu
Chinese Academy of Agricultural Sciences
Rong Liu
Chinese Academy of Agricultural Sciences
Xuxiao Zong
Chinese Academy of Agricultural Sciences
Tao Yang (yangtao02@caas.cn)
Chinese Academy of Agricultural Sciences

Research Article

Keywords: remote sensing, hyperspectral images, stacking ensemble, UAS

Posted Date: September 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3327958/v1

Additional Declarations: No competing interests reported.

Early estimation of faba bean yield based on unmanned aerial systems hyperspectral images and stacking ensemble

Yuxing Cui [1]†, Yishan Ji [1]†, Shuaipeng Fei [2], Zehao Liu [1], Rong Liu [1], Xuxiao Zong [1*], Tao Yang [1*]

[1] National Key Facility for Crop Gene Resources and Genetic Improvement/Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[2] College of Land Science and Technology, China Agricultural University, Beijing 100193, China

*Address correspondence to: yangtao02@caas.cn

†These authors contributed equally to this work

Abstract

Faba bean is a vital legume crop, and its early yield estimation can improve field management practices. In this study, unmanned aerial system (UAS) hyperspectral imagery was used for the first time to estimate faba bean yield early. Different basic algorithms, including random forest (RF), support vector machine (SVM), k-nearest neighbor (KNN), partial least squares regression (PLS), and eXtreme Gradient Boosting (XGB), were employed along with stacking ensemble learning to construct the faba bean yield model and investigate factors influencing model accuracy. The results are as follows: when using the same algorithm and growth period, integrating texture information into the model improved the estimation accuracy compared to using spectral information alone. Among the base models, the XGB model performed the best in the context of growth period consistency. Moreover, the stacking ensemble significantly improved model accuracy, yielding satisfactory results, with the highest model accuracy (R^2) reaching 0.76. Model accuracy varied significantly for models based on different growth periods using the same algorithm. The accuracy of the model gradually improved during a single growth period, but the rate of improvement decreased over time. Data fusion of growth period data helped enhance model accuracy in most cases. In conclusion, combining UAS-based hyperspectral data with ensemble learning for early yield estimation of faba beans is feasible, therefore, this study would offer a novel approach to predict faba bean yield.

Keywords: remote sensing, hyperspectral images, stacking ensemble, UAS

1 Introduction

Faba bean is a highly nutritious and economically valuable crop of exceptional quality that is extensively cultivated in Eurasia. Yield serves as a crucial indicator of plant phenomics and represents the ultimate goal of breeding. Accurate, efficient, and timely yield estimation can significantly enhance the sustainability of the industry and provide growers with essential information regarding fertilizer application, irrigation, pesticide use, and more [1-3]. Currently, two main traditional methods are employed for yield assessment: one relies on accumulated years of experience combined with a general forecast of the current year's weather, whereas the other involves yield estimation based on field data surveys requiring substantial human and material resources [4]. The former demands an exceedingly high level of expertise from researchers, and factors such as weather and soil environment variations also lead to fluctuations in assessment accuracy. The latter is time-consuming, labor-intensive, and often results in irreversible crop damage due to human and mechanical interventions, making it impractical for large-scale crops [3].

In recent times, spectroscopy has emerged as a pivotal technology, focusing on the generation of spectra from various substances and their interactions. Researchers in agriculture have adapted this technology to create a new discipline known as agricultural spectroscopy, quantifying phenotypic characteristics through the interaction between spectra and plants [5-6]. With the advent of remote sensing technology, especially through unmanned aerial systems equipped with spectral imaging capabilities, an increasing number of researchers are investigating the collection of phenotypic characteristics of crops. These studies encompass crop phenotype estimation through remote sensing models [7], water and fertilizer management [8-9], and environmental stress [10-11], all yielding promising results. For instance, Li et al. [7] fused light detection and ranging (LiDAR) image data and multispectral data to construct a machine learning model that accurately predicted (R^2 = 0.891) wheat yields two months before harvest. In another study, hyperspectral imaging was utilized to establish a nitrogen content estimation model based on the partial least squares regression algorithm, achieving an estimation accuracy of 0.85 [12]. These findings demonstrate the potential of remote sensing-based models for applications in agricultural production. Sensors commonly used for spectral data acquisition are RGB [13], multispectral [14], and thermal infrared [15]. However, hyperspectral sensors offer a significant advantage, as they cover numerous continuous wavelength bands, providing detailed spectral information to quantify the biochemical and physical characteristics of crops [16-17]. Additionally, the application of second- or multiple-order derivatives can reveal hidden information related to crop growth [5]. Although researchers have demonstrated the utility of hyperspectral imaging in the study of crop phenotypes [18-19], limited research has been conducted on the phenotypic characteristics of faba bean. Moreover, some studies have integrated spectral features with texture features to construct models, achieving improved estimation performance [20]. Building upon these studies, we employed machine learning algorithms to create a yield estimation model for faba bean, utilizing canopy spectral features and texture features as inputs.

Commonly used machine learning algorithms include random forest (RF) [14,21], support vector machine (SVM) [22], k-nearest neighbor (KNN) [23], partial least squares regression (PLS) [24], and eXtreme Gradient Boosting (XGB) [25]. These algorithms have been successfully applied in studies on crop phenotypes, including yield [14], biomass [26], chlorophyll content [24], and leaf area index [27]. While these studies have demonstrated promising results, each individual model has some drawbacks: RF may overfit noisy classification or regression problems, SVM can be challenging to implement with large training samples, and KNN exhibits low estimation accuracy for rare categories when the sample sizes are unbalanced [28-29].

In light of selecting early faba bean canopy spectra and texture features as inputs, we employed stacking ensemble learning to construct crop estimation models. This approach addresses the limitations of standalone machine learning models and can enhance the accuracy of model estimation [29-31]. Ensemble learning has been shown to improve the generalization ability of models by combining the strengths of different models in previous studies, effectively mitigating the risk of overfitting through cross-validation during model construction [29,31-32]. However, to the best of our knowledge, the stacking ensemble approach has rarely been applied to faba bean yield estimation.

Based on the aforementioned considerations, this study has three main objectives: (1) to evaluate the potential of hyperspectral imagery-based and ensemble learning in faba bean yield estimation; (2) to explore the impact of continuous period data fusion on estimation model accuracy; and (3) to investigate the effect of spectral features and structural features on model estimation accuracy.

2 Materials and Methods

2.1 Experimental design

The experiment was conducted at the Xinxiang Research Site of the Chinese Academy of Agricultural Sciences (113°45' 40" E, 35°8' 11" N, Figure 1) from November 17, 2021, to June 10, 2022. The region has a warm temperate continental monsoon climate, with hot and rainy summers, cold and dry winters, and rain-heat extremes.

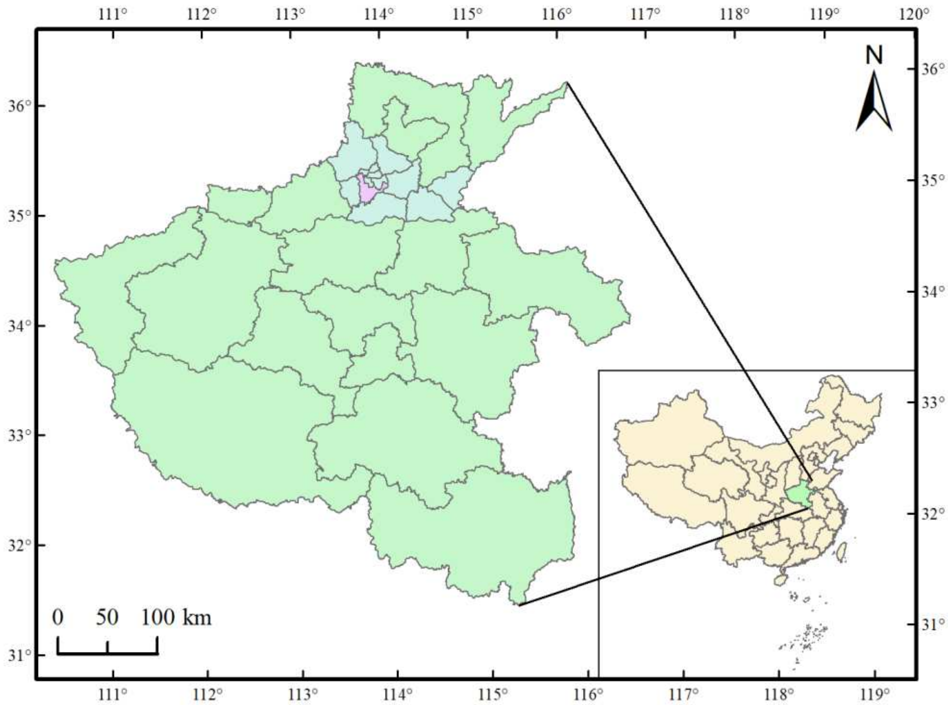

Figure 1. Specific location of field experiments

In this study, 48 faba bean varieties were selected for three replicated experimental groups. A total of 144 experimental plots were planted, following a completely randomized group design (Figure 1). Each plot measured 3.33 m in length and 3 m in width, with a total area of 10 m^2. Within each experimental plot, 10 rows were planted, each containing 25 faba bean seeds. Field management practices, including weeding and irrigation, were performed according to the routine practices of local farmers. Additionally, to enhance the accuracy of subsequent image mosaicking, a total of 16 ground control points (GCPs) were strategically positioned and evenly distributed across the experimental area (Figure 2c).

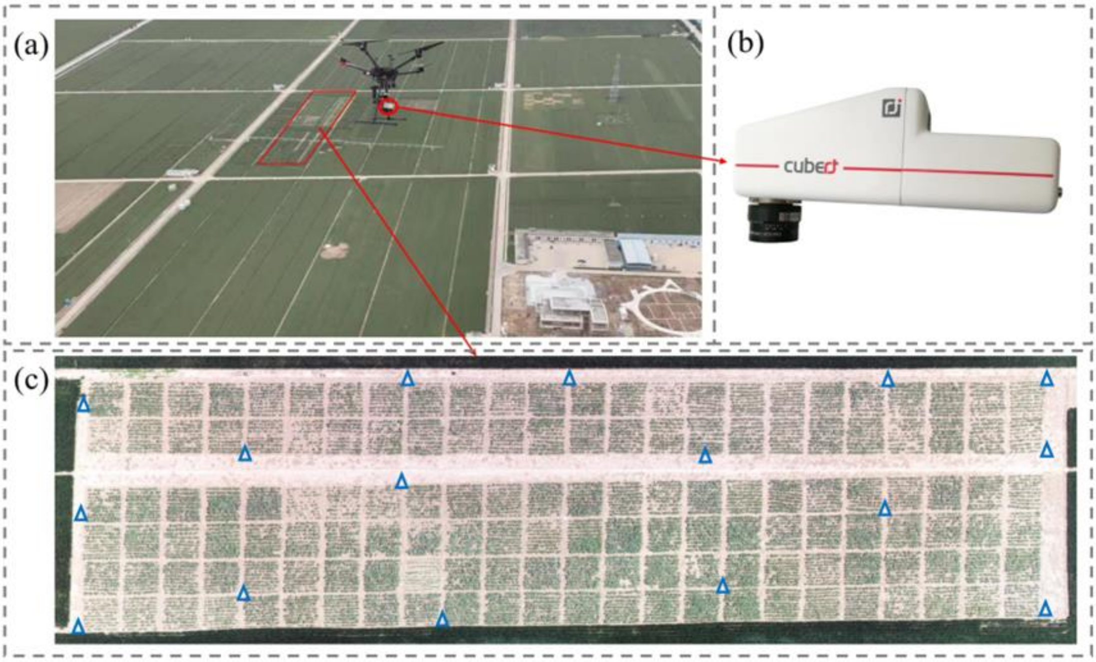

Figure 2. The experimental field, unmanned aerial systems, and hyperspectral camera.
Note: (a) the experimental field and UAS; (b) hyperspectral camera; (c) design of the experimental field; blue triangles indicate the location of ground control points.

2.2 UAS platform and flight mission

A Matrice 600 Pro Hexacopter (DJI, Shenzhen, China) equipped with a UHD185 hyperspectral imaging spectrometer (Cubert GmbH, Ulm, Germany; Figure 2b) was utilized to capture images in 125 spectral bands, spanning from the visible to the near-infrared range (450–950 nm). The UHD185 spectrometer offers a spectral resolution of 4 nm, with a 1000 × 1000 pixel configuration and a focal length of 23 mm [33]. Before the UAS-based hyperspectral imaging system operation, the spectrometer was calibrated using an equipped reference plate. For accurate data acquisition, clear and windless weather conditions were chosen during midday (11:00–14:00). Each flight was planned at an altitude of 30 m, with 85% forward and lateral overlap and minimal variation in sunlight intensity. The experimental workflow of this study is depicted in Figure 3.

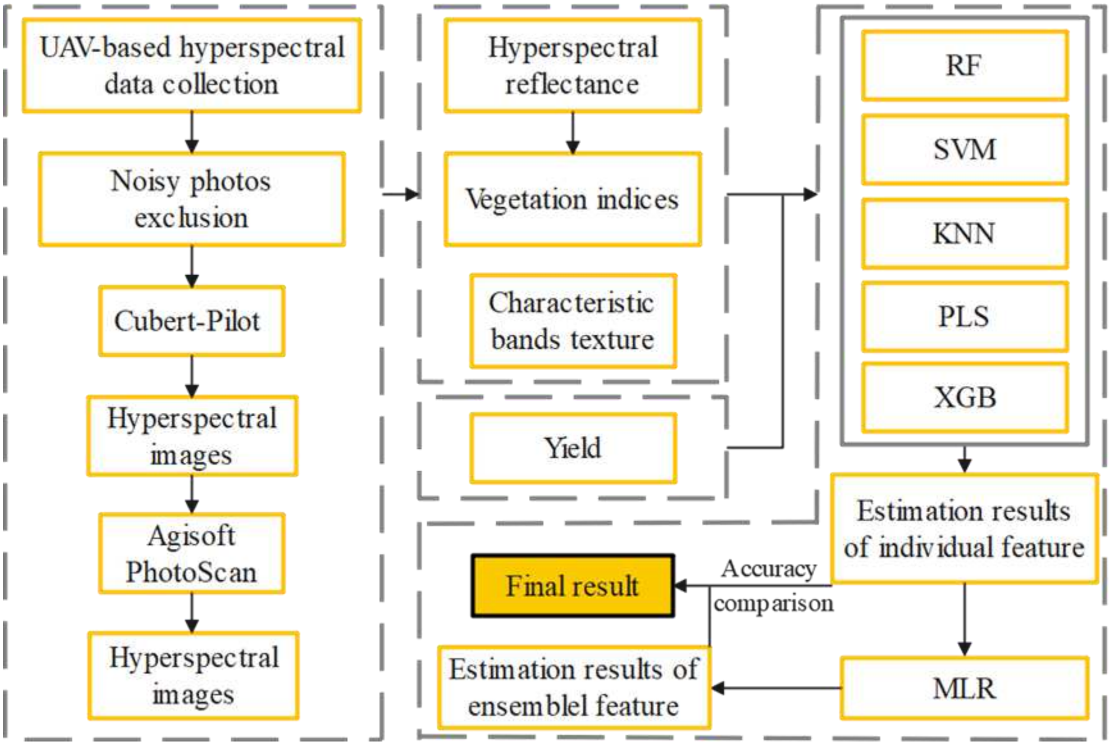

Figure 3. The workflow of this study

2.3 Data collection

The UAS-based hyperspectral data were collected on three dates: March 24, 2022 (P1), April 2, 2022 (P2), and April 11, 2022 (P3). Ground measurements consisted of cropped yield data, recorded as the weight of all seeds in the plot at the time of harvest on June 10, 2022, and are expressed as $t{\cdot}ha^{-1}$ (Table 1).

Table 1. Statistics of the measured yield

Date	Sample size	Max	Min	Mean	SD	CV (%)
10 June 2022	144	5.900	0.300	2.893	1.321	45.657

2.4 Image processing and data extraction

Raw image data were processed using the following steps: (1) Data preprocessing: Noisy photos, such as those taken during take-off, landing, and turning, were removed. Cubert-Pilot software was used to fuse each hyperspectral image with its corresponding panchromatic image. (2) Image mosaicking: Agisoft PhotoScan software (Agisoft LLC, St. Petersburg, Russia) was employed for adding and aligning photos; importing GCP coordinates; generating dense point clouds, grids, and textures; changing paths after stitching full-color images; converting to hyperspectral image stitching; and creating a digital orthophoto map.

For data extraction, ENVI 5.3 (Exelis Visual Information Solutions, Inc., Boulder, CO, USA) and ArcGIS 10.6 (Environmental Systems Research Institute, Inc., Redlands, CA, USA) were utilized. The soil background was eliminated from the stitched spectral images, considering the low vegetation canopy cover during the early stages of bean growth, which could significantly impact vegetation information. The hyperspectral images were split into single-band images by using the “Split to Multiple Single-Band Files” tool in ENVI. Subsequently, these single-band images were imported into the ArcGIS "Model Builder" to remove the soil background. The average spectral reflectance of all bands for all faba bean plants in each sample plot was then determined using ArcGIS.

Before texture extraction, characteristic bands were selected based on high information content, low correlation, and good separability. In this study, a band selection method based on the amount of information was employed. The optimal index factor (OIF) was used as the evaluation method, which is a multiband combination approach. The OIF formula was executed in PyCharm Community Edition 2022.3.3 (Python) to derive the optimal bands. The principle of the OIF is that the weaker the correlation between the bands and the larger the standard deviation of the bands, the more informative band combination, which implies that the amount of information in a band combination is inversely proportional to the correlation coefficient between the wavelengths and directly proportional to the standard deviation of the bands. The OIF was calculated [34] as follows:

$$OIF = \sum_{j=1}^{5}(S_i / \sum_{j=1}^{5} |R_{ij}|) \tag{1}$$

where Si is the standard deviation of the ith band and Rij is the correlation coefficient of the two bands.

After obtaining the optimal band combination, texture calculation and extraction were performed in ENVI by creating a texture image by using the “co-occurrence” tool, selecting the region of interest, calculating the texture, and exporting the data.

2.5 Vegetation indices

The spectral resolution of the UHD185 spectrometer was 4 nm, with a discontinuous band, resulting in missing data between bands. To address this problem, the linear difference method was employed to numerically estimate the unrecorded bands by interpolating data based on the left and right proximity of the points needing interpolation in a one-dimensional data sequence, which was calculated as follows [35]:

$$y = y_0 + \frac{(x - x_0)y_1 - (x - x_0)\, y_0}{x_1 - x_0} \tag{2}$$

where x_0 and x_1 are the wavelengths of the known proximity bands, and y_0 and y_1 are the spectral reflectance of the corresponding bands.

Subsequently, vegetation indices were calculated using the spectral data with reasonable compensation. A total of 53 vegetation indices (Additional file: Table S1) were constructed using the “hsdar” package [36] in R for the estimation of faba bean yield.

2.6 Modeling

2.6.1 Model construction

For model construction, this study adopted a stacked integration approach comprising primary and secondary models [37,38]. The primary model utilized five widely used machine learning methods as base learners to train on the dataset, and the results from this primary model were used as input to train the secondary models. To ensure model stability, the construction involved both 5-fold external cross-validation and 10-fold internal cross-validation. Each machine learning method underwent 10 iterations of 5-fold external cross-validation by using the “repeat” function. Specifically, the training dataset was randomly divided into five equal sets, of which four were used for training and one was used for validation; this process was repeated 10 times, with all data serving as training or validation samples. Concurrently, the hyperparameters of each machine learning algorithm were screened. This step was repeated 10 times for 10-fold internal cross-validation. The training dataset used in the external cross-validation was randomly divided into 10 folds, of which nine were used for training and one was used for validation in each data partition, and this process was repeated 10 times. For each machine learning algorithm, all hyperparameter combinations underwent full 10-fold internal cross-validation, and the combination with the best average accuracy was selected for training the external cross-validation dataset. Five machine learning models were tested for accuracy on the initial test set, and an out-of-sample estimation matrix was constructed for the validation set. After completing the 5-fold cross-validation, the mean of the five sets of estimations generated by each model was considered the new estimation matrix and further used as a test set for the secondary models. Machine learning regression (MLR) was used as a secondary model for dataset training, and 5-fold cross-validation, along with the out-of-sample estimation matrices, was used to train the MLR model. The specific stacking steps for this ensemble learning approach are illustrated in Figure 4.

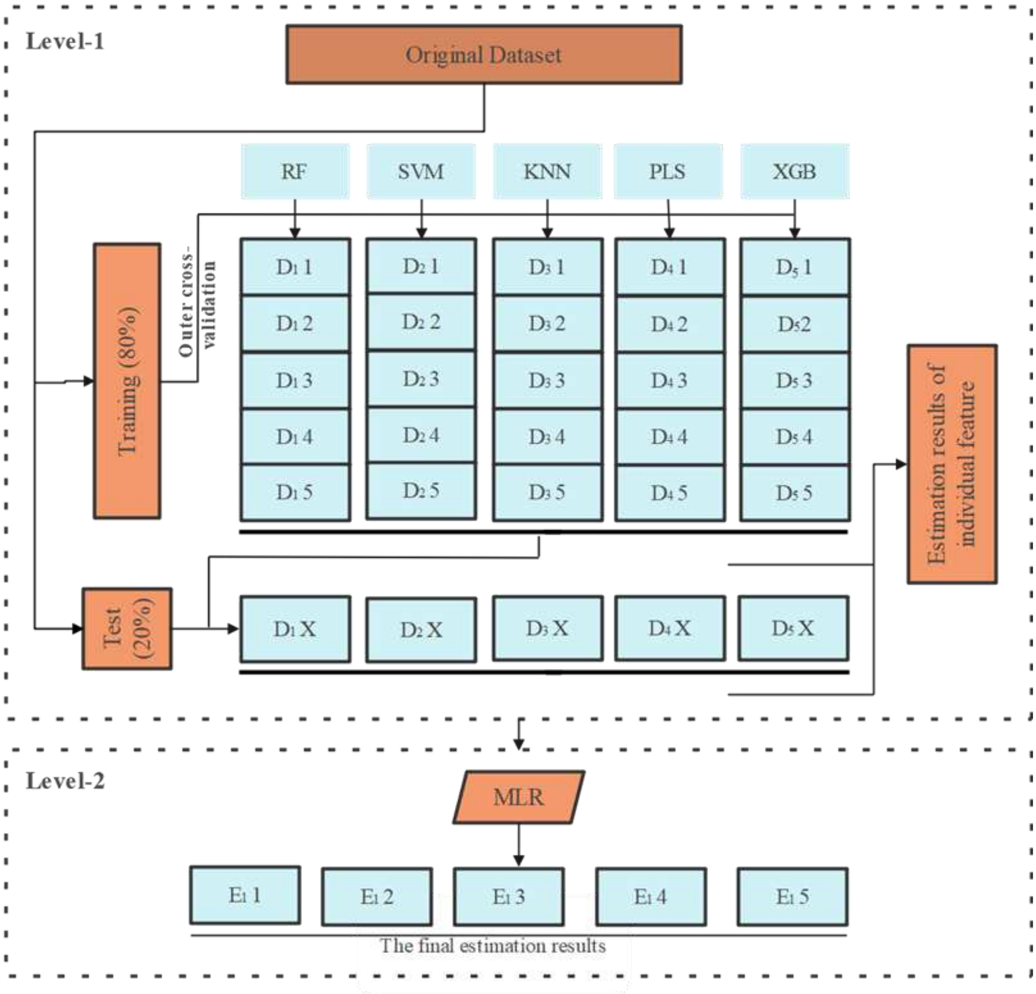

Figure 4. The workflow of the stacking ensemble learning model

2.6.2 Model evaluation

Two commonly used accuracy statistics, namely, the coefficient of determination (R^2) and mean absolute error (MAE), were employed for the quantitative assessment of the performance of all yield forecasting models. These metrics were calculated as follows:

$$R^2 = 1 - \sum_{i=1}^{n}(x_i - y_i)^2 \,/\, \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i) \tag{4}$$

where x_i is the measured yield of faba beans, y_i indicates the yield estimated by the model, $\bar{x}$ indicates the mean of the measured yield, and n represents the total number of tested samples.

3 Results

3.1 Phenotypic analysis and selection of feature bands

3.1.1 Analysis of measured yield data and spectral reflectance

To ensure data authenticity, crop yield data and spectral data were analyzed. A total of 144 yield values were measured in the experiment, and the analysis results are presented in Figure 5. The skewness of the yield dataset was found to be 0.0562, and the kurtosis was −0.571, indicating that the data approximately followed a normal distribution.

The spectral reflectance of the faba bean canopy, supplemented by the linear interpolation algorithm, was illustrated in Figure 6. The results showed significant variations in the spectral reflectance of the faba bean canopy across different growth periods and varieties. Nevertheless, comparing Figure 6(a), Figure 6(b), and Figure 6(c), it could be observed that the fundamental spectral patterns of the crop remained consistent, despite variations in canopy structure due to different varieties and growth periods. As the growth period progressed, the spectral reflectance in the near-infrared (NIR) band of the faba bean canopy gradually increased, indicating that NIR band spectral differences were evident across different growth periods.

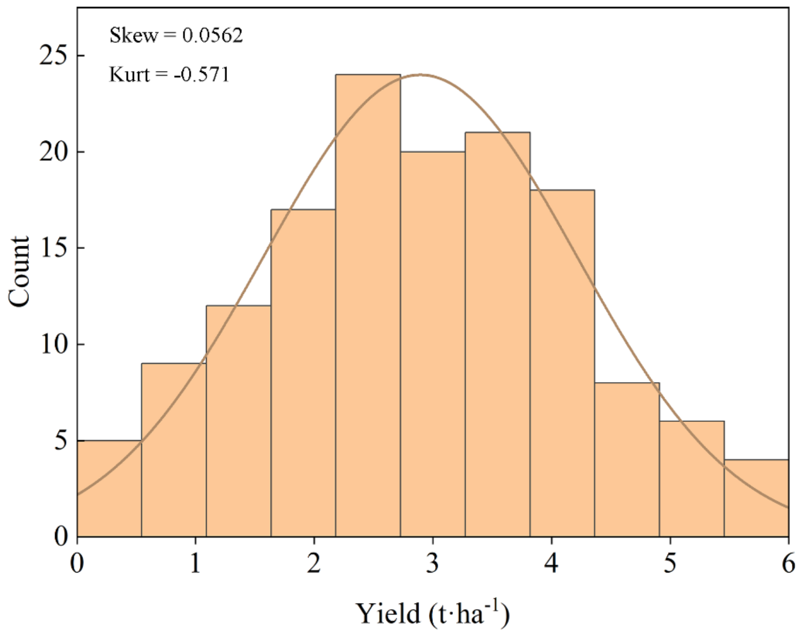

Figure 5. Frequency distribution of yield

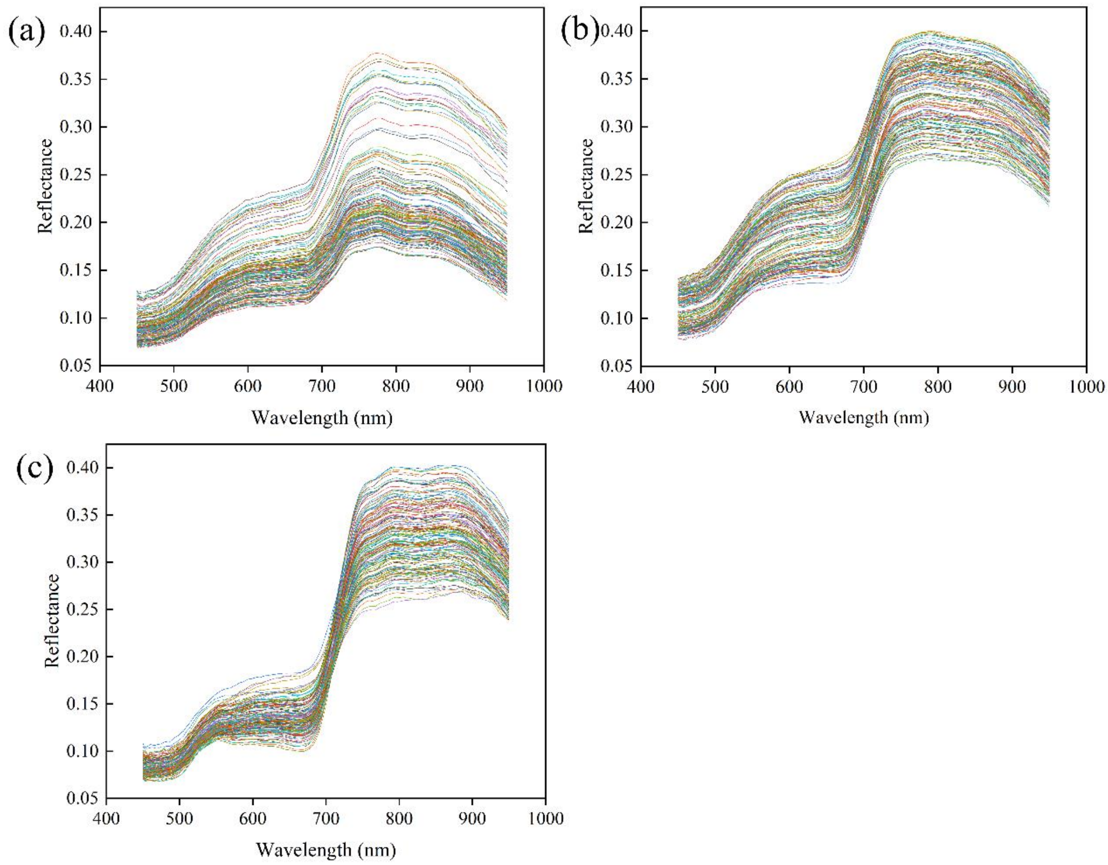

Figure 6. Spectral reflectance of all samples under different growth periods

Note: (a), (b), and (c) represent the spectral reflectance of P1, P2, and P3, respectively.

3.1.2 Selection of the feature bands

The results of the OIF calculation for screening the feature bands are presented in Table 2. These results indicate that the combination of bands 470 nm, 474 nm, 478 nm, 482 nm, and 790 nm within the P1 hyperspectral image has the highest OIF value of 3.2305; the combination of bands 790 nm, 798 nm, 802 nm, 806 nm, and 810 nm within the P2 hyperspectral image has the highest OIF value of 0.8953; and the combination of bands 670 nm, 674 nm, 786 nm, 790 nm, and 794 nm within the P3 hyperspectral image has the highest OIF value of 0.5524. These combinations were found to contain the most information among all respective hyperspectral image band combinations. Therefore, 470 nm, 474 nm, 478 nm, 482 nm, 670 nm, 674 nm, 790 nm, 794 nm, 798 nm, 802 nm, 806 nm, and 810 nm were selected as the feature bands for this study.

Table 2. Top five Optimal Index Factor values

Order	P1		P2		P3	
	Bands	OIF	Bands	OIF	Bands	OIF
1	470,474,478,482,790	3.2305	790,798,802,806,810	0.8953	670,674,790,794,798	0.5524
2	462,470,474,478,826	3.2304	794,802,806,810,814	0.8951	670,674,786,794,798	0.5523
3	470,474,482,486,826	3.2304	798,802,806,810,822	0.8950	670,674,786,790,798	0.5521
4	470,474,478,482,790	3.2303	798,802,806,814,818	0.8950	670,674,790,794,802	0.5520
5	462,470,474,478,790	3.2303	794,798,802,806,818	0.8948	670,674,782,790,794	0.5517

3.2 Performance of models for yield estimation

The faba bean yield estimation models were constructed using different features extracted from hyperspectral data and five base learners (RF, SVM, KNN, PLS, and XGB), along with the stacking ensemble. The test accuracy statistics of the different models are presented in Table 3. Model variables included growth periods, spectral features, texture features, and the algorithms used for yield estimation. The results indicate that for the same growth period and algorithm (Figure 6), the inclusion of texture information improved model accuracy by compensating for missing information in the spectra. For instance, the model fusing texture information based on P1 and RF exhibited better performance (R^2 = 0.548, MAE = 0.694 $t{\cdot}ha^{-1}$). Among the base learners, the XGB yield model performed best under the same growth period. The stacking ensemble approach further improved the model accuracy (Figure 7). Additionally, when the effects of different growth periods on model accuracy were compared, it was observed that model accuracy improved as the growth periods progressed, with the P3-based model being more accurate than the P1- and P2-based models (Figure 8). Furthermore, the fusion of growth periods contributed to the enhanced accuracy of the model (Figure 9).

Table 3. Test accuracy statistics of different models for yield estimation

Period	Data type	Metric	Base learners					Stacking
			RF	SVM	KNN	PLS	XGB	
P1	S	R^2	0.430	0.386	0.387	0.438	0.501	0.541
		MAE	0.777	0.819	0.828	0.807	0.745	0.709
P2	S	R^2	0.591	0.494	0.395	0.577	0.607	0.627
		MAE	0.660	0.723	0.794	0.630	0.642	0.604
P3	S	R^2	0.644	0.543	0.456	0.585	0.670	0.705
		MAE	0.583	0.718	0.741	0.622	0.527	0.494
P1+P2	S	R^2	0.599	0.559	0.465	0.572	0.624	0.640
		MAE	0.612	0.648	0.732	0.667	0.570	0.572
P2+P3	S	R^2	0.690	0.620	0.655	0.672	0.700	0.717
		MAE	0.559	0.631	0.583	0.554	0.526	0.493
P1	S+T	R^2	0.548	0.516	0.401	0.497	0.524	0.572
		MAE	0.694	0.640	0.779	0.713	0.718	0.687
P2	S+T	R^2	0.705	0.686	0.443	0.601	0.674	0.747
		MAE	0.577	0.582	0.781	0.629	0.619	0.544
P3	S+T	R^2	0.710	0.562	0.523	0.580	0.726	0.760
		MAE	0.514	0.570	0.636	0.629	0.501	0.474

3.3 Effects on model accuracy of spectral and texture information fusion

Figure 6 illustrates the effects of spectral data and the fusion of spectral and texture data on the estimation accuracy of different ML models. For the RF algorithm model, the model with fused texture data performed better in P1, with an increase in the R^2 value of 0.118 and a decrease in the MAE value of 0.083 $t{\cdot}ha^{-1}$, than the model based on individual spectral data. Similar results were obtained for P2 and P3; the model based on fused texture data performed better, with R^2 values 0.114 and 0.125 higher than those of the model based on individual spectral data in P2 and P3, respectively. However, an exception was noted in some ML models. The PLS model performed slightly worse in P3 than that based on individual spectral data after fusing texture information, with a reduction in model accuracy (R^2) of 0.005 and an increase in MAE of 0.007 $t{\cdot}ha^{-1}$. On the other hand, based on stacking ensemble learning, the model accuracy improved for all three growth periods (P1, P2, and P3) to a certain extent. For P1, the R^2 value of the model with fused texture data increased by 0.031 compared with the model based on spectral data, and the MAE value of the model with fused texture data decreased by 0.022 t·ha^{-1} compared with the model based on spectral data. P2 and P3 showed similar results to P1, with R^2 values 0.114 and 0.125 higher than the model based on individual spectral data, respectively, and MAE values 0.114 and 0.125 lower than the model based on individual spectral data. These results confirmed that in most cases, the estimation accuracy of the models based on the fusion of spectral and texture data was higher than that of the individual spectrum-based models.

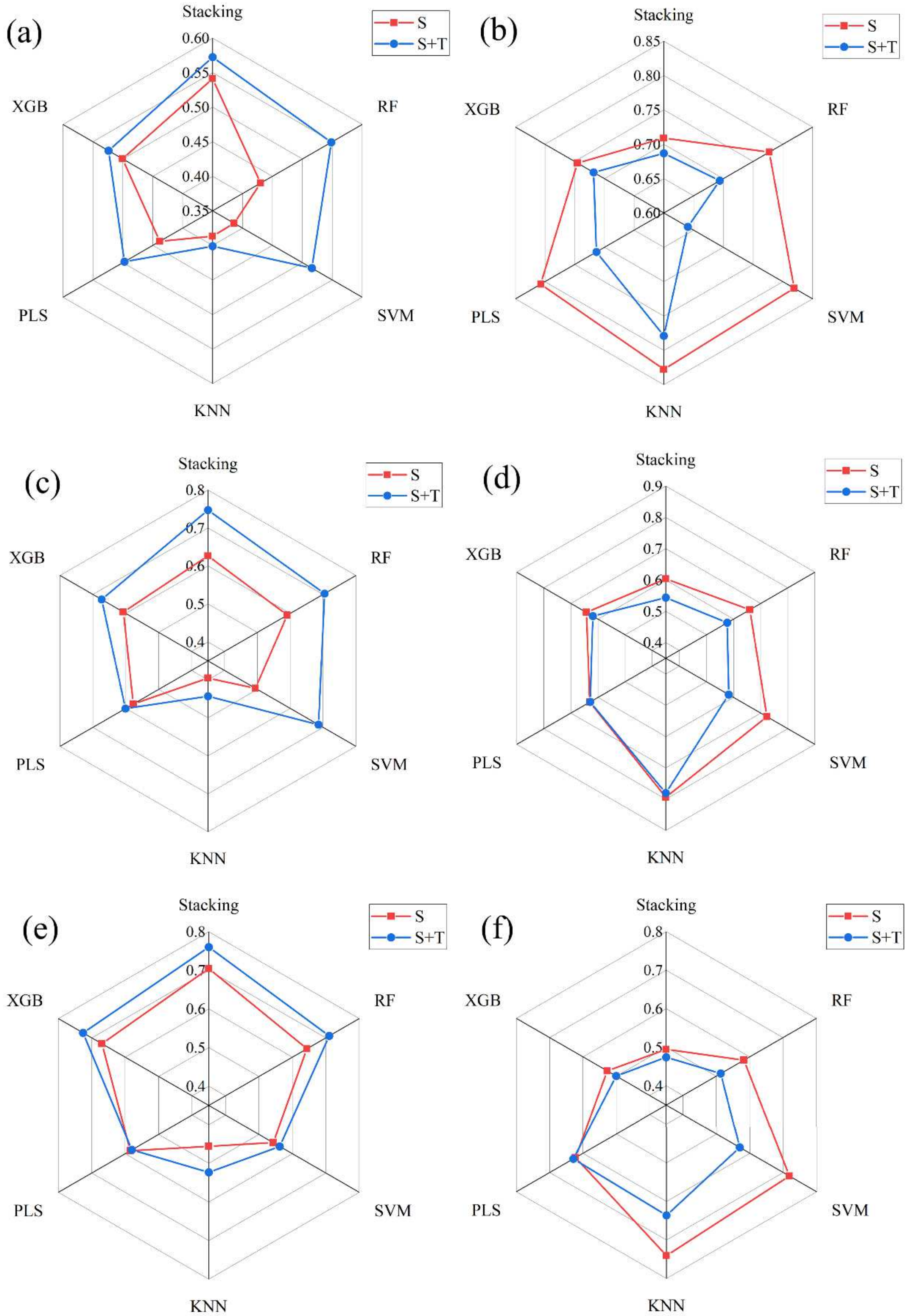

Figure 6. Comparison of the estimation accuracies between the spectral data-based model and the model based on the fusion of spectral and texture data

Note: (a) R^2 based on P1; (b) MAE based on P1; (c) R^2 based on P2; (d) MAE based on P2; (e) R^2 based on P3; (f) MAE based on P3.

3.4 Selection of optimal ML algorithm for yield estimation

To explore the effect of different base learners and ensemble learning on model accuracy, different models were constructed for yield estimation based on the same growth period and five algorithms. The model performance results were shown in Figure 7. In P1, the XGB model (R^2 = 0.501, MAE = 0.807 $t \cdot ha^{-1}$) performed best among the primary models, with R^2 values that were 0.071, 0.115, 0.114, and 0.063 times higher than those of the RF-, SVM-, KNN-, and PLS-based models, respectively, and error percentages that were 4.118%, 9.035%, 10.024%, and 7.683% lower than those of the RF-, SVM-, KNN-, and PLS-based models, respectively. Similarly, in terms of P2 and P3 (Figure 7b, Figure 7c), the XGB model outperformed the other primary models, followed by the RF model, whereas the KNN model showed the lowest performance among the primary models. Figure 7 demonstrates that the accuracy of ensemble learning-based models improved further compared with that of the best-performing primary model (XGB). The R^2 values of the ensemble learning-based models improved by 7.984%, 3.295%, and 5.224% compared with those of the XGB model in P1, P2, and P3, respectively, and their MAE values decreased by 4.832%, 5.919%, and 6.262%, respectively. In summary, the XGB model exhibited the best performance among the primary models in this study, and stacking ensemble learning contributed to improving the accuracy of the yield estimation model.

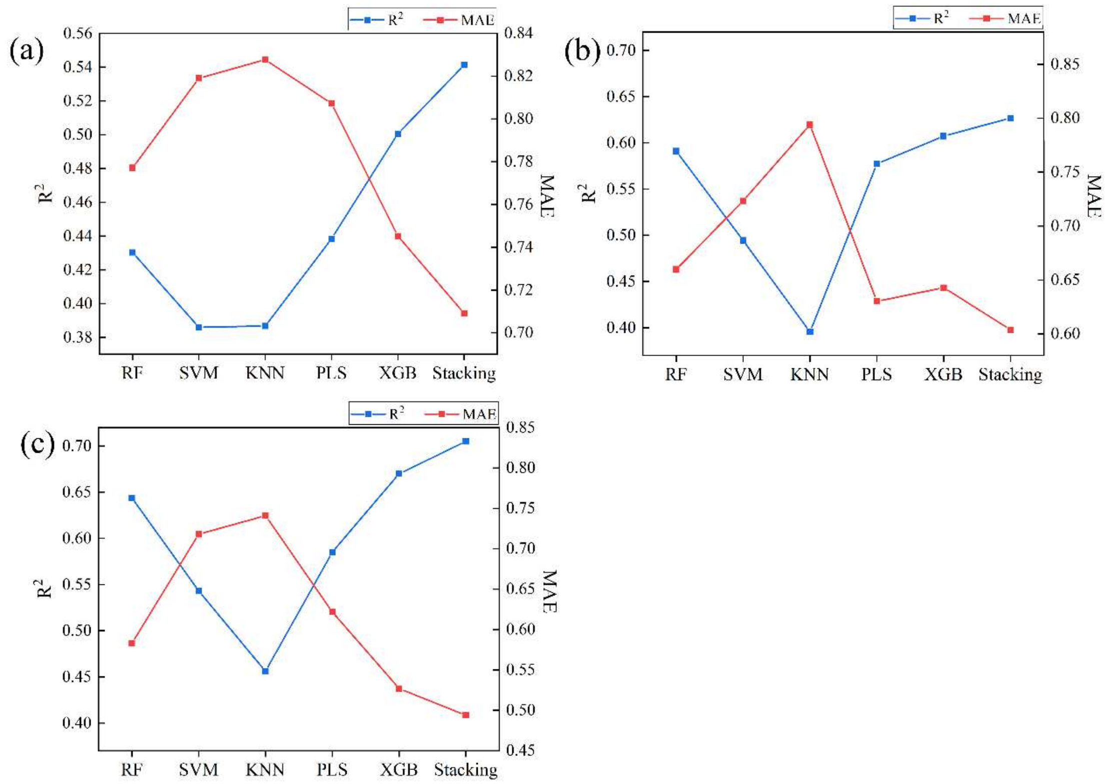

Figure 7. Effects on model accuracy of different ML algorithms

Note: (a) P1-based model statistical metrics of accuracy; (b) P2-based model statistical metrics of accuracy; (c) P3-based model statistical metrics of accuracy.

3.5 Effects of different growth periods on model accuracy

The effect of different growth periods on model accuracy was investigated, and the results were presented in Figures 8 and 9. Figure 8 illustrates the impact of different growth periods on the yield estimation model accuracy. For the RF algorithm, the accuracy (R^2) of the P3-based model was 8.968% higher than that of the P2-based model, and the P2-based model's accuracy was 37.442% higher than that of the P1-based model. Similar trends were observed for the other algorithms, with an increase in accuracy as the growth period advanced. However, it is worth noting that the rate of accuracy improvement in most algorithmic models decreased as the growth period proceeded. Figure 9 depicts the effect of fusing different growth periods on the yield estimation model. For the RF algorithm, the P1+P2-based model outperformed the P1- and P2-based models, and the fusion of P2 and P3 also contributed to enhancing the accuracy of the original model. The accuracy changes of the models based on other algorithms followed a similar pattern to that of the RF model, except for the PLS algorithm. In the case of the PLS algorithm, the R^2 of the P2-based model was slightly higher than that of the P1+P2-based model, and the MAE of the P2-based model was 0.037 $t \cdot ha^{-1}$ lower than that of the P1+P2-based model. In summary, based on the same algorithm, the model accuracy continuously improved with the growth periods. Furthermore, the results showed that fusing growth periods generally improved the model accuracy in most cases. However, there was a noticeable decrease in the rate of accuracy improvement for each model as the growth period progressed.

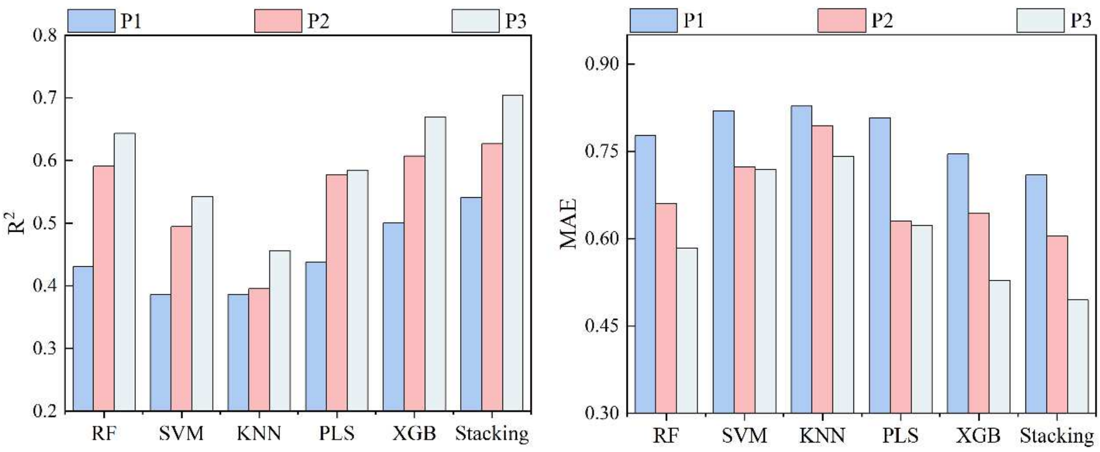

Figure 8. Effects of different growth periods on model accuracy

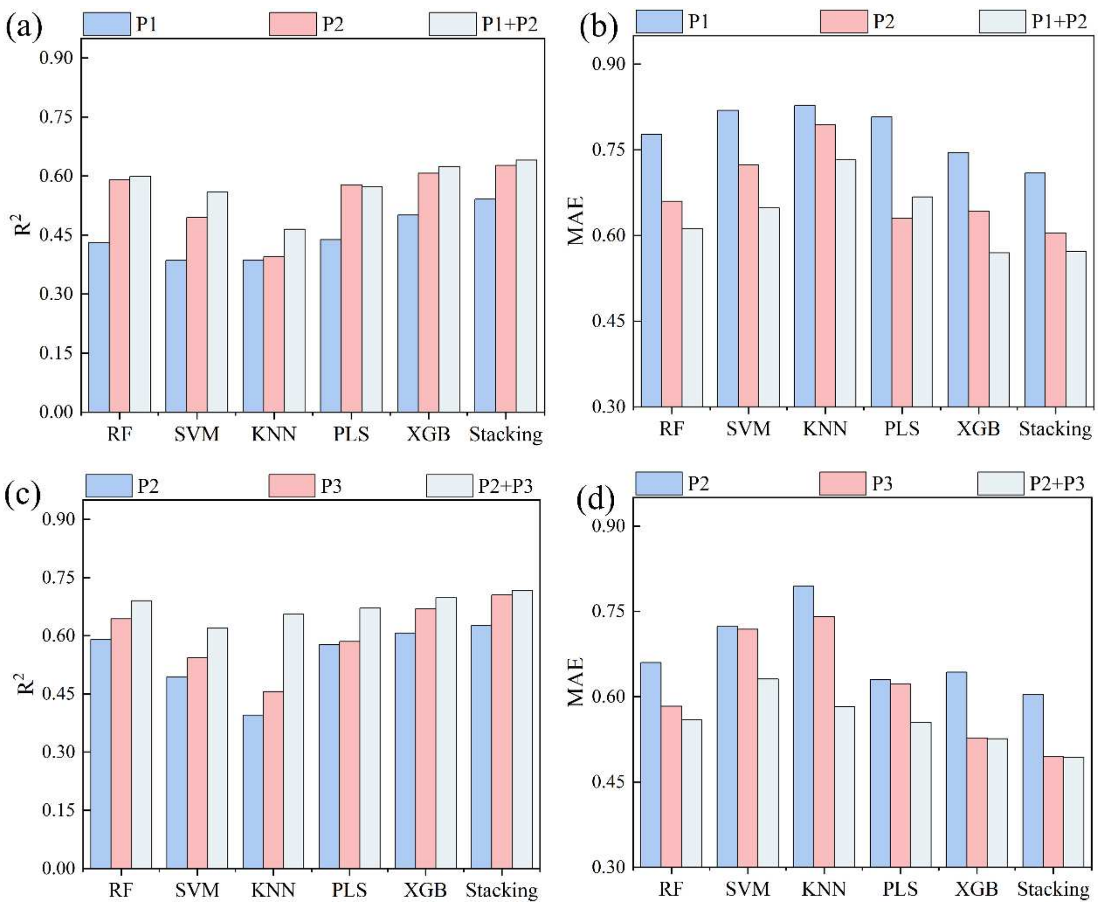

Figure 9. Effects of growth period fusion on model accuracy

Note: (a) R^2 of the P1-, P2-, and P1+P2-based models; (b) MAE of the P1-, P2-, and P1+P2-based models; (c) R^2 of the P2-, P3-, and P2+P3-based models; (d) MAE of the P2-, P3-, and P2+P3-based models.

4 Discussion

Faba bean is a valuable protein resource, and yield is a crucial phenotypic parameter for its evaluation. Early yield estimation can aid in developing accurate field planning, optimizing field production, and controlling cost inputs. The application of UAS-based imaging technology has proven effective for crop phenotype collection and estimation. Moreover, hyperspectral imagery has significantly enriched information content and holds unparalleled potential for impact and development compared to previous technologies. However, there has been a lack of studies focusing on the application of UAS-based hyperspectral technology to the phenotypic estimation of faba bean. In this study, we conducted an exploration of the factors influencing the yield estimation model of faba bean based on UAS hyperspectral images using stacking ensemble learning.

4.1 Contribution of texture information to yield estimation model accuracy

Texture information has been successfully applied in previous studies as structural information in crop phenotype estimation. Similar to the results of this study, the fusion of texture information and spectral information helped improve the accuracy of yield estimation models. This improvement can be attributed to the fact that texture information includes crop details that may not be present in spectral information alone [39]. Therefore, the inclusion of texture information complements spectral data and leads to enhanced model estimation accuracy [39]. However, there was an exception in this study where for P3 and PLS, the model based on the fusion of spectral and texture information performed slightly worse than the model based solely on spectral data. One possible explanation could be that the PLS algorithm is relatively simple and more suitable for small sample datasets, making it more susceptible to overfitting. With increased input data, the generalization ability of the PLS model may be reduced [40,41]. Moreover, the PLS algorithm is less noise-resistant and may not be well suited for the work of repeated observations of remote sensing. The fused texture information might contain more noise or irrelevant information, leading to a slight decrease in model estimation accuracy instead.

4.2 Effects of machine learning algorithms and ensemble learning algorithm on the performance of the estimation model

The results of this study revealed that the XGB model consistently outperformed other primary models, followed by the RF model, which aligns with the findings of Kavzoglu and Teke's study [42]. The superiority of XGB can be attributed to several advantages: (1) XGB is essentially a gradient boosting algorithm that iteratively trains multiple weak learners to minimize the loss function using gradient descent [43]. This iterative approach gradually reduces the residuals to better fit the data. (2) XGB provides a range of regularization techniques, such as regularization terms and pruning operations, which help prevent overfitting when dealing with multiple noises in hyperspectral data [44]. (3) The study involved multiple cross-validations, and XGB offers a rich set of parameter tuning options for selecting the best parameter configuration through cross-validation [45]. This allows for better tuning of the model to fit the data and improve performance.

Previous studies have typically used individual machine learning algorithms for crop phenotype estimation. However, individual ML algorithms have limitations in handling different types of data [46]. Issues such as an unsuitable sample size or the inability to handle noisy problems can lead to significant variations in model accuracy, which can result in different production decisions in the field. Therefore, the need to explore estimation methods with higher accuracy arises [14]. In this study, MLR was used as a secondary learner for stacked learning and model construction. Consistent with previous research findings, the ensemble model outperformed the individual algorithm models under various modeling conditions, demonstrating the feasibility and benefits of the ensemble learning approach [14,21,47].

4.3 Analysis of the effects of different growth periods on the performance of the estimation model

The ensemble models constructed based on different growth periods generally achieved satisfactory results (with some exceptions, such as the relatively poor performance of the KNN models). The findings demonstrated that the model estimation accuracy gradually improved as the growth period progressed under the same modeling conditions. However, it was observed that the rate of accuracy improvement decreased as the growth period progressed, consistent with the findings of previous studies and in accordance with crop growth patterns [48,49]. This phenomenon can be attributed to the fact that in the early growth period, the correlation between crop characteristic data and yield is weak. Additionally, early growth period images may contain more noisy or uncertain data, leading to less accurate estimation of crop phenotypes [50,51]. As the growth period advances, the learning capacity of the model may reach a saturation point, where the model has already learned the relevant features from the available data, and further improvement becomes limited, resulting in a gradual decrease in the rate of accuracy improvement.

Growth period data fusion was also explored in this study, and the results indicated that in most cases, the accuracy of the model based on fused multi-growth period data was higher than that of the model based on a single growth period. This observation can be attributed to the interconnectedness and dependence of stages in the growth process. The model benefits from contextual information provided by data from multiple growth periods, which enhances its ability to accurately estimate crop phenotypes [52]. Additionally, multiple factors and characteristics may exist within a single growth period, but without data from other growth periods for comparison, the model may not be able to accurately distinguish which characteristics are most influential in predicting the target yield [53]. This limitation can be overcome by using data from multiple growth periods, allowing the model to utilize information from various features across different fertility periods, thereby improving estimation accuracy. The study demonstrates the feasibility of constructing models using multiple sources of data, leading to enhanced accuracy in crop yield estimation.

4.4 Limitations and implications

Before this study, few researchers had explored faba bean phenotypes using hyperspectral data and integrated learning methods. Therefore, this study's construction of a yield estimation model using early growth period data based on canopy hyperspectral data and ensemble learning methods represents a significant step forward, yielding satisfactory results. These findings indicate the feasibility of using canopy hyperspectral data for faba bean yield model construction and early yield estimation, paving the way for further applications of hyperspectral imaging in early crop yield estimation.

However, it is essential to acknowledge some limitations in this study. While the early yield estimation models performed well, it remains to be investigated whether similar approaches yield better results when applied to other phenotypic traits. Future research could explore the fusion of data from LiDAR, hyperspectral, IKONOS, and other imaging sources for model construction, providing a comprehensive understanding of different crop phenotypes [54,55].

5 Conclusions

In this study, based on hyperspectral images, the effects of texture information, stacking ensemble learning and different growth periods on model accuracy were explored separately. The results of the study are as follows:

(1) Effects on the model accuracy of spectral and texture information fusion were explored in this study. In most cases, the model accuracy after fusion of texture information exhibited higher estimation accuracy than the model based on spectral data individually.

(2) Under the same modeling context of growth period and spectra, the XGB model performed the best among the base models; stacking ensemble learning improved model accuracy more significantly, and all ensemble learning-based models had satisfactory accuracy (the values of R^2 ranged from 0.541 to 0.760).

(3) The effects of data from different growth periods on model accuracy were explored. In the same modeling context with the same algorithm and spectral information, model accuracy increased as growth periods progressed, while the increase rate gradually decreased; in most cases, fusion of growth period data helped to improve model accuracy.

This article was the first study to apply hyperspectral image data to a model for phenotype estimation of faba bean. This study provides a new idea for the early estimation of faba bean yield and provides a reference for faba bean breeding work and other subsequent studies.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

We affirm that all authors have reviewed and approved the final version of the manuscript for publication in Plant Methods.

Availability of data and materials

The original data and materials presented in the study are included in the article, and further inquiries can be directed to the corresponding authors.

Competing interests

The authors declare that there are no conflicts of interest regarding the publication of this article.

Funding

This work was funded by China Agriculture Research System of MOF and MARA (CARS-08-G11), National Crop Genebank project from the Ministry of Science and Technology of China (NCGRC-2023-7), Project of accurate identification of faba bean germplasm resources, Agricultural Science and Technology Innovation Program in CAAS.

Author contributions

Y.C. analyzed the data and wrote the manuscript, Y.J. designed the study, S.F. offered the code of model construction, Z.L. helped with image processing, R.L. participated in data collection, and T.Y. and X.Z. offered the experimental platforms and research ideas.

Acknowledgments

We thank the other teachers for their contributions, and we are also grateful for the linguistic editing and proofreading during the preparation of the manuscript by TopEdit LLC.

Supplementary Materials

Additional file: Table S1. Vegetation indices used in this study.

References

[1] Everingham Y, Sexton J, Skocaj D, Inman-Bamber G. Accurate prediction of sugarcane yield using a random forest algorithm. Agron Sustain Dev. 2016;36:27.

[2] Rahman MM, Robson A. Integrating Landsat-8 and Sentinel-2 Time Series Data for Yield Prediction of Sugarcane Crops at the Block Level. Remote Sens. 2020;12(8):1313.

[3] Poudyal C, Costa L, Sandhu HS, Ampatzidis Y, Odero DC, Arbelo OC, et al. Sugarcane yield prediction and genotype selection using UAV-based hyperspectral imaging and machine learning. Agron J. 2022;114:2320–2333.

[4] Shafiee S, Lied LM, Burud I, Dieseth JA, Alsheikh M, Lillemo M. Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield

prediction based on UAV imagery. Comput. Electron. Agric. 2021;183:106036.
[5] Fei S, Li L, Han Z, Chen Z, Xiao Y. Combining novel feature selection strategy and hyperspectral vegetation indices to predict crop yield. Plant Methods. 2022;18(1):119.
[6] Montesinos-López OA, Montesinos-López A, Crossa J, Campos GDL, Alvarado G, Suchismita M, et al. Predicting grain yield using canopy hyperspectral refectance in wheat breeding data. Plant Methods. 2017;13:4.
[7] Qing L, Shi J, Jing Z, Xiao W, Zhuang S, Zi L, et al. Deciphering the contributions of spectral and structural data to wheat yield estimation from proximal sensing. Crop J. 2020;10:1334–1345.
[8] Li ZH, Li ZH, Fairbairn D, Li N, Xu B, Feng HK, et al. Multi-LUTs method for canopy nitrogen density estimation in winter wheat by field and UAV hyperspectral. Comput Electron Agric. 2019;162;174–182.
[9] Traore A, Ata-Ul-Karim ST, Duan AW, Soothar MK, Traore S, Zhao B. Predicting Equivalent Water Thickness in Wheat Using UAV Mounted Multispectral Sensor through Deep Learning Techniques. Remote Sens. 2021;13 (21):4476.
[10] Appeltans S, Guerrero A, Nawar S, Pieters J, Mouazen AM. Practical recommendations for hyperspectral and thermal proximal disease sensing in potato and leek fields. Remote Sens. 2020;12(12):1939.
[11] Zubler AV, Yoon J. Proximal methods for plant stress detection using optical sensors and machine learning. Biosensors. 2020;10(12):193
[12] Wang S, Guan K, Wang Z. Airborne hyperspectral imaging of nitrogen deficiency on crop traits and yield of corn by machine learning and radiative transfer modeling. Int. J. Appl. Earth Obs. Geoinf. 2021;105:102617.
[13] Li B, Xu X, Zhang L, Han J, Bian C, Li G, et al. Above-ground biomass estimation and yield prediction in potato by using UAV-based RGB and hyperspectral imaging. ISPRS J Photogramm Remote Sens. 2020;162:161–172.
[14] Fei S, Hassan MA, Xiao Y, Su X, Chen Z, Cheng Q, et al. UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat. Precis Agric. 2023;24:187–212.
[15] Ji Y, Chen Z, Cheng Q, Liu R, Li M, Yan X, et al. Estimation of plant height and yield based on UAV imagery in faba bean (Vicia faba L.). Plant Methods. 2022;18:1–13.
[16] Prasad N, Semwal M, Kalra A. Hyperspectral vegetation indices offer insights for determining economically optimal time of harvest in Mentha arvensis. Ind Crops Prod. 2022;180:114753.
[17] Roberts DA, Roth KL, Wetherley EB, Meerdink SK, Perroy RL. Hyperspectral vegetation indices. In: Hyperspectral Indices and Image Classifications for Agriculture and Vegetation. Boca Raton (America) CRC Press, 2018, pp. 3–26.
[18] Shu M, Zhou L, Chen H, Wang X, Meng L, Ma Y. Estimation of amino acid contents in maize leaves based on hyperspectral imaging. Front. Plant Sci. 2022;13:885794.
[19] Sarić R, Nguyen VD, Burge T, Berkowitz O, Trtílek M, Whelan J, et al. Applications of hyperspectral imaging in plant phenotyping. Trends Plant Sci. 2022;27(3):301–315.
[20] Wu L, Gong Y, Bai X, Wang W, Wang Z. Nondestructive Determination of Leaf Nitrogen Content in Corn by Hyperspectral Imaging Using Spectral and Texture Fusion. Appl Sci. 2023;13(3):1910.
[21] Fu P, Meacham-Hensold K, Guan K, Bernacchi CJ. Hyperspectral Leaf Reflectance as Proxy

for Photosynthetic Capacities: An Ensemble Approach Based on Multiple Machine Learning Algorithms. Front Plant Sci. 2019;10:730.
[22] Kaur S, Pandey S, Goel S. Semi-automatic leaf disease detection and classification system for soybean culture. IET Image Process. 2018;12:1038–1048.
[23] Mohan KJ, Balasubramanian M, Palanivel S. Detection and recognition of diseases from paddy plant leaf images. Int J Comput Appl. 2016;144(12):34–41.
[24] Bhadra S, Sagan V, Maimaitijiang M, Maimaitiyiming M, Newcomb M, Shakoor N, et al. Quantifying Leaf Chlorophyll Concentration of Sorghum from Hyperspectral Data Using Derivative Calculus and Machine Learning. Remote Sens. 2020;12(13):2082.
[25] Herrero-Huerta M, Rodriguez-Gonzalvez P, Rainey, KM. Yield prediction by machine learning from UAS-based mulit-sensor data fusion in soybean. Plant Methods. 2020;16:78.
[26] Jin X, Li Z, Feng H, Ren Z, Ren Z, Li S. Deep neural network algorithm for estimating maize biomass based on simulated Sentinel 2A vegetation indices and leaf area index. Crop J. 2020;8:87–97.
[27] Zhang Y, Yang Y, Zhang Q, Duan R, Liu J, Qin Y, et al. Toward Multi-Stage Phenotyping of Soybean with Multimodal UAV Sensor Data: A Comparison of Machine Learning Approaches for Leaf Area Index Estimation. Remote Sens. 2023;15(1):7.
[28] Chen J, Zeb A, Nanehkaran YA, Zhang D. Stacking ensemble model of deep learning for plant disease recognition. J Ambient Intell Human Comput. 2022;14:12359–12372.
[29] Feng L, Zhang Z, Ma Y, Du Q, Williams P, Drewry J, et al. Alfalfa Yield Prediction Using UAV-Based Hyperspectral Imagery and Ensemble Learning. Remote Sens. 2020;12(12):2028.
[30] Divina F, Gilson A, Goméz-Vela F, Torres MG, Torres JF. Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting. Energies. 2018;11(4):949.
[31] Ribeiro MHDM, Coelho LDS. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Appl Soft Comput. 2020;86:105837.
[32] Soares E, Costa Jr P, Costa B, Leite D. Ensemble of evolving data clouds and fuzzy models for weather time series prediction. Appl Soft Comput. 2018;64:445–453.
[33] Wang Q, Che Y, Shao K, Zhu J, Wang R, Sui Y, et al. Estimation of sugar content in sugar beet root based on UAV multi-sensor data. Comput Electron Agric. 2022;203:107433.
[34] Chavez PS, Berlin GL. Statistical method for selecting Landsat MSS ratios. J Appl Photogr Eng. 1982;8(1):23–30.
[35] Noor M, Yahaya AS, Ramli NA, Al Bakri AM. Filling Missing Data Using Interpolation Methods: Study on the Effect of Fitting Distribution. Key Engineering Materials. 2013;594:889–895.
[36] Lehnert L, Meyer H, Obermeier W, Silva B, Regeling B, Bendix J. Hyperspectral Data Analysis in R: The hsdar Package. Journal of statistical software. 2019;89:1–23.
[37] Wolpert DH. Stacked generalization. Neural Network. 1992;5:241–259.
[38] Fei S, Hassan MA, He Z, Chen Z, Shu M, Wang J, et al. Assessment of Ensemble Learning to Predict Wheat Grain Yield Based on UAV-Multispectral Reflectance. Remote Sens. 2021;13(12):2338.
[39] Zheng H, Ma J, Zhou M, Li D, Yao X, Cao W, et al. Enhancing the Nitrogen Signals of Rice Canopies across Critical Growth Stages through the Integration of Textural and Spectral Information from Unmanned Aerial Vehicle (UAV) Multispectral Imagery. Remote Sens. 2020;12(6):957.

[40] Jia W, Zhao D, Ding L. An optimized RBF neural network algorithm based on partial least squares and genetic algorithm for classification of small sample. Appl Soft Comput. 2016;48:373–384.
[41] Yue J, Feng H, Yang G, Li Z. A Comparison of Regression Techniques for Estimation of Above-Ground Winter Wheat Biomass Using Near-Surface Spectroscopy. Remote Sens. 2018;10(1):66.
[42] Kavzoglu T, Teke A. Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost). Arab J Sci Eng. 2022;47: 7367–7385.
[43] Elavarasan D, Vincent PM. Reinforced XGBoost machine learning model for sustainable intelligent agrarian applications. J. Intell. Fuzzy Syst. 2020; 39:7605-7620.
[44] Chen Tianqi, Guestrin Carlos. XGBoost: A Scalable Tree Boosting System. KDD ’16. San Francisco, CA, USA. 2016; pp785-794.
[45] Carmona P, Climent F, Momparler A. Predicting bank failure in the U.S. banking sector: An extreme gradient boosting approach. Int Rev Econ Financ. 2019;61:304–323.
[46] Matese A, Di Gennaro SF. Beyond the traditional NDVI index as a key factor to mainstream the use of UAV in precision viticulture. Sci Rep. 2021;11:2721.
[47] Ji Y, Liu R, Xiao Y, Cui Y, Chen Z, Zong X, et al. Faba bean above-ground biomass and bean yield estimation based on consumer-grade unmanned aerial vehicle RGB images and ensemble learning. Precision Agric. 2023;24:1439–1460.
[48] Feng P, Wang B, Liu DL, Waters CM, Xiao D, Shi L, et al. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agr and Forest Meteorol. 2020:107922.
[49] Al-Shammari D, Fuentes I, Whelan BM, Filippi P, Bishop TFA. Mapping of Cotton Fields Within-Season Using Phenology-Based Metrics Derived from a Time Series of Landsat Imagery. Remote Sens. 2020;12(18):3038.
[50] Basnyat P, Mcconkey BG, Lafond G, Moulin A, Pelcat Y. Optimal time for remote sensing to relate to crop grain yield on the Canadian prairies. Can J Plant Sci. 2004;84:97–103.
[51] Holman FH, Riche AB, Michalski A, Castle M, Wooster MJ, Hawkesford MJ. High Throughput Field Phenotyping of Wheat Plant Height and Growth Rate in Field Plot Trials Using UAV Based Remote Sensing. Remote Sens. 2016;8(12):1031.
[52] Gonzalez N, Vanhaeren H. Leaf size control: Complex coordination of cell division and expansion. Trends Plant Sci. 2012;17(6):332–340.
[53] Zhao B, Zhang J, Yang C, Zhou G, Ding Y, Shi Y, et al. Rapeseed Seedling Stand Counting and Seeding Performance Evaluation at Two Early Growth Stages Based on Unmanned Aerial Vehicle Imagery. Front Plant Sci. 2018;9:1362.
[54] Debes C, Merentitis A, Heremans R, Hahn JT, Frangiadakis N, Kasteren TV, et al. Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest. IEEE J Sel Top Appl Earth Obs Remote Sens. 2014;7(6):2405–2418.
[55] Sohn G, Dowman IJ. Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction *. ISPRS J Photogramm Remote Sens. 2007;62:43–63.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.docx