COUPLING UAV MULTISPECTRAL IMAGERY AND MACHINE LEARNING TO CONSTRUCT A MONITORING AND PREDICTION MODEL FOR SOYBEAN GRAIN MOISTURE CONTENT AT MATURITY

/

基于无人机多光谱数据和机器学习的成熟期大豆籽粒含水率的监测预测模型

Lulu LV[1)], Chengqian JIN [*1),2)], Tengxiang YANG[2)], Anqi JIANG[1)], Han YAN[1)]

[1)] College of Agricultural Engineering and Food Science, Shandong University of Technology, Zibo 255000, China

[2)] Nanjing Institute of Agricultural Mechanization, Ministry of Agriculture and Rural Affairs, Nanjing 210014, China

**Corresponding authors: Chengqian JIN; Tel: +86-15366092900; E-mail: jinchengqian@caas.cn*

DOI: https://doi.org/10.35633/inmateh-78-04

***Keywords**: UAV multispectral, soybean, grain moisture content, feature selection, machine learning*

ABSTRACT

Soybean (Glycine max) grain moisture content (MC) at harvest affects yield, storage, and processing quality, but traditional measurements are laborious and unsuitable for large-scale monitoring. This study aimed to develop an efficient method for estimating soybean (Glycine max) grain moisture content (MC) at maturity, addressing the limitations of traditional labor-intensive measurements. UAV-based multispectral imagery from a DJI Mavic 3M was used to extract spectral reflectance and vegetation indices (VIs). Three feature selection techniques (SHAP, RFA, ReliefF) and six regression models (PLSR, SVR, MLR, RFR, XGBoost, RR) were applied to identify key predictors and optimize performance. Results showed that SVR using spectral reflectance achieved the highest accuracy (R^2 = 0.763, RMSE = 1.473), while RFR performed best for combined spectral and VI features. The RE and NIR bands were the most sensitive to MC. The findings demonstrate that integrating UAV multispectral data with machine learning and feature selection enables accurate, rapid, and non-destructive prediction of soybean MC, supporting precision harvest and crop management.

摘要

大豆（Glycine max）收获时的籽粒含水量（MC）影响产量、储存和加工质量，但传统测量费时费力，不适用于大规模监测。本研究旨在开发一种高效的方法，用于估算大豆（Glycine max）在成熟期的籽粒含水量（MC），以解决传统人工测量方法的局限性。使用大疆 Mavic 3M 无人机获取的多光谱影像提取了光谱反射率和植被指数（VIs）。应用了三种特征选择技术（SHAP、RFA、ReliefF）和六种回归模型（PLSR、SVR、MLR、RFR、XGBoost、RR）来识别关键预测因子并优化性能。结果表明，使用光谱反射率的 SVR 模型获得了最高精度（R^2 =0.763，RMSE=1.473），而 RFR 模型在结合光谱和 VI 特征时表现最佳。红边（RE）和近红外（NIR）波段对 MC 最敏感。研究结果表明，将无人机多光谱数据与机器学习和特征选择相结合，能够实现对大豆 MC 的准确、快速、无损预测，为精准收获和作物管理提供支持。

INTRODUCTION

Soybean (*Glycine max*), originating in China, is a crucial grain and economic crop valued for its high protein content and broad industrial applications. With the rising global demand for soybean-based products, increasing attention has been focused on improving both yield and quality. Grain moisture content (MC) at harvest is a critical parameter influencing mechanical harvesting efficiency, grain loss, storage stability, and processing quality; thus, accurate determination of the optimal harvest window is essential to ensure yield and quality *(Alabi, et al. 2022)*. As a major food and oil crop, soybean grain MC also serves as a key indicator of crop maturity and an important reference for determining the appropriate harvest time. However, conventional moisture determination methods, though accurate, are labor-intensive and time-consuming, limiting their application in large-scale precision agriculture. Therefore, developing rapid, accurate, and non-destructive methods for monitoring soybean grain MC is vital for optimizing harvest timing and reducing post-harvest losses. In this context, unmanned aerial vehicle (UAV)-based remote sensing has shown significant potential due to its high spatial resolution, flexibility, and efficiency in agricultural monitoring. Particularly, multispectral UAV imagery enables accurate assessment of crop reflectance and moisture status, providing an innovative, non-destructive solution for identifying the optimal harvest window of soybean *(Shi, et al. 2024)*.

Significant progress has been achieved globally in the remote sensing-based retrieval of crop water content. Existing studies have predominantly relied on satellite remote sensing or ground-based hyperspectral observations, often combined with machine learning methods to establish quantitative relationships between spectral features and measured data. For example, *Yang et al., (2025a)*, estimated rice grain moisture content using UAV multispectral imagery and machine learning algorithms; *Sun et al., (2020)*, employed hyperspectral imaging and developed support vector regression (SVR) and partial least squares regression (PLSR) models to achieve non-destructive detection of barley grain moisture; *Shu et al., (2025)*, proposed an integrated UAV hyperspectral and machine learning framework for wheat water monitoring; *Yang et al., (2025b)*, combined deep learning with high-resolution UAV imagery to assess water status in winter wheat and summer maize. In addition, *Yakubu et al., (2024)*, and *Ren et al., (2025)*, also demonstrated promising results in monitoring water content in wheat and maize, respectively. Collectively, these studies indicate that remote sensing combined with machine learning holds substantial potential for crop water content estimation *(Tulu, et al. 2025)*. Nevertheless, the majority of current research has been directed toward the estimation of leaf- and canopy-level water content or biomass, while UAV-based multispectral inversion of grain moisture content at harvest is still scarce.

To address this gap, the present study investigates soybean grain moisture content during the harvest period. High-resolution multispectral imagery was acquired using a DJI Mavic 3M UAV platform, and multiple machine learning algorithms were employed for model development and comparative evaluation. Three feature selection methods were applied to spectral and vegetation index variables, followed by grid search optimization to improve predictive accuracy and model robustness. The objective of this study is to establish a rapid and reliable soybean grain moisture prediction approach based on UAV multispectral data, thereby providing theoretical and technical support for scientific harvest decision-making and advancing the intelligent and precise development of soybean production.

MATERIALS AND METHODS

Overview of the study area

The study was conducted in Suixi County, Huaibei City, Anhui Province, located in the central part of the Huaibei Plain. The terrain is relatively flat, with an elevation ranging from 23.5 to 32.4 m, and the geographical coordinates are 116°36′–116°37′E and 35°48′N (Fig. 1). This region belongs to the Yellow River alluvial plain, with soil types dominated by clay loam. The soil exhibits uniform texture, favorable irrigation and drainage conditions, and medium-to-high fertility, making it suitable for the cultivation of a variety of crops. The area has a warm temperate monsoon continental climate, characterized by four distinct seasons and a coincidence of rainfall and heat. The major crops cultivated in this region include wheat, maize, soybean, and rice.

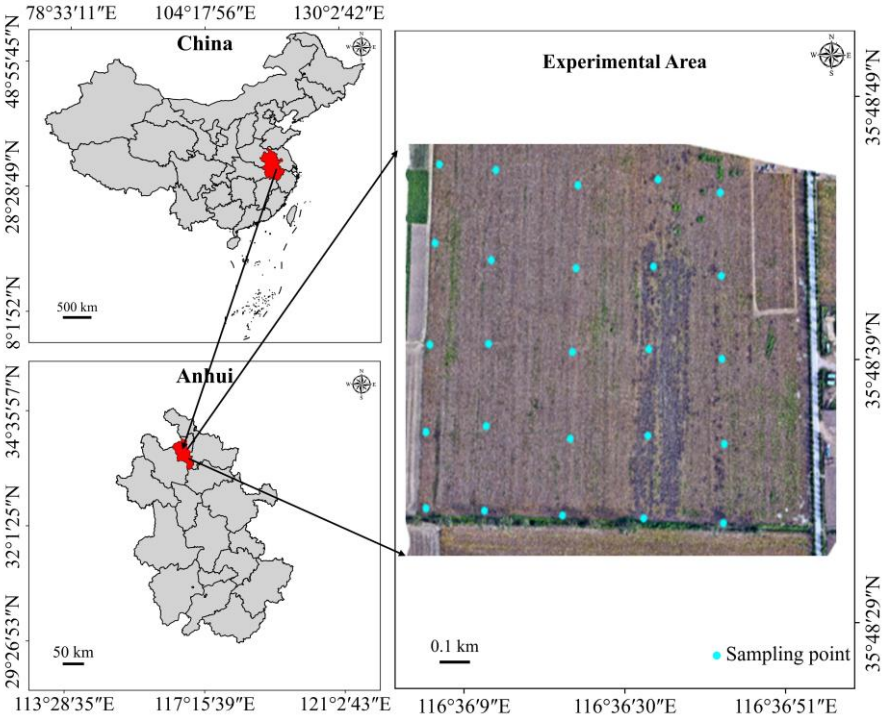


Fig. 1 - Geographical location of the study area

In this study, the soybean cultivar *Sudou 21* was selected as the experimental material. The sowing date was June 10, 2024, with a planting density of 225,000 plants per hectare, a row spacing of 40 cm, and a plant spacing of 11 cm. The harvest period was from October 14 to October 16, 2024.

UAV image data acquisition

In this study, a DJI Mavic 3M multispectral UAV (DJI Innovations, Shenzhen, China) was used for data acquisition, equipped with four 5-megapixel single-band sensors (Green, Red, red edge, NIR), a 20-megapixel RGB camera, an RTK positioning module, and a top-mounted irradiance sensor to improve reflectance accuracy *(Barbosa Júnior, et al. 2025)*. The main UAV parameters are summarized in Table 1. Flights were conducted from 10:00–12:00, October 14–16, 2024, under clear, calm conditions at 80 m altitude, with 2000×1500-pixel resolution, 80% forward and side overlap, and a 2 s acquisition interval. Each exposure captured four multispectral TIF images and one RGB image. A total of 577 datasets (2885 images) were collected, providing the basis for spectral feature extraction and soybean grain moisture content estimation.

Table 1

Parameters of the Mavic 3 M multispectral camera

Band	Name	Wavelength (nm)	Exposure adjustment
1	Green	560 nm	16 nm
2	Red	650 nm	16 nm
3	RE	730 nm	16 nm
4	NIR	860 nm	26 nm

Soybean moisture content determination

Field experiments were conducted during soybean maturity (October 14–16, 2024) with sampling points at 70 m intervals. Fresh weights were measured, and samples were oven-dried to determine dry weight. Grain moisture content was calculated using Equation 1, with GPS coordinates recorded for spatial accuracy (Table 2, Fig. 2).

$$C_W = \frac{L_W - L_d}{L_W} \times 100\% \quad (1)$$

where C_W represents the soybean moisture content (%), L_W is the fresh weight of the soybean sample at the time of collection (mg), and L_d is the dry weight after oven-drying (mg).

Fig. 2 - Actual determination of soybean grain moisture content

Table 2

Statistics of soybean grain moisture content

Date	Number of samples	Maximum (%)	Minimum (%)	Average value (%)
10.14	25	25.15	16.25	20.86
10.15	25	23.19	14.46	18.52
10.16	25	21.25	12.44	16.23

Multispectral image processing

During data acquisition, the integrated solar irradiance sensor of the Mavic 3M recorded ambient light conditions, providing essential information for subsequent radiometric calibration and effectively improving the accuracy and stability of reflectance calculations. After image collection, the data were imported into DJI Terra V4.2 software.

By selecting the multi-camera system layout, the software automatically aligned and arranged images across all spectral bands. Radiometric calibration was then performed by enabling the solar sensor option, completing the basic reflectance correction. Additional optimization parameters were applied to compensate for camera model errors and enhance the precision of RTK positioning data. To obtain the true ground spectral reflectance, each channel of the orthomosaic imagery was radiometrically corrected to convert the digital numbers (DN values) recorded by the UAV sensors into physically meaningful surface reflectance. The radiometric calibration equation is presented in Equation 2. By utilizing calibration panels with known reflectance, raw digital values are converted to standardized spectral reflectance and adjusted for each spectral band to ensure consistency and comparability across images *(Wang, et al. 2024)*.

$$\frac{D_{NS}}{R_S} = \frac{D_N}{R} \tag{2}$$

where D_{NS} is the DN value of the calibration panel captured by the UAV (pixel brightness in remote sensing imagery), R_S is the known reflectance of the calibration panel, D_N is the DN value of the soybean field obtained by the UAV, and *R* represents the calibrated reflectance of the soybean experimental plots.

Following radiometric calibration of the orthomosaic images, the processed images were imported into ArcMap software for spectral data extraction. The workflow involved cropping the imagery to the extent of the soybean experimental fields and extracting reflectance values at the exact locations corresponding to field sampling points using their geographic coordinates. The extracted spectral data served as critical input for subsequent modeling and analysis *(Khose and Mailapalli 2024)*.

Vegetation index selection

Vegetation exhibits pronounced differences in reflectance and absorption across spectral regions. Specifically, healthy Green vegetation strongly absorbs incident radiation in the blue and red wavelengths, while displaying markedly high reflectance in the Green and near-infrared (NIR) regions. Such spectral characteristics constitute the theoretical foundation for the development of vegetation indices (VIs). By mathematically integrating information from different spectral bands, VIs enhance the sensitivity of vegetation signals while effectively suppressing background effects and environmental noise, thereby serving as a fundamental tool for quantitative vegetation monitoring in remote sensing applications *(Wang, et al. 2022)*.

Over the past decades, numerous VIs have been proposed and widely applied in studies addressing crop growth monitoring, pest and disease detection, drought assessment, and the estimation of yield and moisture content. In the present study, building upon the spectral response characteristics of soybean canopies and insights from previous research, 16 VIs closely associated with soybean grain moisture content were selected to provide a concise yet reliable metric of canopy condition *(Wang, et al. 2025)*.

Spectral Outlier handling

To ensure model stability and predictive accuracy, the identification and removal of outliers from spectral data is a crucial preprocessing step. Previous studies have demonstrated that the boxplot method is effective in detecting and eliminating anomalous observations, thereby improving model robustness. In this study, boxplot-based statistical analysis was first applied to the collected multispectral reflectance data of soybean canopies to evaluate variability and dispersion in the raw dataset. The results revealed considerable differences in reflectance distributions across spectral bands, with the interquartile ranges (25%–75%) of the Green, Red, RE, and NIR channels measured as 56.586–507.420, 71.958–703.155, 119.999–997.877, and 192.315–1160.934, respectively, reflecting both distributional characteristics and variation patterns across bands.

Outlier detection was conducted using the interquartile range (IQR) criterion, defined as the difference between the upper (Q3) and lower (Q1) quartiles. Data points falling above the threshold (Q3 + 1.5IQR) or below the threshold (Q1 – 1.5IQR) were classified as outliers *(Roma, et al. 2025)*. The analysis indicated that abnormal values were present in all four channels (Green, Red, RE, and NIR). To further improve data reliability, the three-sigma (3σ) rule was additionally employed, whereby samples outside the range of μ ± 3σ (μ representing the mean and σ the standard deviation) were excluded. This approach, under the assumption of approximate normal distribution, effectively balanced data completeness with reliability *(Sim, et al. 2005)*. Boxplots before and after filtering are presented in Fig. 3.

Following this two-step screening and correction procedure, 73 valid samples were retained for subsequent analysis and modeling. This preprocessing not only eliminated the adverse effects of outliers but also provided a more stable basis for input features, thereby enhancing both the accuracy and generalization ability of the estimation models.

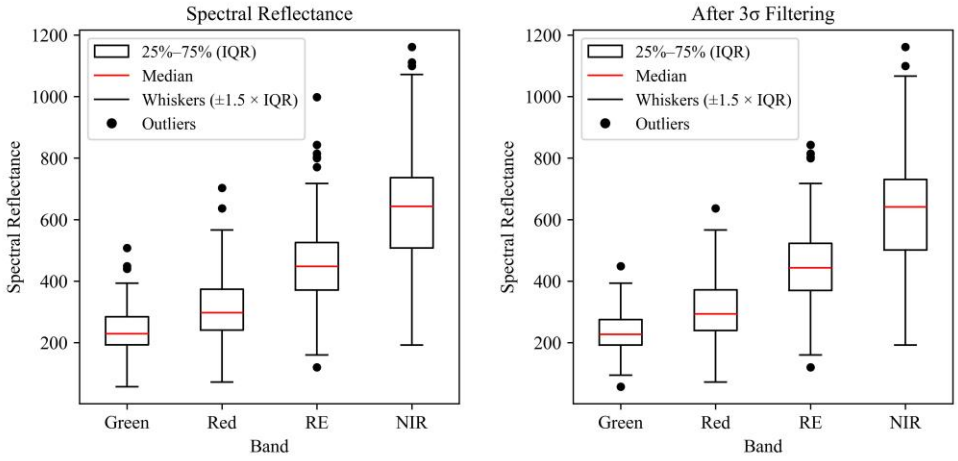


Fig. 3 - Spectral reflectance box line diagram

Model construction and accuracy testing

Six regression algorithms were employed to construct models for estimating soybean grain moisture content at maturity, including Partial Least Squares Regression (PLSR) *(Burnett, et al. 2021)*, Support Vector Regression (SVR) *(Sun, et al. 2025)*, Multiple Linear Regression (MLR) *(Kokaly and Clark 1999)*, Random Forest Regression (RFR) *(Liu, et al. 2025)*, Extreme Gradient Boosting (XGBoost) *(Qi, et al. 2025)*, and Ridge Regression (RR) *(Imani and Ghassemian 2015)*.

PLSR is a classical technique for high-dimensional datasets with severe multicollinearity, especially when predictors exceed samples and variables are strongly correlated. It projects both independent (X) and dependent (y) variables into a latent space, extracting components that maximize covariance, thereby enabling simultaneous dimensionality reduction and regression modeling. Widely applied in spectral analysis and agricultural remote sensing, its formulation and final regression expressions are shown in Equations (3)-(5).

$$X = TP^T + E \tag{3}$$

$$y = Tq + f \tag{4}$$

$$\hat{y} = XW(P^TW)^{-1}q \tag{5}$$

where T denotes the score matrix, P the loading matrix, q the response loading vector, and E and f represent residual terms.

SVR is an extension of the support vector machine framework for nonlinear regression. Its core principle is to construct a regression function in the feature space that tolerates deviations within an ε-insensitive margin while maximizing model generalization. This makes SVR highly effective for modeling complex nonlinear relationships. The optimization objective and constraints are expressed in Equations (6) and (7).

$$\min_{W,b,\delta_i,\delta_i^*} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}(\delta_i + \delta_i^*) \tag{6}$$

Constraints:

$$\begin{cases} y_i - (W^T\emptyset(x_i) + b) \le \varepsilon + \delta_i \\ (W^T\emptyset(x_i) + b) \le \varepsilon + \delta_i^* \\ \delta_i, \delta_i^* \ge 0 \end{cases} \tag{7}$$

where $\emptyset(\cdot)$ denotes the kernel function, C the penalty parameter, and ε the error tolerance.

MLR is one of the most fundamental regression approaches, assuming a linear relationship between predictors and the response variable. Coefficients are estimated using ordinary least squares. Although MLR is valued for simplicity and interpretability, it is sensitive to multicollinearity, often requiring feature selection or dimensionality reduction in spectral applications. The general formulation is given in Equation (8).

$$y = X\beta + \varepsilon \tag{8}$$

where X represents the input matrix, β the regression coefficients, and ε the error term.

RFR is an ensemble learning algorithm based on the bagging strategy. It constructs multiple regression trees and averages their predictions, thereby improving model stability and reducing the risk of overfitting. The predictive expression is shown in Equation (9).

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} T_m(x) \tag{9}$$

where T_m denotes the m-th regression tree and M the total number of trees.

XGBoost is a highly efficient and scalable implementation of gradient boosting decision trees (GBDT). It incrementally builds multiple regression trees by fitting to the residuals of previous trees, while introducing regularization terms to prevent overfitting. The predictive model, objective function, and regularization expressions are provided in Equations (10)–(12).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F \tag{10}$$

$$L(\emptyset) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{11}$$

$$\Omega\left(f_k\right) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{12}$$

where:

$\hat{y}_i$ represents the predicted value of the i-th sample, K the total number of trees, and $f_k(x_i)$ the prediction of the k-th tree for sample x_i, the function space F includes all possible CART trees, $l(\cdot)$ denotes the loss function, and Ω represents the regularization term, T denotes the number of leaves, w the leaf weights, γ the penalty coefficient for the number of leaves, and λ the L2 regularization coefficient for leaf weights.

RR extends linear regression by introducing an L2 regularization term into the loss function. By penalizing the squared magnitude of regression coefficients, RR effectively mitigates the effects of multicollinearity and reduces the risk of overfitting, thereby improving model stability. The optimization objective is defined in Equation (13).

$$\hat{\beta} = arg\ \min_{\beta}\{\|y\text{-}X\beta\|^2 + \lambda\|\beta\|^2\} \tag{13}$$

where:

λ is the regularization parameter controlling penalty strength. When $\lambda = 0$, RR degenerates into ordinary least squares regression.

After excluding outliers, 73 samples were retained for analysis. Due to the limited dataset, model performance was evaluated using five-fold cross-validation, in which the data were partitioned into five subsets, four for training and one for validation, and the average results were reported to reduce variance and mitigate overfitting *(Bin, et al. 2025)*. Although this approach improves reliability under small-sample conditions, the limited number of observations may still constrain the generalization capacity of complex models such as XGBoost and Random Forest Regression (RFR). Therefore, model performance should be interpreted with consideration of potential overfitting risks. Spectral reflectance, vegetation indices, and their combinations served as predictor features, with performance quantified by the coefficient of determination (R^2) and root mean square error (RMSE) *(Xie, et al. 2025)*. Higher R^2 values indicate stronger explanatory power, while lower RMSE values reflect greater predictive accuracy. The formulations of R^2 and RMSE are provided in Equations (14) and (15).

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

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

where:

n denotes the sample size, y_i the observed value, $\hat{y}_i$ the predicted value, and $\bar{y}_i$ the sample mean.

RESULTS

Correlation analysis between spectral bands and moisture content

The relationships between soybean grain moisture content and spectral reflectance from four multispectral bands (Green, Red, RE, and NIR) were assessed using the Pearson correlation coefficient (PCC) *(Trentin, et al. 2025)*. RE and NIR exhibited the strongest negative correlations with moisture content, with coefficients of −0.836 and −0.781, respectively, whereas Green and Red showed moderate negative correlations of −0.731 and −0.662. All correlations were statistically significant at the 0.01 level, indicating that these spectral features reliably reflect variations in grain moisture.

Notably, the RE band demonstrated the highest sensitivity (r = −0.836), likely due to its position at the transition between the strong chlorophyll absorption region in the red spectrum and the high-reflectance NIR region, rendering it highly responsive to subtle physiological changes. NIR and Green also exhibited pronounced correlations, consistent with the influence of water content on cellular structures and mesophyll scattering. The Red band, although weaker, still captured indirect water-related effects via chlorophyll absorption.

Collectively, RE and NIR are primary indicators for estimating soybean grain moisture, with Green and Red providing complementary information. Incorporating multiple bands not only improves predictive robustness but also establishes a solid spectral basis for subsequent feature selection and modeling.

Correlation analysis between vegetation index and soybean moisture content

In quantitative remote sensing studies of vegetation water content, correlating grain moisture with vegetation indices (VIs) is a standard approach to identify spectral indicators sensitive to water dynamics. Here, a suite of multispectral VIs was evaluated against field-measured soybean grain moisture content at maturity using Pearson correlation analysis *(Shafiee, et al. 2023)*. The results are presented in Fig. 4.

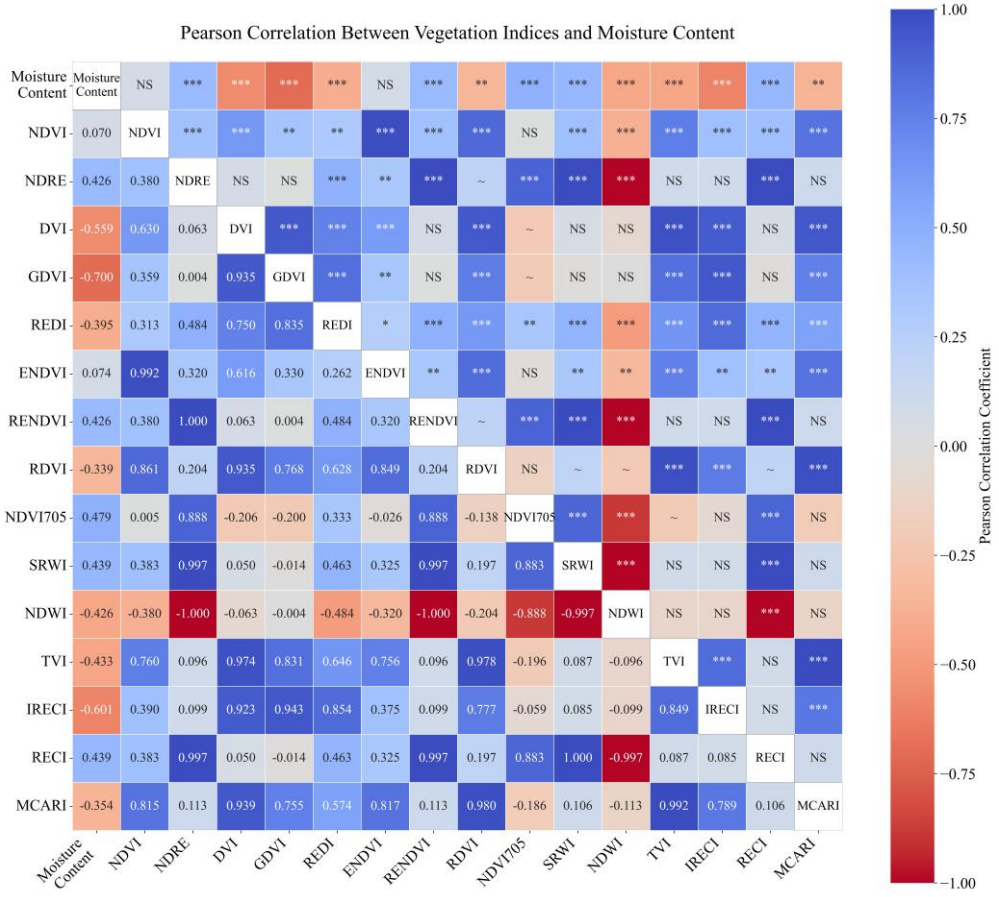


Fig. 4 - Pearson correlation coefficient matrix

****, **, *, ~, and NS indicate statistical significance at P < 0.001, P < 0.01, P < 0.05, and P < 0.10, and non-significance.*

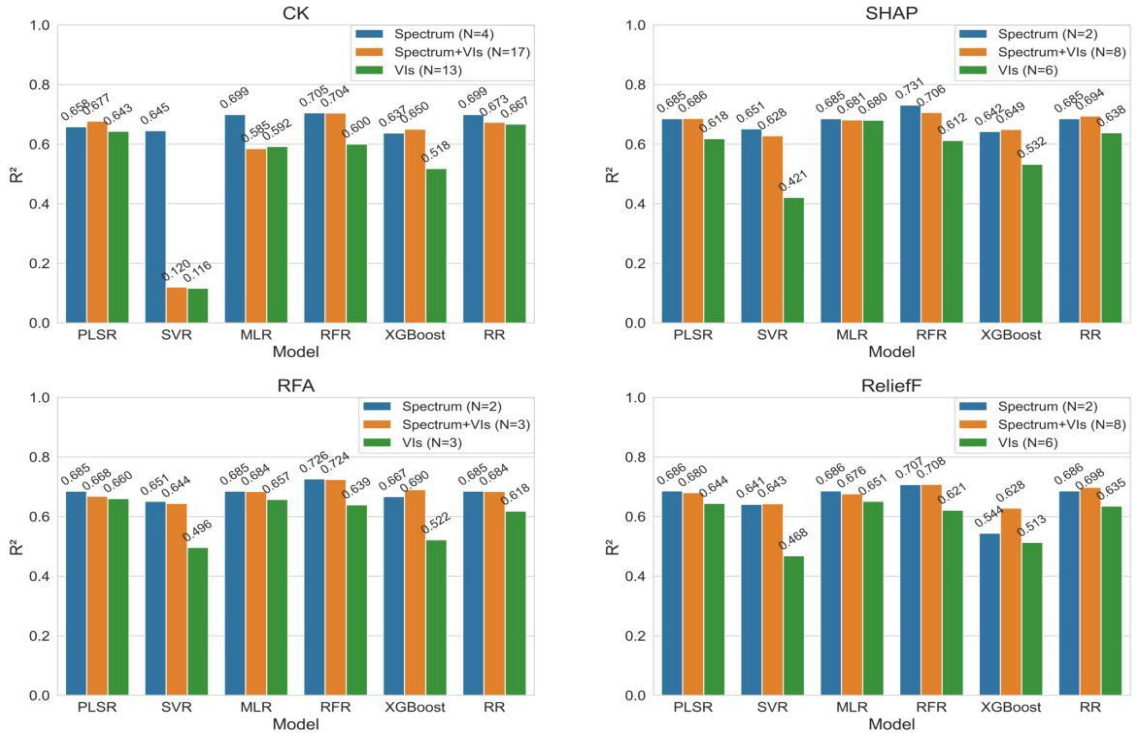
The analysis indicated that most VIs exhibited statistically significant correlations with grain moisture (P < 0.01), although the strength and direction of the relationships varied. Notably, GDVI and IRECI showed the highest negative correlations, with coefficients of −0.700 and −0.601, respectively, highlighting their strong responsiveness to moisture variation. In contrast, NDVI and ENDVI were not significantly correlated (P > 0.05), suggesting limited sensitivity at the mature stage. Mechanistically, GDVI (Green Difference Vegetation Index) integrates green and NIR bands, capturing dual spectral responses associated with leaf water content, thereby providing a comprehensive signal of moisture-induced spectral variation.

IRECI (Red Edge Chlorophyll Index) exploits the RE band's sensitivity to pigment concentration and leaf structural changes, which are often accompanied by water content variation. By contrast, conventional indices such as NDVI and ENDVI commonly suffer from saturation under high canopy coverage and late growth stages, reducing their utility in moisture monitoring.

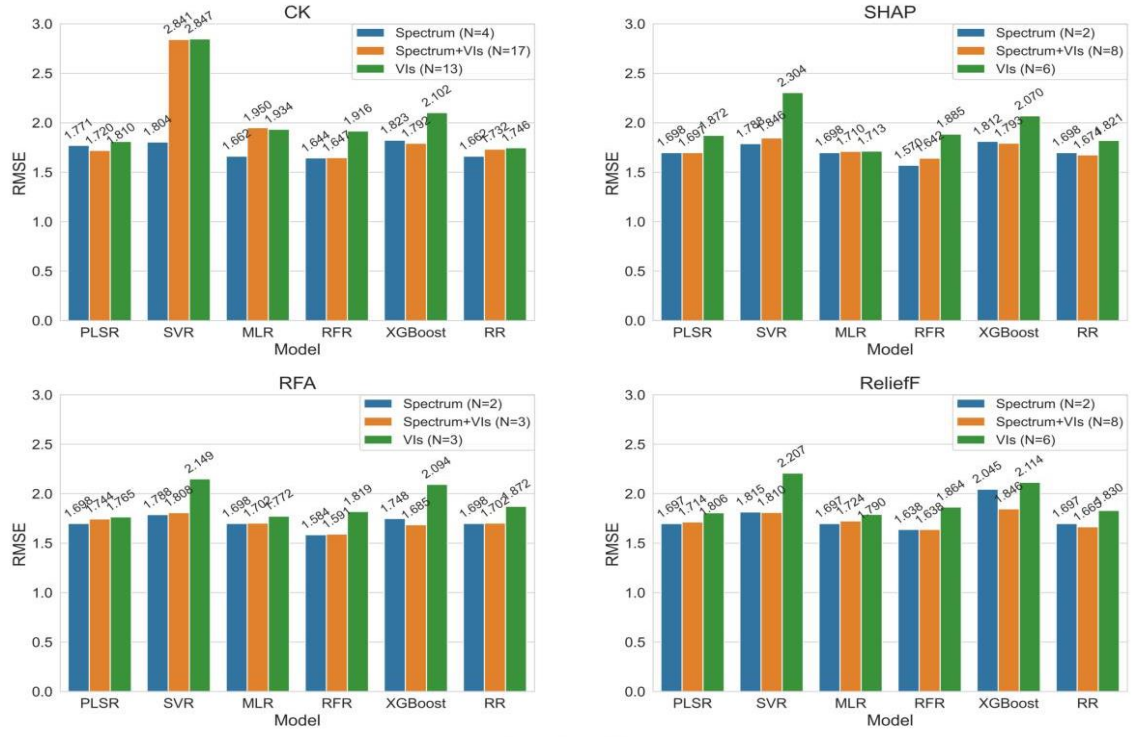
Overall, except for NDVI and ENDVI, the remaining indices effectively reflected grain moisture variation and were retained as candidate features for modeling. However, potential multicollinearity among VIs may compromise model stability; thus, subsequent feature selection is necessary to optimize the variable set and ensure robust predictive performance *(Shu, et al.)*.

Feature screening method analysis

To enhance the generalization ability and computational efficiency of the models, this study employed three feature selection methods—SHAP (SHapley Additive exPlanations) *(Feng, et al. 2025)*, RFA (Recursive Feature Addition) *(Li, et al. 2025)*, and ReliefF (Relief-based Feature Selection) *(Xu, et al. 2024)*—to extract spectral features.



(a) R²



(b) RMSE

Fig. 5 - Comparison of model accuracy of different feature screening methods

SHAP, derived from the Shapley value in game theory, quantifies each feature's average marginal contribution across all possible feature subsets, offering both global importance and local interpretability, thereby balancing stability and explainability. In contrast, RFA operates as a wrapper-based method that iteratively adds features while evaluating model performance to determine the optimal subset, whereas ReliefF functions as a filter-based algorithm that assigns feature weights based on differences among neighboring samples, effectively identifying variables highly correlated with the target.

The optimal features selected by SHAP, RFA, and ReliefF were used to construct predictive models, with performance evaluated using feature number (N), coefficient of determination (R^2), and root mean square error (RMSE), as shown in Fig. 5. The unselected feature set served as the control (CK). Overall, all three feature selection methods effectively reduced spectral dimensionality while maintaining or improving prediction accuracy, although performance varied with feature composition and regression algorithm.

When spectral reflectance alone was used, SHAP-selected features yielded the best performance for SVR and RFR, confirming SHAP's ability to identify informative bands with minimal redundancy. XGBoost achieved slightly higher accuracy with RFA-selected features, whereas ReliefF showed marginal advantages for PLSR. Notably, SHAP retained only two key spectral bands, reducing multicollinearity and enhancing model stability, which is particularly beneficial under limited sample conditions.

For combined spectral and vegetation index (VI) features, RFA consistently outperformed SHAP and ReliefF in PLSR, SVR, MLR, and RFR models, demonstrating strong adaptability to nonlinear and correlated variables. This advantage stems from RFA's ability to preserve complementary information while suppressing redundancy. ReliefF proved most effective for RR, whereas XGBoost performed best under the CK condition. This suggests that, for ensemble models such as XGBoost, excessive feature reduction may remove complementary predictors, and retaining the full feature set can be advantageous under small-sample conditions.

For vegetation indices alone, RFA provided the most consistent and robust results across models, surpassing SHAP and ReliefF in stability and accuracy. This indicates that RFA is particularly suitable for index-based features characterized by strong intercorrelations and nonlinear moisture responses.

Overall, RFA achieved the most favorable balance between dimensionality reduction and predictive accuracy, confirming its superior capacity to extract essential spectral information related to soybean grain moisture content. Selecting two optimal spectral bands via RFA thus represents the most effective and practical strategy for accurately estimating soybean grain moisture content at maturity.

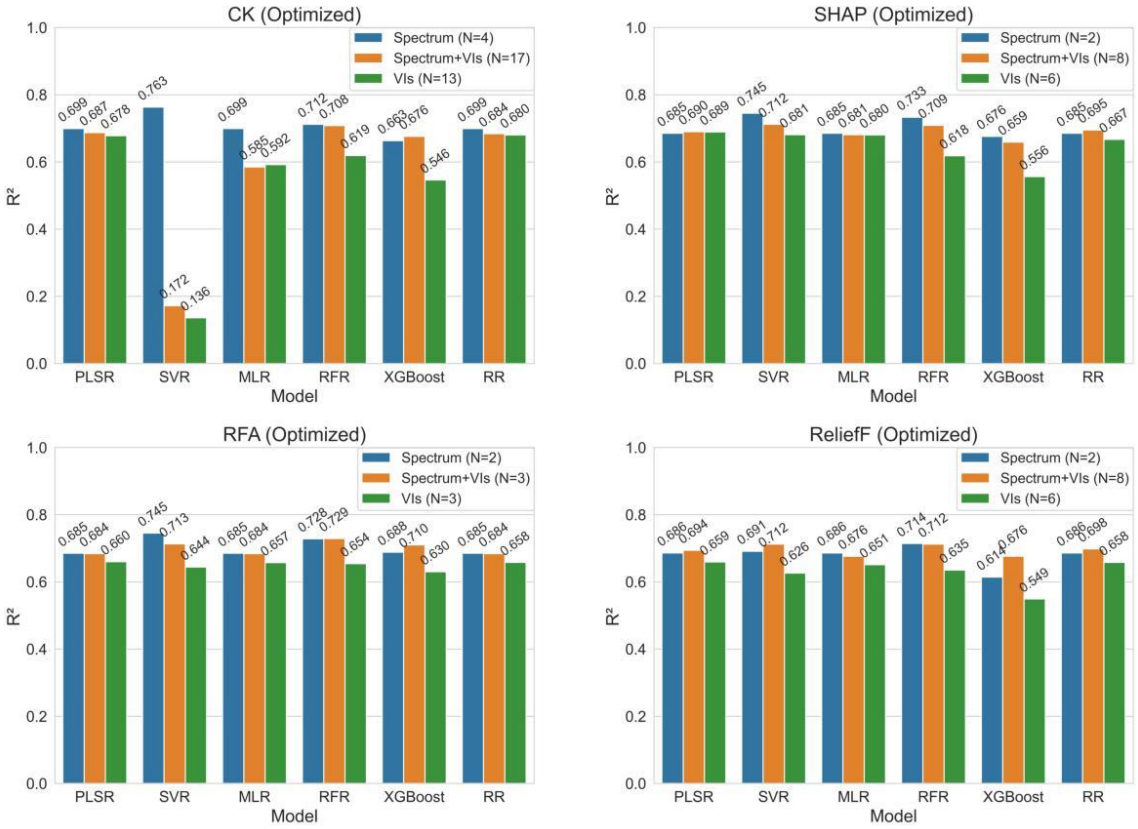
Model building

This study evaluated three feature sets—spectral reflectance, vegetation indices, and their combination—using six regression models (PLSR, SVR, MLR, RFR, XGBoost, RR) with grid search optimization for predicting soybean grain moisture content at maturity *(Ecke, et al. 2024)*, the parameters are shown in Table 4. Figure 6 presents the results of grid search optimization for the three feature combinations under the baseline condition without feature selection (CK) and following feature selection using SHAP, RFA, and ReliefF. The MLR model was not subjected to optimization.

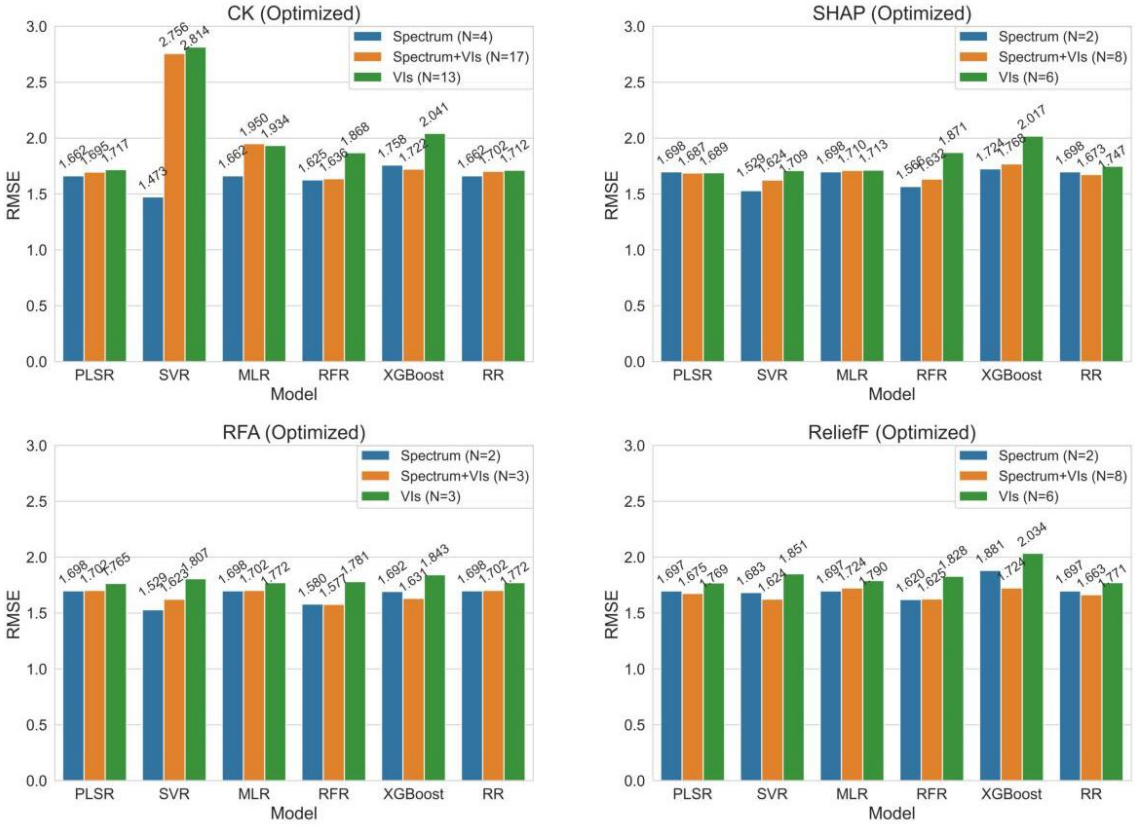
Table 4

Grid search results of main parameters of each model

Model	Parameter	Numeric Value
PLSR	n_components	4
SVR	C	10
	gamma	scale
RFR	max_depth	3
	n_estimators	100
XGBoost	learning_rate	0.05
	max_depth	3
	n_estimators	50
RR	alpha	100



(a) R²



(b) RMSE

Fig. 6 - Comparison of prediction accuracy of the models after optimization

Model comparison analysis

In the model optimization stage, grid search was applied to optimize parameters across different feature combinations and regression algorithms. Using spectral reflectance alone, the Support Vector Regression (SVR) model demonstrated the best performance after hyperparameter tuning, achieving an R^2 of 0.763 and RMSE of 1.473. Compared with the pre-optimized model, R^2 increased by 16.54% and RMSE decreased by 17.06%, indicating that SVR effectively captured complex nonlinear relationships between spectral reflectance and soybean grain moisture content. The optimized SVR also improved generalization capability under high-dimensional spectral conditions, even with a limited sample size.

The Random Forest Regression (RFR) model also performed well (R^2 = 0.711, RMSE = 1.628), confirming the robustness of ensemble learning methods in mitigating overfitting. For combined spectral and VI features, RFA-based models exhibited consistent performance improvement. The optimized RFR model achieved the highest accuracy (R^2 = 0.734, RMSE = 1.562), reflecting RFA's effectiveness in reducing redundancy and enhancing computational efficiency. However, incorporating vegetation indices slightly increased model complexity without yielding substantial accuracy gains relative to spectral-only models.

When using VIs alone, the SVR model based on ReliefF-selected features performed best (R^2 = 0.732, RMSE = 1.569), highlighting its capability to handle nonlinear and collinear predictors. Meanwhile, RFA-selected features improved the stability of linear models such as MLR (R^2 = 0.690, RMSE = 1.686). As summarized in Fig. 6, integrating feature selection with grid search optimization significantly enhanced both model accuracy and interpretability. Among all combinations, SVR based solely on spectral reflectance consistently achieved superior performance, indicating that sensitive bands such as RE and NIR provide sufficient information for reliable soybean grain moisture estimation. The fitted results (Fig. 7) further validate the robustness of the optimized models, with predicted values closely aligned to the 1:1 line.

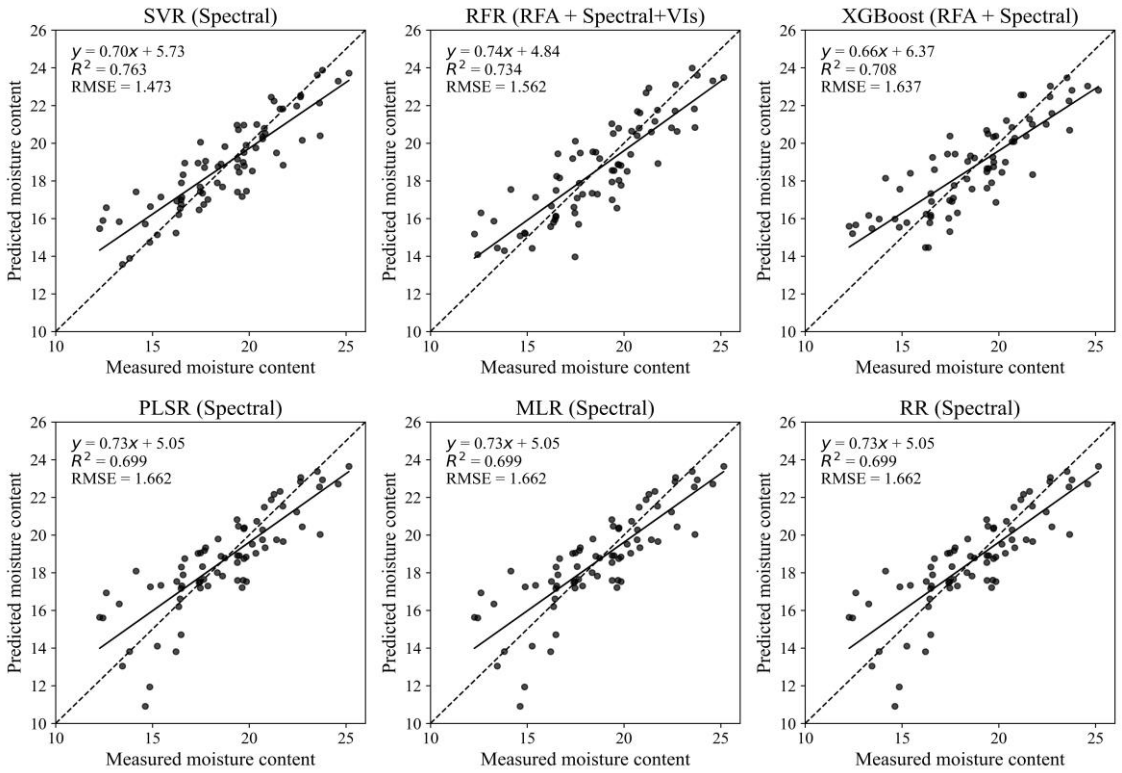


Fig. 7 - Comparison of measured and predicted MC values of soybean grains at maturity

CONCLUSIONS

This study employed a DJI Mavic 3M multispectral UAV to investigate soybean grain moisture content (MC) at maturity, combining spectral reflectance, vegetation indices, and machine learning models. Raw multispectral bands, particularly RE and NIR, exhibited the strongest correlations with MC (r = −0.836 and −0.781), while Green and Red also showed notable associations (r = −0.731 and −0.662). Among vegetation indices, GDVI and IRECI achieved the highest correlations (r = −0.700 and −0.601), whereas traditional NDVI was largely saturated in dense canopies, and NDRE partially mitigated this limitation (r = 0.4262). ENDVI showed low applicability under the experimental conditions.

Models based on raw spectral reflectance outperformed those using vegetation indices alone. SVR captured complex nonlinear relationships effectively (R^2 = 0.763, RMSE = 1.473), whereas RFR handled multicollinearity in combined features through latent variable extraction (R^2 = 0.734, RMSE = 1.562). Feature selection methods—SHAP, RFA, and ReliefF—successfully reduced redundancy and multicollinearity, with RFA maintaining high predictive accuracy while minimizing model complexity.

It should be noted that the relatively small sample size (n = 73) imposes inherent limitations on training complex models such as XGBoost and RFR. Although five-fold cross-validation was employed to alleviate overfitting, model generalizability may still be constrained. This limitation underscores the importance of cautious interpretation when extrapolating results beyond the experimental conditions. Nevertheless, the consistent performance trends observed across feature selection methods indicate that the identified spectral–moisture relationships are structurally stable, even under limited sample conditions.

From an application perspective, the proposed framework offers a rapid, non-destructive, and cost-effective solution for estimating soybean grain moisture content, supporting timely harvest decision-making and precision agricultural management. By enabling field-scale moisture assessment, this approach has practical implications for reducing harvest losses and improving operational efficiency. Future research should focus on expanding sample size, validating model robustness across diverse environments, and integrating multi-source data and advanced learning strategies to further enhance prediction scalability and reliability.

ACKNOWLEDGEMENT

This research was funded by the National Key Research and Development Plan project (2021YFD2000503), the National Natural Science Foundation of China (No. 32171911), the National Soybean Industry Technology System (CARS-04), the Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-SAE-202301).

REFERENCES

[1] Alabi T. R., Abebe A. T., Chigeza G., & Fowobaje K. R. (2022). Estimation of soybean grain yield from multispectral high-resolution UAV data with machine learning models in West Africa. *Remote Sensing Applications: Society and Environment*, 27: 100782. https://doi.org/10.1016/j.rsase.2022.100782.

[2] Barbosa Júnior M. R., Sales L. d. A., Santos R. G. d., Vargas R. B. S., Tyson C., & Oliveira L. P. d. (2025). Forecasting yield and market classes of Vidalia sweet onions: A UAV-based multispectral and texture data-driven approach. *Smart Agricultural Technology*, 10: 100808. https://doi.org/10.1016/j.atech.2025.100808.

[3] Bin W., Fan L., Xu B., Yang J., Zhao R., Wang Q., Ai X., Zhao H., & Yang Z. (2025). UAV-based LiDAR and multispectral sensors fusion for cotton yield estimation: Plant height and leaf chlorophyll content as a bridge linking remote sensing data to yield. *Industrial Crops and Products*, 230: 121110. https://doi.org/10.1016/j.indcrop.2025.121110.

[4] Burnett A. C., Anderson J., Davidson K. J., Ely K. S., Lamour J., Li Q., Morrison B. D., Yang D., Rogers A., & Serbin S. P. (2021). A best-practice guide to predicting plant traits from leaf-level hyperspectral data using partial least squares regression. *Journal of Experimental Botany*, 72(18): 6175-6189. https://doi.org/10.1093/jxb/erab295.

[5] Ecke S., Stehr F., Dempewolf J., Frey J., Klemmt H.-J., Seifert T., & Tiede D. (2024). Species-specific machine learning models for UAV-based forest health monitoring: Revealing the importance of the BNDVI. *International Journal of Applied Earth Observation and Geoinformation*, 135: 104257. https://doi.org/10.1016/j.jag.2024.104257.

[6] Feng Z., Yang Z., Suo L., Wu K., Ma Z., Zhang H., Duan J., & Feng W. (2025). Enhancing maize above-ground biomass estimation through multispectral, digital and LiDAR fusion on UAV platforms. *Agricultural Water Management*, 315: 109551. https://doi.org/10.1016/j.agwat.2025.109551.

[7] Imani M., & Ghassemian H. (2015). Ridge regression-based feature extraction for hyperspectral data. *International Journal of Remote Sensing*, 36(6): 1728-1742. https://doi.org/10.1080/01431161.2015.1024894.

[8] Khose S. B., & Mailapalli D. R. (2024). Spatial mapping of soil moisture content using very-high resolution UAV-based multispectral image analytics. *Smart Agricultural Technology*, 8: 100467. https://doi.org/10.1016/j.atech.2024.100467.

[9] Kokaly R. F., & Clark R. N. (1999). Spectroscopic Determination of Leaf Biochemistry Using Band-Depth Analysis of Absorption Features and Stepwise Multiple Linear Regression. *Remote Sensing of Environment*, 67(3): 267-287. https://doi.org/10.1016/S0034-4257(98)00084-4.

[10] Li J., Xue F., Li G., Zhang M., Tian J., & Zhang H. (2025). SPRC: A novel Sentinel-1/-2 Phenology-enhanced framework for automated paddy rice mapping. *International Journal of Applied Earth Observation and Geoinformation*, 143: 104772. https://doi.org/10.1016/j.jag.2025.104772.

[11] Liu Y., Fan K., Meng L., Nie C., Liu Y., Cheng M., Song Y., & Jin X. (2025). Synergistic use of stay-green traits and UAV multispectral information in improving maize yield estimation with the random forest regression algorithm. *Computers and Electronics in Agriculture*, 229: 109724. https://doi.org/10.1016/j.compag.2024.109724.

[12] Qi Z., Feng Y., Wang S., & Li C. (2025). Enhancing hydropower generation Predictions: A comprehensive study of XGBoost and Support Vector Regression models with advanced optimization techniques. *Ain Shams Engineering Journal*, 16(1): 103206. https://doi.org/10.1016/j.asej.2024.103206.

[13] Ren Y., Zhang W., Wang H., Zhang Z., Sheng W., Qiu R., & Zhang M. (2025). Estimation models for maize leaf water content at various stages using near-infrared spectroscopy. *Infrared Physics & Technology*, 145: 105732. https://doi.org/10.1016/j.infrared.2025.105732.

[14] Roma E., Orlando S., Carella A., Lo Bianco R., Massenti R., & Catania P. (2025). Fraction cover estimation using drone-based multispectral images in six olive cultivars and different planting systems: a case study in Sicily. *Smart Agricultural Technology*, 12: 101323. https://doi.org/10.1016/j.atech.2025.101323.

[15] Shafiee S., Mroz T., Burud I., & Lillemo M. (2023). Evaluation of UAV multispectral cameras for yield and biomass prediction in wheat under different sun elevation angles and phenological stages. *Computers and Electronics in Agriculture*, 210: 107874. https://doi.org/10.1016/j.compag.2023.107874.

[16] Shi W., Li Y., Zhang W., Yu C., Zhao C., & Qiu J. (2024). Monitoring and zoning soybean maturity using UAV remote sensing. *Industrial Crops and Products*, 222: 119470. https://doi.org/10.1016/j.indcrop.2024.119470.

[17] Shu M., Ge Z., Li Y., Yue J., Guo W., Fu Y., Dong P., Qiao H., & Gu X. (2025). A novel canopy water indicator for UAV imaging to monitor winter wheat water status. *Smart Agricultural Technology*, 12: 101160. https://doi.org/10.1016/j.atech.2025.101160.

[18] Shu M., Shen M., Zuo J., Yin P., Wang M., Xie Z., Tang J., Wang R., Li B., Yang X., & Ma Y. The Application of UAV-Based Hyperspectral Imaging to Estimate Crop Traits in Maize Inbred Lines. *Plant Phenomics*, 2021. 10.34133/2021/9890745.

[19] Sim C. H., Gan F. F., & Chang T. C. (2005). Outlier Labeling With Boxplot Procedures. *Journal of the American Statistical Association*, 100(470): 642-652. https://doi.org/10.1198/016214504000001466.

[20] Sun H., Zhang L., Rao Z., & Ji H. (2020). Determination of moisture content in barley seeds based on hyperspectral imaging technology. *Spectroscopy Letters*, 53(10): 751-762. https://doi.org/10.1080/00387010.2020.1832531.

[21] Sun J., Shu S., Hu H., Deng Y., Li Z., Zhou S., Liu Y., Dang M., Huang W., Hou Z., Yin X., Zhang R., Yang C., Jing W., Yang J., & Zhou C. (2025). Location optimization of unmanned aerial vehicle (UAV) drone port for coastal zone management: The case of Guangdong coastal zone in China. *Ocean & Coastal Management*, 262: 107576. https://doi.org/10.1016/j.ocecoaman.2025.107576.

[22] Trentin C., Ampatzidis Y., Tasioulas S., & Tsouvaltzis P. (2025). Optimizing tomato yield prediction using phenologically timed UAV-based spectral data and machine learning. *Smart Agricultural Technology*, 12: 101158. https://doi.org/10.1016/j.atech.2025.101158.

[23] Tulu B. B., Teshome F., Ampatzidis Y., Hailegnaw N. S., & Bayabil H. K. (2025). AgriSenAI: Automating UAV thermal and multispectral image processing for precision agriculture. *SoftwareX*, 30: 102083. https://doi.org/10.1016/j.softx.2025.102083.

[24] Wang C., Zhang X., Zhang N., Guo H., Wu H., & Wang X. (2025). Optimizing the estimation of cotton leaf SPAD and LAI values via UAV multispectral imagery and LASSO regression. *Smart Agricultural Technology*, 12: 101098. https://doi.org/10.1016/j.atech.2025.101098.

[25] Wang N., Guo Y., Wei X., Zhou M., Wang H., & Bai Y. (2022). UAV-based remote sensing using visible and multispectral indices for the estimation of vegetation cover in an oasis of a desert. *Ecological Indicators*, 141: 109155. https://doi.org/10.1016/j.ecolind.2022.109155.

[26] Wang Y., Kootstra G., Yang Z., & Khan H. A. (2024). UAV multispectral remote sensing for agriculture: A comparative study of radiometric correction methods under varying illumination conditions. *Biosystems Engineering*, 248: 240-254. https://doi.org/10.1016/j.biosystemseng.2024.11.005.

[27] Xie S., Wang X., Zhu X., & Li Y. (2025). Prediction of soil organic matter content in winter wheat jointing stage based on UAV multispectral and machine learning. *Measurement*, 256: 118508. https://doi.org/10.1016/j.measurement.2025.118508.

[28] Xu T., Wang F., Shi Z., & Miao Y. (2024). Multi-scale monitoring of rice aboveground biomass by combining spectral and textural information from UAV hyperspectral images. *International Journal of Applied Earth Observation and Geoinformation*, 127: 103655. https://doi.org/10.1016/j.jag.2024.103655.

[29] Yakubu A. U., Xiong S., Jiang Q., Zhao J., Wu Z., Wang H., Ye X., & Wangsen H. (2024). Fuzzy-based thermal management control analysis of vehicle air conditioning system. *International Journal of Hydrogen Energy*, 77: 834-843. https://doi.org/10.1016/j.ijhydene.2024.06.030.

[30] Yang M.-D., Hsu Y.-C., Tseng W.-C., Tseng H.-H., & Lai M.-H. (2025a). Precision assessment of rice grain moisture content using UAV multispectral imagery and machine learning. *Computers and Electronics in Agriculture*, 230: 109813. https://doi.org/10.1016/j.compag.2024.109813.

[31] Yang N., Zhang Z., Yang X., Dong N., Xu Q., Chen J., Sun S., Cui N., & Ning J. (2025b). Evaluation of crop water status using UAV-based images data with a model updating strategy. *Agricultural Water Management*, 312: 109445. https://doi.org/10.1016/j.agwat.2025.109445.