Received: 17 March 2025 | Accepted: 13 July 2025

DOI: 10.1002/ppj2.70037

The Plant Phenome Journal OPEN ACCESS

ORIGINAL ARTICLE

Enhancing yield and protein content estimation in dry peas through multispectral data integration and ensemble learning model

Aliasghar Bazrafkan[1,2] | **Hannah Worral**[3] | **Nonoy Bandillo**[4] | **Paulo Flores**[1]

[1]Department of Agricultural and Biosystems Engineering, North Dakota State University, Fargo, North Dakota, USA

[2]Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, Michigan, USA

[3]Department of Plant Science, North Dakota State University, Fargo, North Dakota, USA

[4]Department of Crop and Soil Sciences, North Carolina State University, Raleigh, North Carolina, USA

Correspondence
Paulo Flores, Department of Agricultural and Biosystems Engineering, North Dakota State University, Fargo, ND 58102, USA.
Email: paulo.flores@ndsu.edu

Assigned to Associate Editor Sara Tirado Tolosa.

Funding information
North Dakota Department of Agriculture, Grant/Award Numbers: 21-316, 20-489; USDA-NIFA, Grant/Award Number: ND01488 ND01513; the U.S. Department of Agriculture, Agricultural Research Service, Grant/Award Number: 58-6064-8-023

Abstract

Accurately predicting yield and protein content in dry peas (*Pisum sativum* L.) is essential for improving agricultural productivity, sustainability, and nutritional outcomes. Traditional methods for estimating these traits are labor-intensive and limited in scope, while remote sensing technologies and machine learning offer scalable, nondestructive alternatives. This study addresses the challenge of integrating multispectral data and advanced machine learning models, specifically stacked meta-models, to enhance prediction accuracy for yield and protein content in dry peas. Data were collected from 860 genotypes across two North Dakota locations over three growing seasons using uncrewed aerial system-mounted multispectral sensors. The results revealed significant variability in yield and protein content due to location and year, with the meta-model outperforming individual machine learning models for both traits. The meta-model achieved a root mean squared error (RMSE) of 396.28 kg ha^{-1} and an coefficient of determination (R^2) of 0.77 for yield, demonstrating strong predictive performance. However, protein content predictions were moderate, with an RMSE of 1.28% and an R^2 of 0.54, highlighting limitations in the correlation between spectral features and protein synthesis. Growth stage analysis emphasized the critical importance of the maturity stages to predict the yield and protein content in dry peas. These findings support the potential of integrating multispectral data and meta-modeling for accurate agricultural predictions. Future research should focus on incorporating additional data sources, refining feature engineering techniques, and tailoring models to specific growth stages to further improve predictive accuracy. This study provides a foundation for advancing precision agriculture practices in pulse crops.

Abbreviations: DT, decision tree; GB, gradient boosting; KNN, *K*-nearest neighbors; LGBM, light gradient boosting machine; MAE, mean absolute error; MLP, multilayer perceptron; PCA, principal component analysis; RF, random forest; RMSE, root mean squared error; UAS, uncrewed aerial system; VI, vegetation index.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2025 The Author(s). *The Plant Phenome Journal* published by Wiley Periodicals LLC on behalf of American Society of Agronomy and Crop Science Society of America.

Plain Language Summary

Predicting crop yield and protein content is essential for improving food production and sustainability. Traditional methods require time-consuming manual sampling, but modern technology, like satellite and drone-based remote sensing combined with artificial intelligence, can make this process faster and more accurate. This study used machine learning models to analyze multispectral images collected from dry pea fields in North Dakota. The goal was to improve yield and protein content predictions by integrating multiple models into a single, more accurate "meta-model." Results showed that this approach significantly improved prediction accuracy compared to individual models. While predicting yield was highly successful, predicting protein content was more challenging due to weaker correlations with spectral data. The study highlights the importance of selecting the right data sources and modeling approaches to enhance agricultural predictions.

1 | INTRODUCTION

Studying the yield and protein content of dry peas (*Pisum sativum* L.) is critical due to their nutritional value, agricultural benefits, and economic importance. Dry peas provide a major source of plant-based protein, ranging from 17% to 30%, which is essential for food security, particularly in regions with vegetarian diets (Boye et al., 2010; Thavarajah et al., 2023). Their functional properties also make dry peas valuable in food applications, such as in bakery products and snacks (Shanthakumar et al., 2022; X. Zhu, Li, et al., 2023). Additionally, dry peas contribute to sustainable agriculture by fixing atmospheric nitrogen, reducing the need for chemical fertilizers, and improving soil health (K. Liu et al., 2020). Optimizing dry peas yields can enhance economic returns for farmers while lowering the environmental impact of agriculture through reduced greenhouse gas emissions (L. Liu et al., 2021). Thus, studying pulse crops including dry peas supports better nutrition, sustainability, and economic resilience. To fully harness these benefits, accurate methods for estimating yield and protein content are essential.

Traditional and remote sensing methods for estimating yield and protein content in plants differ in process, efficiency, and accuracy. Traditional methods involve manual sampling, which is labor-intensive, time-consuming, and destructive. Though precise, they rely on limited samples, may not represent entire fields, and are costly due to the need for specialized equipment and manpower (Yunlu et al., 2005). In contrast, remote sensing methods use nondestructive technologies, like hyperspectral imaging, LiDAR, and multispectral cameras to assess crops over large areas efficiently (Z. Li, Wang, et al., 2015). Remote sensing allows for real-time data collection, reducing time and costs while enabling frequent monitoring throughout the growing season (B. Zhu et al., 2023). Additionally, it can be integrated with models like decision support systems and machine learning to estimate yield and protein content with high precision using biophysical and spectral variables (Z. Li, Jin, et al., 2015; Zhou et al., 2021). While remote sensing methods provide broader coverage and efficiency, integrating machine learning with uncrewed aerial system (UAS) remote sensing offers even greater accuracy and predictive capabilities.

Machine learning models combined with UAS remote sensing greatly enhance the accuracy of yield and protein content estimation in plants (Yang et al., 2024). UASs with multispectral, hyperspectral, and thermal sensors capture high-resolution data on plant health and canopy structure, which are key for yield predictions (Zhou et al., 2021). Machine learning algorithms process features like vegetation indices (VIs) and structural data, outperforming traditional methods, as seen in soybean yield predictions with random forest (RF) models (Ren et al., 2023). These models handle complex, nonlinear relationships, improving prediction accuracy across growth stages (Zha et al., 2020). Data fusion from UASs, combining structural and spectral data, further boosts prediction accuracy in crops like maize and ryegrass (Pranga et al., 2021), offering scalable, noninvasive solutions for agriculture. Although machine learning models combined with UAS remote sensing have greatly enhanced yield and protein content estimation, there are still several research gaps that need to be addressed to further improve accuracy and applicability.

Research on estimating plant yield and protein content has several gaps that need to be addressed. First, the influence of growth stages on prediction accuracy, especially for protein content, remains underexplored. This gap aligns with findings from the literature (Astaoui et al., 2021; C. Zhang et al., 2024) and highlights the need for feature selection analyses tailored to specific growth stages, including vegetative, flowering, maturity, and harvest in dry peas studies. Additionally,

a review of the current literature reveals significant gaps in parameter tuning for machine learning models based on multispectral information, particularly for dry pea yield and protein content estimation. While machine learning has demonstrated potential in agricultural predictions (Zhou et al., 2021), there is a lack of focused studies exploring the optimization of parameters across a diverse range of models using spectral data in dry peas. Moreover, machine learning advancements, particularly deep and ensemble learning, remain underutilized for predicting multiple traits simultaneously. Stacking models, which could enhance predictions by combining base models and meta-models, are lacking in research on pulse crops. Furthermore, most models lack region-specific calibration and are not optimized for specific crops or environments, such as dry peas, limiting their generalizability (Sharifi, 2021). Finally, while yield prediction research is prevalent, there is insufficient focus on protein content estimation, a key nutritional trait that requires more attention (B. Zhu, Chen, et al., 2023). Addressing these gaps, particularly through multispectral data integration and advanced machine learning models, can significantly advance the field.

In response to these challenges, the main goal of this study is to propose a new approach for accurately predicting yield and protein content in dry peas by integrating multispectral information with machine learning models. What makes our approach unique is the idea of using predictions from multiple models as additional features in subsequent rounds of prediction, particularly for protein content and yield. This approach, known as ensemble learning, combines the strengths of multiple models to improve overall predictive accuracy and reduce errors (Seni & Elder, 2010). Specifically, we use a stacking technique where predictions from several base models are input into a meta-model, allowing more robust and generalized predictions. This concept is similar to stacking in ensemble learning, where predictions from base models are used as inputs for a meta-model. As secondary goals, we aim to identify the most important spectral features and growth stage influencing the yield and protein content estimation in dry peas. Finally, we aim to determine the optimal configurations for machine learning models based on the dataset of dry peas.

2 | MATERIALS AND METHODS

Data acquisition, dataset creation and cleaning, feature selection, parameter tuning, training various machine learning models, developing a stacked meta-model, and model evaluation were the key steps of this study. Figure 1 illustrates the workflow for estimating yield and protein content in dry pea plots.

2.1 | Study area and data acquisition

To investigate the effectiveness of using UAS-mounted sensors to estimate yield and protein content in dry peas, image datasets were gathered over three growing seasons (2022–2024) at two locations across North Dakota (Carrington—489548 m E, 5263199 m N; and Minot—328694 m E, 5339621 m N). There were 330 plots at each location, though plot dimensions varied slightly between sites (1.22 m × 2.74 m in Carrington and 1.52 m × 3.35 m in Minot). These field experiments were conducted as part of North Dakota State University's pulse crops breeding program. Over the 3-year period, across both environments, a total of 860 distinct genotypes were evaluated, distributed across 793 plots per season. The discrepancy between the total number of plots and the expected 660 plots per season (330 per location) is due to some missing plots or data exclusions resulting from field conditions, management decisions, or data quality issues.

After harvest, seeds underwent a purification process to remove visually defective seeds, such as splits, diseased, or off-type seeds. Germination tests were conducted, and seed quantities for each test location were calculated based on germination rates, seed weights, and plot sizes. Prior to planting, seeds were treated with fungicide (Obvius Plus, BASF Corporation) to prevent disease and ensure uniform emergence. Finally, seeds were packaged individually and prepared for distribution to testing sites.

2.2 | Aerial data collection and image preprocessing

Image collection for this study was conducted in both environments (Carrington and Minot) at early-pea (May to June), mid-pea (July to August), and late-pea (August to September) growth stages. The image data were acquired using the RedEdge-MX dual camera system (MicaSense), which was mounted on a DJI Matrice 200 (DJI Inc.) UAS. The DJI Matrice 200 (M200), an autonomous aircraft, was controlled using DJI Pilot version 2.5.1.15 for flight planning and execution. To reduce the impact of shadows cast by the plants, data collection flights were scheduled around solar noon. The RedEdge-MX dual camera system includes two multispectral cameras—MSblue and MSred, which are held next to each other using a mounting bracket—that capture images across 10 spectral bands ranging from 444 to 840 nm. Additionally, the system features a downwelling light sensor, which measures ambient light and allows for adjustments to be made to account for changes in lighting conditions. All flights were carried out at an altitude of 50 m, with 80% side and forward overlaps, and a flight speed of 5 m s^{-1}. The resulting imagery had an average ground sample distance of 3 cm per pixel.

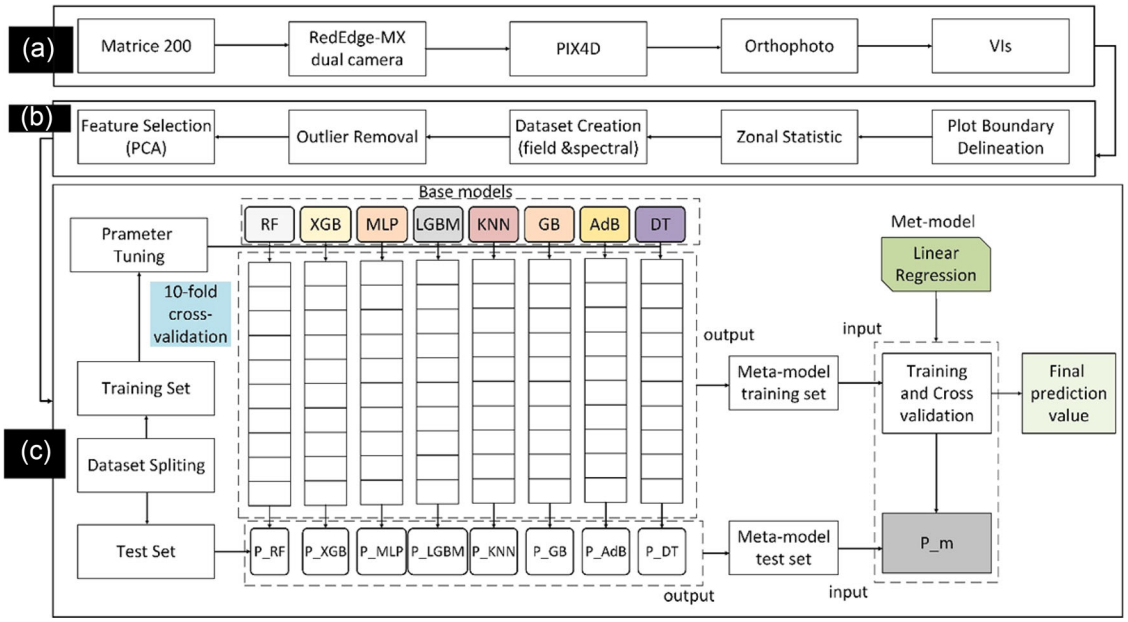


FIGURE 1 Workflow of estimating yield and protein contents in dry pea plots. (a) Data acquisition; (b) dataset creation, cleaning, and feature selection; and (c) modeling and evaluation. AdB, AdaBoost; DT, decision tree; GB, gradient boosting; KNN, *K*-nearest neighbors; LightGBM, light gradient boosting machine; MLP, multilayer perceptron; P_AdB, prediction from AdaBoost model; P_DT, prediction from decision tree model; P_GB, prediction from gradient boosting model; P_KNN, prediction from KNN model; P_LGBM, prediction from LightGBM model; P_m, final prediction from meta-model; P_MLP, prediction from multilayer perceptron model; P_RF, prediction from random forest model; P_XGB, prediction from XGBoost model; PCA, principal component analysis; RF, random forest; VI, vegetation index; XGB, extreme gradient boosting.

During the growth cycle of dry peas, UAS flights were conducted to collect multispectral data at key physiological stages to monitor development. The first flights occurred during the vegetative stage (3–6 weeks after sowing, early June–late June) when there is rapid expansion of stem and leaves, and when the root system begins to fixate nitrogen. The second flights, during the flowering stage (6–8 weeks after sowing, early July to mid-July), documented the reproductive phase when self-pollinating flowers emerge, marking a critical transition to seed production. A third round of flights, during the maturity stage (12–15 weeks after sowing, mid-August to early September), captured the final seed development and pod maturation stages. Finally, flights during the senescence and harvest stage (15+ weeks after sowing, early September) recorded the drying of plants, signaling readiness for harvest. These strategically timed flights allowed for comprehensive monitoring of pea growth and development.

2.3 | Data preprocessing and image processing

Radiometric calibration was performed using a manufacturer-provided calibrated reflective panel (version 2.34.56). This calibration process occurred during the generation of orthorectified products and was essential for calculating reflectance values for each spectral band. The calibration was automatically applied by the Pix4Mapper software (version 4.8.1) from Pix4D. Reflectance data from these bands were then used to compute various VIs (Table 1), which were subsequently analyzed for each flight date. The selection of VIs was guided by their application in prior studies on yield and protein content in various crops (Fei et al., 2021; Sarkar et al., 2024). A semiautomated method in QGIS (QGIS 3.26; QGIS Development Team) was employed for plot-level analysis, which involved delineating plot boundaries based on the red–green–blue (color model) mosaic generated by the RedEdge-MX sensor. This method utilized the coordinates of the lower-left corner of the plots, along with their dimensions and the vertical and horizontal spacing between them, to automatically outline the boundaries of each plot. Each plot was then assigned a unique ID for analysis across all imagery and flight dates.

2.4 | Data preparation and cleaning

The dataset used in this study (793 rows of data) contained various predictor variables such as spectral bands and VIs (a total of 172 features), with the yield and protein content as the response variables. Initially, the dataset was loaded using *Pandas* package (Bernard, 2016) in Python (version 3.9.11) to facilitate further analysis. The predictor variables and response (yield and protein content) were separated into

TABLE 1 List of vegetation indices (VIs) calculated from the multispectral imagery collected over the pea trials.

VI	Name	Equation	Reference
CIgreen	Green chlorophyll index	$\frac{NIR}{green} - 1$	(Burns et al., 2022)
ExG	Excess green index	$(2 \times *g \times *r - *b)$	(Jiang et al., 2019)
GLI	Green leaf index	$\frac{(2 \times green - red - blue)}{(2 \times green + red + blue)}$	(Eng et al., 2019)
GNDVI	Green normalized vegetation index	$\frac{(NIR - green)}{(NIR + green)}$	(Jiang et al., 2019)
GRRI	Green–red ratio index	$\frac{green}{red}$	(Lu et al., 2021)
GSAVI	Generalized soil-adjusted vegetation index	$\frac{(NIR - red)}{(NIR + red - green - blue)}$	(J. Zhang et al., 2021)
NDRE	Normalized difference red-edge index	$\frac{(NIR-RE)}{(NIR+RE)}$	(Large, 1954)
NDVI	Normalized difference vegetation index	$\frac{(NIR - red)}{(NIR + red)}$	(Shafian et al., 2018; Stanton et al., 2017; J. Zhang et al., 2021)
$NDVI_{RE}$	Red-edge normalized difference vegetation index	$\frac{(NIR - RE)}{(NIR + RE)}$	(Dong et al., 2019; Xie et al., 2018)
SAVI	Soil adjusted vegetation index	$1.5 \times \frac{(NIR - red)}{(NIR + red + 0.5)}$	(Candiago et al., 2015; Stanton et al., 2017)
SRred	Simple ratio red-edge index	$\frac{RE}{Red}$	(Sims & Gamon, 2002)
VARI	Visible atmospherically resistant index	$\frac{(green - red)}{(green + red - blue)}$	(Eng et al., 2019; Jiang et al., 2019)

Note: $*g = (green/(red + green + blue))$, $*r = (red/(red + green + blue))$, $*b = (blue/(red + green + blue))$.
Abbreviations: NIR, near infrared; RE, red-edge.

X (predictors) and *y* (response), respectively. To ensure data quality, outliers were identified and removed using the *Z-score* (Abdi, 2007) method. Each feature (including both predictors and the response) was standardized to calculate the *Z-score*. A *Z*-score threshold of 2.7 was used, following conventional practice to exclude extreme outliers beyond approximately the 99th percentile. This value was chosen to minimize the influence of anomalies while preserving the natural variability present in agricultural field data. These rows were subsequently removed from the dataset, and the cleaned data were saved for further analysis. This process ensured the removal of extreme values that could have potentially skewed model performance. To account for variability and improve the model's generalizability, a model was trained incorporating location and year as input features. This approach was applied only if a statistically significant difference (p-value < 0.05) was detected in yield and protein content across different years and locations. This process helps ensure the creation of a unified dataset without concerns about variability in yield and protein content across different environments and periods.

The cleaned dataset (680 rows) was split into 462 rows for training, 82 rows for testing, and 136 rows for validation. A random state of 42 was used to ensure reproducibility. The training set was used to train the models, while the test set was used to evaluate model performance. The validation dataset acted as an unseen dataset to test the reliability of the selected model. A variety of machine learning models were employed to predict yield based on the predictor variables. The models included the following: multilayer perceptron (MLP) regressor (Murtagh, 1991), RF regressor (Segal, 2004), XGBoost (Chen & Guestrin, 2016), *K*-nearest neighbors (KNN) (Dhanabal & Chandramathi, 2011), light gradient boosting machine (LightGBM) (Fan et al., 2019), gradient boosting (GB) regressor (Natekin & Knoll, 2013), AdaBoost regressor (Solomatine & Shrestha, 2004), and decision tree (DT) regressor (Xu et al., 2005). The reason for selecting these models is that they have been widely used in studies related to yield (Fei et al., 2021; Q.-C. Li, Xu, et al., 2023; Parganiha & Verma, 2024; Y. Zhang et al., 2022) and protein content (W. Li, Wu, et al., 2023; Wang et al., 2021; Zhou et al., 2021) estimations.

2.5 | Feature selection

In the feature selection process, principal component analysis (PCA) (Abdi & Williams, 2010) was applied to determine the contribution of each predictor to the total explained variance. While PCA was used here to reduce dimensionality and assess variance contributions, we acknowledge that model-based feature importance methods (e.g., random forest) provide supervised insights into feature relevance for

prediction and may offer complementary advantages. Future studies could integrate both approaches to enhance feature selection strategies. Initially, the feature set was standardized using *StandardScaler* from the *scikit-learn* package (version 1.0.2) in Python, as PCA performs optimally with standardized data. PCA was then fitted to the standardized feature matrix without specifying a fixed number of components, allowing us to capture the entire variance structure of the dataset. Feature importance scores were calculated by taking the absolute values of the PCA components and multiplying them by the explained variance of each component. These scores were then normalized to create a total importance score for each feature. Subsequently, features were sorted by importance, with the most significant predictors displayed at the top, allowing prioritization based on their contribution to the overall variance. To determine the optimal number of predictors, a cumulative explained variance curve was plotted. This curve illustrates the cumulative percentage of variance explained as additional features are included, aiding in identifying the "elbow point" where the addition of more features contributes minimally to the variance. This process helped in selecting an efficient subset of features for modeling.

2.6 | Parameter tuning

For each model, a set of hyperparameters was defined for optimization through *GridSearchCV* from the *scikit-learn* package (version 1.0.2) in Python, which performed an exhaustive search over specified hyperparameter combinations with 10-fold cross-validation. The models were evaluated using root mean squared error (RMSE) (Hodson, 2022) and coefficient of determination (R^2) (Cameron & Windmeijer, 1997) metrics on the test set, identifying the best-performing model and optimal hyperparameters for each approach. This systematic approach enabled the selection of the most suitable model configuration to predict dry pea yield and protein content effectively.

After doing parameter tuning, each model was trained on the training set using the *fit* function in Python, and predictions were made on the test set using the *predict* function. The predictions from each model were stored for further evaluation and comparison. To evaluate the performance of the models, RMSE and mean absolute error (MAE) (Hodson, 2022) error metrics were computed. RMSE was computed using the actual and predicted values of the yield and protein content to assess the average error in the model's predictions. MAE was used to evaluate the absolute differences between the predicted and actual yield values. Both RMSE and MAE were calculated for each model on the test set. The models were then compared based on these error metrics, and the results were visualized using bar plots for RMSE and MAE.

2.7 | Ensemble learning

In addition to individual models, a meta-model (ensemble learning) was constructed using linear regression. Ensemble learning is a machine learning approach that combines multiple base models to address regression or classification challenges (Fei et al., 2021). Stacking regression, introduced by Wolpert (1992), is a technique within this paradigm that merges predictions from various models to enhance accuracy. This method is particularly effective for creating ensembles with diverse types of predictors (Fei et al., 2021; Mienye & Sun, 2022). The predictions from the eight base models were stacked together to form a new feature set. This stacked feature set was used as the input to the meta-model, which was trained to predict yield and protein content based on the combined predictions of the base models. The meta-model was evaluated on the test set, and its performance was compared to the base models. To further assess the robustness of the models, a 10-fold cross-validation procedure (Fushiki, 2011) was applied, repeated 100 times for each model. The 10-fold cross-validation procedure was based on stratified random sampling to ensure that each fold maintained a balanced representation of plots from different years and locations. Cross-validation helps reduce model overfitting and provides a more generalizable measure of model performance (Montesinos López et al., 2022). For each fold, RMSE, MAE, and R^2 were computed, and the results were aggregated over 100 runs. The performance of the models was compared based on the distribution of RMSE, MAE, and R^2 values.

Finally, the trained models were applied to a new, unseen dataset to assess their generalization performance. Predictions were made using the trained base models, and their outputs were stacked to form the input for the meta-model. The performance on the unseen dataset was evaluated using RMSE, MAE, and the R^2 score. A scatter plot comparing the actual and predicted values of the meta-model was generated, along with the associated performance metrics. For each model (including the meta-model), the mean and standard deviation of RMSE and MAE were calculated across all cross-validation runs. These values were reported to provide a detailed comparison of the models' accuracy and robustness. To assess the capability of each growth stage in estimating yield and protein content, the most accurate model was trained using the training dataset (85% of the entire dataset) for separate datasets corresponding to each growth stage of dry peas. The average values of RMSE, MAE, and R^2 were recorded, derived from a 10-fold cross-validation repeated 100 times, using the test dataset (15% of the entire dataset). Based on the cross-validation results, the accuracy of each growth stage dataset in estimating yield and protein content was determined.

TABLE 2 Descriptive statistics of yield and protein content at different locations and years.

		Yield (kg ha^{-1})				Protein content (%)			
Location	**Year**	**Min**	**Mean**	**Max**	**SD**	**Min**	**Mean**	**Max**	**SD**
Carrington	2022	1040	3240	5113	612	20.74	24.1	28.14	1.14
	2024	1724	3941	5305	740	20.6	24.22	27.52	1.39
Minot	2022	1649	2640	3309	645	23.27	27.36	31.17	1.17
	2023	1406	3164	5305	873	23.89	27.22	31.12	1.64
	2024	1825	3584	5289	639	24.17	27.11	30.22	1.23

Abbreviations: Max, maximum; Min, minimum; SD, standard deviation.

3 | RESULTS

3.1 | Yield and protein variation across locations and years

The analysis of variance (ANOVA) revealed statistically significant differences in both yield and protein content across locations and years (p-value < 0.05). The full ANOVA results are presented in Table S1. These results indicate that the observed variations in yield and protein content are not due to random chance, but rather suggest meaningful differences influenced by both location and year factors. Descriptive statistics (Table 2) show that the average yield varied significantly between locations. These findings confirm that both location and year have substantial impacts on crop yield and protein content, warranting their inclusion as features in subsequent modeling to capture variability and improve the model's predictive accuracy across different temporal and spatial contexts.

3.2 | Feature selection

Based on PCA, feature importance scores identified key growth stages of dry peas influencing seed yield and protein content. Figure 2a illustrates the contributions of VIs, spectral bands, and their transformations during different growth stages (vegetative, flowering, maturity, and harvest) to yield estimation. The results indicate a uniform distribution of importance across features, with consistent contributions from variables such as the average values of red-edge and red bands at the vegetative and flowering stages. Similarly, for protein content prediction (Figure 2b), features like the red-edge and red bands at the vegetative and flowering stages show notable contributions, highlighting these stages as key for biochemical marker detection. These findings emphasize that integrating spectral data from multiple growth stages enhances model accuracy in capturing the dynamic relationship between crop development, yield, and protein content, underscoring the value of multistage data integration in precision agriculture.

Features were then ranked, and a cumulative explained variance curve was plotted to guide the selection of the optimal number of predictors for predicting yield (Figure 3a) and protein content (Figure 3b). The curve demonstrated a gradual increase in explained variance as additional features were included, with a notable "elbow point" suggesting that the 172 top-ranked features could efficiently represent the majority of variance in the data for both yield and protein contents.

3.3 | Hyperparameter tuning

The parameter tuning process using grid search revealed varying performances across different models and parameters for predicting both pea yield and protein content. Table 3 provides an overview of the models and the best-performing hyperparameters identified for each model and target variable.

3.4 | Machine learning results

Among the models tested for dry peas yield prediction, XGBoost performed best, achieving the lowest error and highest R^2 (0.81), closely followed by RF and GB. Models such as LightGBM, KNN, and AdaBoost showed moderate accuracy, while MLP and DT were the least effective, with MLP notably failing to capture yield patterns. For protein content prediction, LightGBM achieved the highest accuracy, with an R^2 value of 0.55, indicating the best model performance among those tested. RF and XGBoost followed but exhibited higher errors and lower predictive accuracy compared to LightGBM. Additionally, their performance in protein content prediction was poorer compared to their performance in yield estimation. Despite XGBoost's relatively higher accuracy for predicting yield and LightGBM's relatively higher accuracy for predicting protein content, none of the models effectively predicted yield and protein content in dry peas based solely on spectral information, highlighting the limitations of these data for accurate estimation.

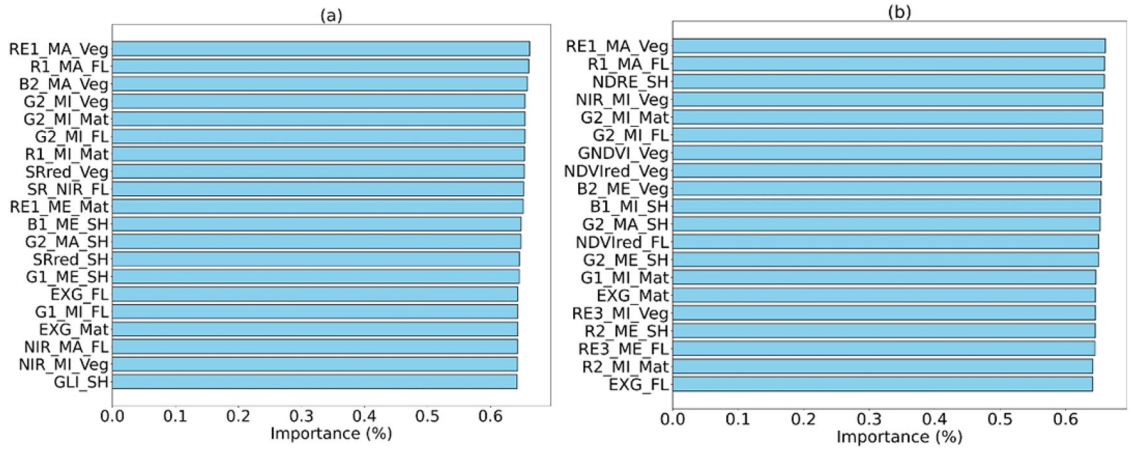


FIGURE 2 Future selection of the top 20 spectral features to predict the yield (a) and protein content (b) based on principal component analyses (PCAs). B2, 475 nm; EXG, excess green index; FL, flowering stage; G2, green 560 nm; GLI, green leaf index; GNDVI, green normalized vegetation index; MA, maximum; Mat, maturity stage; ME, mean; MI, minimum; NDRE, normalized difference red-edge Index; NDVIred, red-edge normalized difference vegetation index; NIR, near infrared; R1, 668 nm; RE1, red-edge 705 nm; RE3, 740 nm; SH, senescence and harvest stage; SR_NIR, simple ratio NIR index; SRred, simple ratio red-edge index; Veg, vegetative stage.

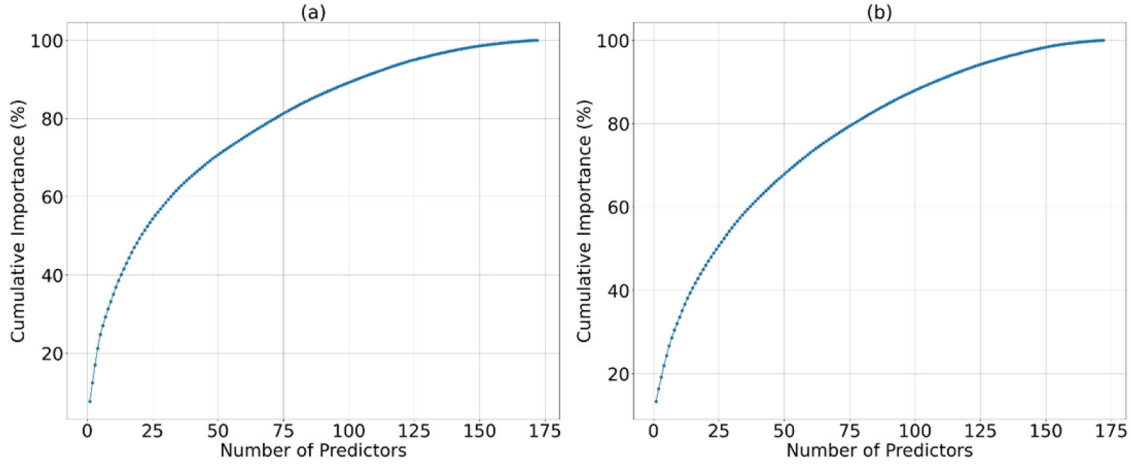


FIGURE 3 Cumulative explained variance curve showing the ideal number of features to predict yield (a) and protein content (b) in dry peas.

The comparison of machine learning models with the meta-model for predicting yield and protein content in dry peas reveals a distinct improvement in performance with the meta-model. For both yield and protein content, the meta-model shows a significant reduction in prediction error compared to individual models, highlighting its effectiveness in synthesizing information from multiple base models (Yang et al., 2024). The results demonstrate that while individual models struggle to accurately predict yield and protein content, the meta-model's integration of diverse predictions enables a more robust and accurate output. This underscores the potential of model stacking to enhance prediction accuracy when single models alone are insufficient. Table 4 shows the results of machine learning model and meta-model to estimate the yield and protein contents in dry peas based on the test datasets.

3.5 | Cross-validation

The 10-fold cross-validation, repeated 100 times, assessed the performance of multiple machine learning models, including a meta-model, for predicting yield. The RF, GB, and XGBoost models consistently performed well, showing low RMSE and MAE, indicating strong generalization and accuracy (Figure 4a). In contrast, DT and LightGBM models exhibited higher variability and occasional outliers, suggesting less consistency. The meta-model, built using the predictions from these models, outperformed the individual models with the lowest RMSE and MAE, highlighting the effectiveness of meta-learning in improving yield prediction accuracy.

Similarly, for protein content prediction (Figure 4b), the results further reinforce the strength of combining multiple

TABLE 3 Models, primary hyperparameters, values tested during grid search, and the optimal value for each parameter based on the tuning results used to evaluate models to predict both pea yield and protein content.

Model	Parameter	Tested values	Optimal value	
			Yield	Protein content
Random forest	Number of estimators	[50, 100, 200, 500, 1000]	100	50
	Maximum depth	[None, 10, 20, 50, 100]	50	20
	Minimum samples split	[2, 5, 10]	5	10
XGBoost	Number of estimators	[50, 100, 200, 500, 1000]	1000	50
	Learning rate	[0.0001, 0.001, 0.01, 0.1, 0.2]	0.2	0.1
	Maximum depth	[3, 5, 7, 9, 11]	3	5
MLP	Hidden layer sizes	[(50,), (100,), (50, 50)][a]	(50,)	(50, 50)
	Activation	['relu', 'tanh']	'relu'	'tanh'
	Alpha	[0.0001, 0.001]	0.001	0.0001
	Learning rate	['constant', 'adaptive']	'adaptive'	'constant'
LightGBM	Number of estimators	[50, 100, 200, 500, 1000]	50	50
	Learning rate	[0.0001, 0.001, 0.01, 0.1, 0.2]	0.1	0.1
	Maximum depth	[None,3, 5, 7, 9, 11]	3	3
	Number of leaves	[31, 50]	31	31
	Minimum data in leaf	[10, 20, 30, 40, 50]	20	10
KNN	Number of neighbors	[3, 5, 7, 9, 11]	9	7
	Weights	['uniform', 'distance']	'distance'	'distance'
Gradient boosting	Number of estimators	[50, 100, 200, 500, 1000]	500	50
	Learning rate	[0.0001, 0.001, 0.01, 0.1, 0.2]	0.2	0.2
	Maximum depth	[3, 5, 7, 9, 11]	3	3
AdaBoost	Number of estimators	[50, 100, 200, 500, 1000]	1000	100
	Learning rate	[0.0001, 0.001, 0.01, 0.1, 0.2]	1	1
Decision tree	Maximum depth	[None, 10, 20, 50, 100]	10	10
	Minimum samples split	[2, 5, 10]	10	10

[a](100,) represents a single hidden layer with 100 neurons. (50, 50) represents a two-layer neural network with 50 neurons in each layer.
Abbreviations: KNN, K-nearest neighbors; LightGBM, light gradient boosting machine; MLP, multilayer perceptron.

TABLE 4 Performance comparison of base models and meta-model to estimate yield and protein contents in dry peas.

Model	Yield (kg ha^{-1})			Protein content (%)		
	RMSE	MAE	R^2 score	RMSE	MAE	R^2 score
MLP	439.70	384.18	0.50	32.45	26.85	0.21
Random forest	328.80	247.07	0.78	1.25	0.89	0.55
XGBoost	307.56	238.23	0.81	1.27	0.87	0.53
KNN	377.86	300.36	0.71	1.29	0.86	0.51
Gradient boosting	341.53	257.32	0.76	1.36	1.03	0.46
AdaBoost	379.47	304.69	0.70	1.27	0.92	0.53
Decision tree	438.36	343.40	0.61	1.49	1.09	0.35
LightGBM	357.29	276.86	0.74	1.25	0.86	0.55
Meta-Model	127.64	105.89	0.98	0.36	0.28	0.96

Abbreviations: KNN, *K*-nearest neighbors; LightGBM, light gradient boosting machine; MAE, mean absolute error; MLP, multilayer perceptron; RMSE, root mean squared error.

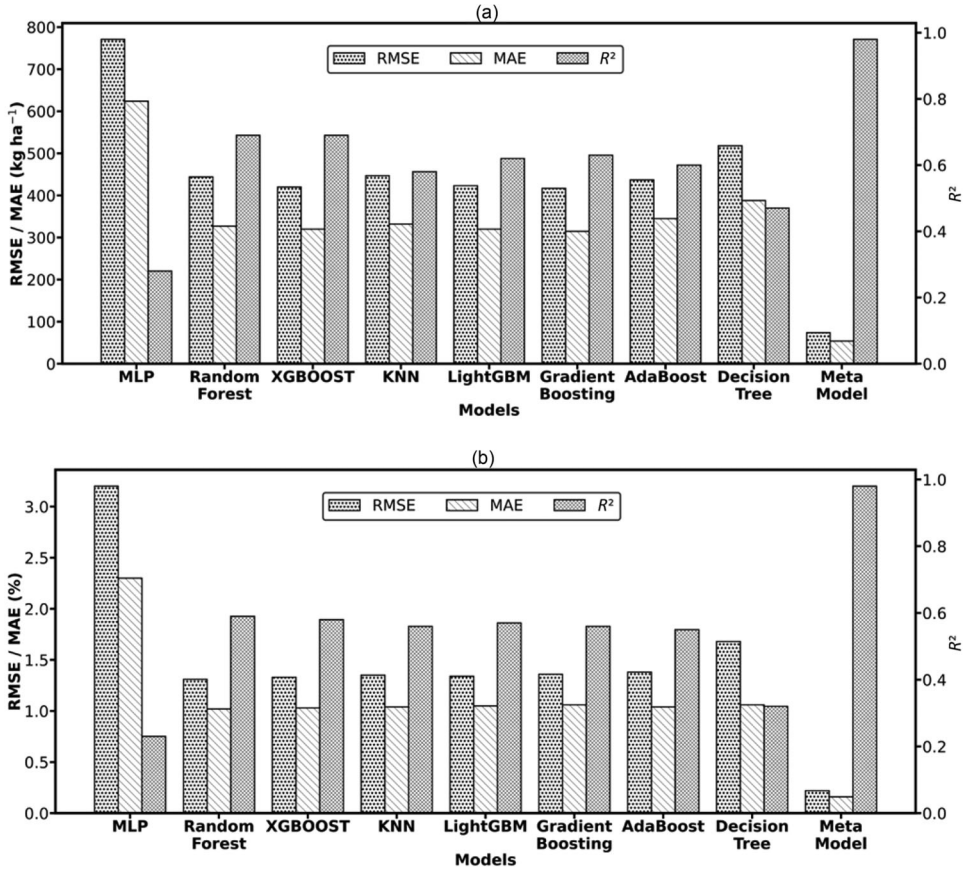


FIGURE 4 Results of a 10-fold cross-validation approach using the test datasets for machine learning models used to estimate yield (a) and protein content (b) in dry peas. KNN, *K*-nearest neighbors; LightGBM, light gradient boosting machine; MAE, mean absolute error; MLP, multilayer perceptron; RMSE, root mean squared error.

models. The RF, XGBoost, and GB models again exhibited strong performance, with low RMSE and MAE values, confirming their capability to generalize well on protein content estimation. In contrast, DT tree and LightGBM models displayed more erratic performance, with larger errors and outliers in their predictions. The meta-model, leveraging the strengths of the base models, again achieved the best results with the lowest RMSE and MAE, demonstrating that integrating multiple models through meta-learning can significantly enhance predictive accuracy for both yield and protein content estimation. These findings suggest that meta-modeling is a valuable technique for improving performance, particularly for complex agricultural tasks involving multiple variables.

3.6 | Model reliability

While all eight individual machine learning models exhibited poor performance in estimating yield and protein content on the unseen dataset (Figure 5), the meta-model demonstrated strong predictive capability for dry peas (Figure 6a), achieving an RMSE of 396.28 kg ha^{-1}, an R^2 of 0.77, and an MAE of 272.4 kg ha^{-1}. This high R^2 value indicates that the meta-model effectively captures a substantial proportion of the yield variability, underscoring its reliability and robustness in generalizing to new data. The low RMSE and MAE further confirm the accuracy of the meta-model, highlighting its ability to synthesize and leverage predictions from multiple base models to enhance predictive performance. These results suggest that the meta-model provides a promising approach for yield estimation in dry peas when using spectral and related features. The meta-model's performance in estimating protein content in dry peas (Figure 6b) is moderate, achieving an RMSE of 1.28% and an R^2 of 0.54. While it captures some of the variability in protein content, the R^2 value is notably lower than that observed for yield estimation (0.77). This suggests that the meta-model is more effective at predicting yield than protein content, potentially due to a lower correlation between spectral features and protein content. The higher RMSE and lower R^2 indicate challenges in accurately modeling protein content using the current dataset, implying that additional data or alternative features may be needed to enhance prediction accuracy for protein content in dry peas. To assess within-field prediction performance, the meta-model was tested separately on individual location-year datasets (Table 5). The results showed

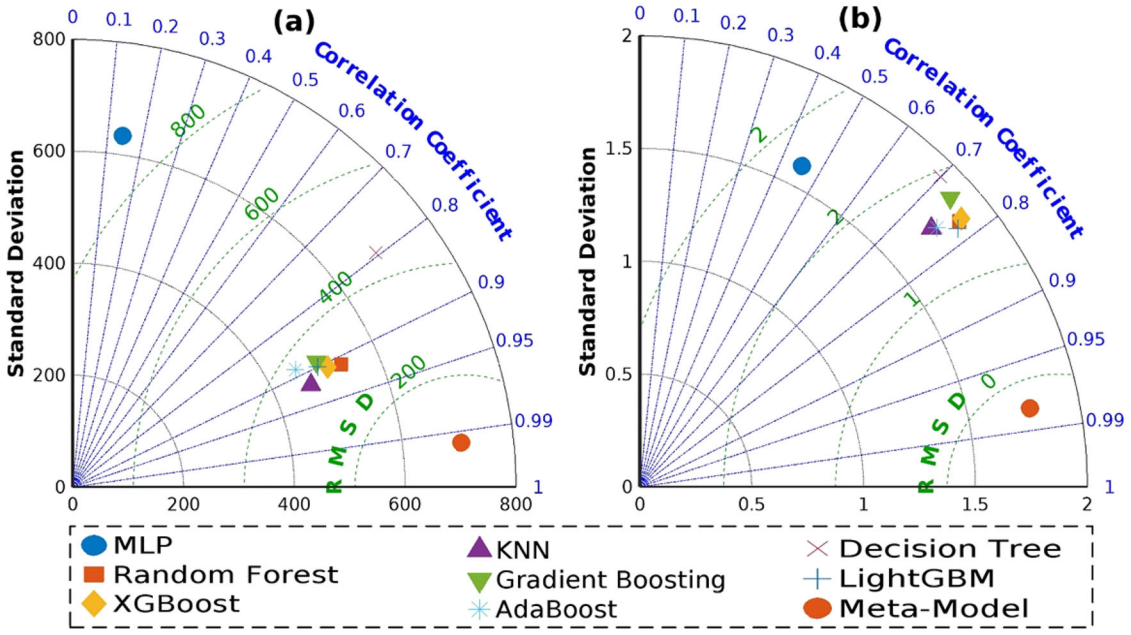


FIGURE 5 The Taylor plot to show the performance of models to estimate the yield (a) and protein content (b) in dry peas for the validation datasets. KNN, *K*-nearest neighbors; LightGBM, light gradient boosting machine; MLP, multilayer perceptron; RMSD, root mean square difference.

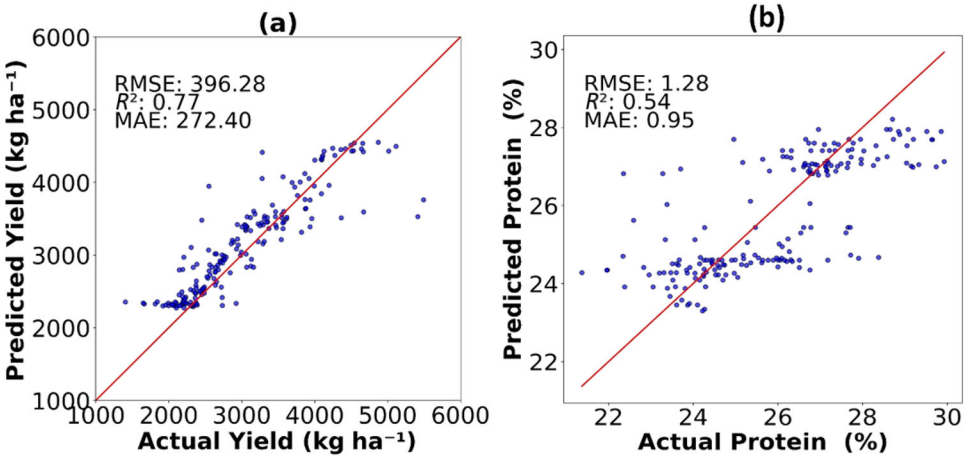


FIGURE 6 Performance of the meta-model to estimate the yield (a) and protein content (b) in dry peas for the validation datasets. MAE, mean absolute error; RMSE, root mean squared error.

TABLE 5 Prediction performance of the meta-model for yield and protein content across individual location-year field environments. Metrics include R^2 (coefficient of determination), RMSE (root mean squared error), and MAE (mean absolute error), showing slightly reduced performance within fields compared to the full dataset, especially for protein content.

		Yield (kg ha^{-1})			Protein (%)		
Location	**Year**	**RMSE**	**MAE**	**R^2 score**	**RMSE**	**MAE**	**R^2 score**
Carrington	2022	452.1	342.6	0.67	1.45	1.09	0.42
	2024	388.3	299.7	0.74	1.37	1.02	0.48
Minot	2022	417.5	311.4	0.71	1.41	1.04	0.45
	2023	360.2	281.6	0.79	1.26	0.95	0.53
	2024	375.8	293	0.76	1.32	0.98	0.5

R^2 values ranging from 0.67 to 0.79 across fields, indicating slightly reduced but still strong predictive performance when models are applied within specific environments. These results suggest that while the meta-model generalizes well, localized factors may introduce additional variability.

While the meta-model demonstrated strong performance in yield prediction, its accuracy in predicting protein content was noticeably lower, with an R^2 of 0.54. This suggests only a modest correlation between spectral features and protein synthesis. Given the complexity of protein accumulation, which is influenced by factors such as nitrogen availability, genotype, and environmental interactions, it is possible that the observed predictive performance is partly incidental or driven by indirect associations. These findings indicate that the current methodology is more reliable for yield estimation, whereas protein content prediction remains a secondary outcome with potential but requires further refinement and supporting data sources for improvement.

After identifying the meta-model as the most accurate model for estimating yield and protein content in dry peas, its capability was assessed for each growth stage. The meta-model was trained using the training dataset for separate datasets corresponding to each growth stage of dry peas. The cross-validation results of the meta-model, tested with data from each growth stage, indicate that the maturity stage emerged as the most effective for estimating yield and protein content in dry peas based solely on spectral information. Figure 7 illustrates the cross-validation results of the meta-model when fed with datasets from each growth stage of dry peas.

4 | DISCUSSION

The significant differences in crop yield and protein content across locations and years highlight the critical role of environmental and temporal factors in agricultural outputs. These findings align with previous studies emphasizing the influence of site-specific conditions such as soil fertility, climate, and water availability, as well as seasonal climate variability. For example, some studies (Basso et al., 2009; Huang et al., 2021; Příkopa et al., 2005; Rong et al., 2021) reported significant location-based yield variations due to soil and climate differences, while others (Heil et al., 2020; Lobell & Field, 2007) linked annual climate fluctuations to protein content. Similarly, Astrauskas et al. (2022) observed higher yields in winter wheat in years with optimal rainfall than in drought-impacted years. Incorporating location and year as variables in predictive models is widely supported in agricultural research, as it enhances precision by capturing spatiotemporal variability. Some studies (Prey et al., 2023; Stepanov et al., 2020) demonstrated that such models significantly improve yield prediction accuracy. Our findings reinforce this approach, supporting the integration of environmental and temporal variables in predictive models to enhance adaptability and precision across diverse regions and years. This study reinforces the importance of multi-stage and multi-index approaches for predicting yield and protein content, aligning with prior research. Consistent with Li-Hong et al. (2007) and Zhou et al. (2021), key predictors like the red-edge and red bands demonstrate the significance of spectral indices across growth stages. The relevance of nitrogen-sensitive indices, such as the normalized difference red-edge Index, noted by Colovic et al. (2022), is evident, particularly for protein content prediction. Similarly, our findings affirm the value of integrating spectral data across stages, as highlighted by Darra et al. (2024) and Thenkabail et al. (2000), for capturing crop growth dynamics and enhancing biochemical analysis for precision agriculture applications.

The superior performance of the meta-model in predicting both yield and protein content can be attributed to its ability to combine the strengths of multiple machine learning models, mitigating individual model weaknesses. Models like the RF, GB, and XGBoost inherently excel at capturing complex, nonlinear interactions between variables, which are critical in agricultural predictions involving diverse environmental, genetic, and management factors (Bazrafkan et al., 2024). However, their individual predictions may still be influenced by specific biases or limitations. By leveraging meta-learning, the meta-model synthesizes these individual predictions, effectively reducing overfitting and enhancing generalization, as evidenced by the lowest RMSE and MAE values observed. These results align with previous findings where ensemble and stacking approaches demonstrated improved predictive performance in complex agricultural datasets. For instance, Gharakhanlou and Perez (2024) highlighted that ensemble methods significantly enhance yield prediction by integrating diverse data sources and models. Similarly, Fei et al. (2021) and Q.-C. Li, Xu, et al. (2023) found that the RF model performs robustly in yield predictions, but their accuracy further improves when used in combination through meta-modeling. The relatively poorer performance of DT and LightGBM models in this study, marked by higher variability and outliers, may stem from their sensitivity to noisy data or inadequate handling of extreme cases (Atla et al., 2011). These findings reinforce the growing consensus that meta-learning approaches are particularly effective for agricultural predictions, where the interplay of numerous complex factors requires robust, flexible modeling techniques.

The meta-model's strong performance in predicting yield for dry peas, with an RMSE of 396.28 kg ha^{-1}, an R^2 of 0.77, and an MAE of 272.4 kg ha^{-1}, underscores its robustness and ability to generalize effectively to unseen data. The high R^2 value reflects the model's capacity to capture the majority of yield variability, likely due to the integration of

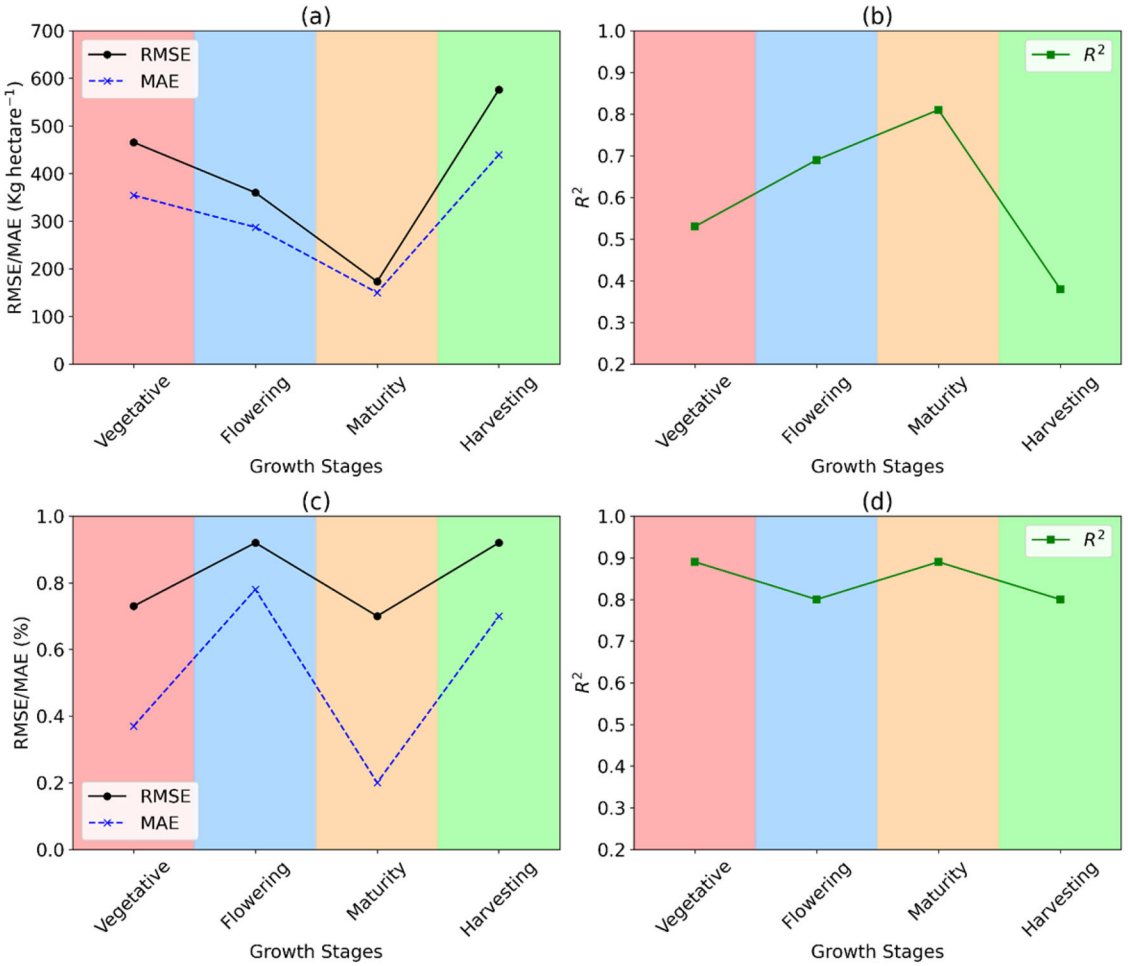


FIGURE 7 Performance of the meta-model in estimating yield and protein content in dry peas across growth stages: (a) error metrics for yield estimation, (b) R^2 values for yield estimation, (c) error metrics for protein content estimation, and (d) R^2 values for protein content estimation. MAE, mean absolute error; RMSE, root mean squared error.

diverse base models, which collectively leverage the spectral and related features to provide comprehensive insights. This aligns with findings from Parashar et al. (2024) and Shahhosseini et al. (2020), who demonstrated that meta-models combining ensemble methods outperform single models by synthesizing predictions and capturing complex relationships in agricultural datasets. The low RMSE and MAE further highlight the meta-model's precision, making it a promising tool for yield prediction in dry peas.

In contrast, the meta-model's moderate performance in protein content prediction, with an RMSE of 1.28% and an R^2 of 0.54, suggests limitations in capturing protein variability. The lower R^2 indicates that the spectral features used in the model are less correlated with protein content compared to yield, potentially due to the more complex or less direct relationship between protein synthesis and the spectral indices. This mirrors findings by Wu et al. (2020), who noted that stacking approaches can face challenges when feature relevance is lower for the target variable. This research highlights that spectral reflectance indices exhibit varying degrees of correlation with yield (Figure 8a) and protein content (Figure 8b). While some spectral indices show strong positive correlations with yield, those most correlated with protein content tend to exhibit negative relationships. These differences may contribute to variations in model performance when predicting protein content compared to yield. These findings align with the results of Sandhu et al. (2021) and Zhou et al. (2021). To improve protein content predictions, additional data sources, such as biochemical or soil nutrient measurements, or alternative feature engineering methods may be necessary. These results highlight the differential predictive capabilities of the meta-model across traits and the importance of tailored datasets for specific agricultural predictions.

The findings underscore that the meta-model's performance varies notably across the growth stages of dry peas, with the maturity stage proving to be the most effective for estimating both yield and protein content based on spectral information. This suggests that the spectral characteristics

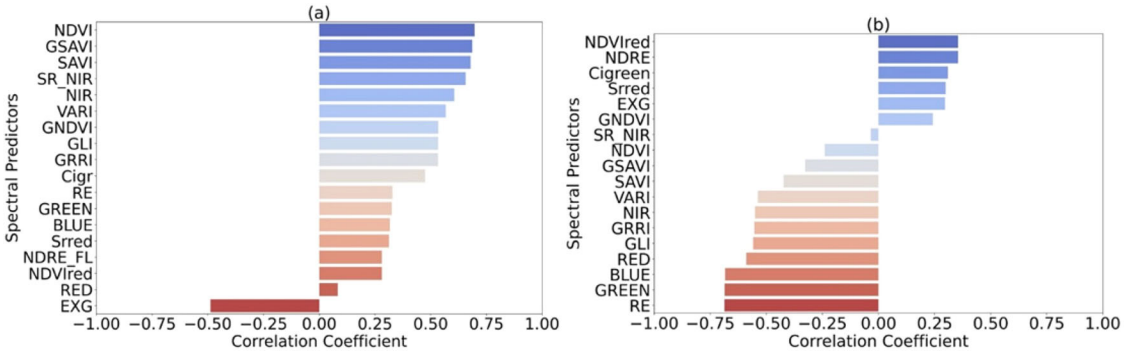


FIGURE 8 The relationship between spectral information and yield (a) and protein content (b) in dry peas at the maturity stage of growth, represented by R^2 values.

captured at maturity provide the most reliable indicators for these agronomic traits, likely due to the fully developed physiological and biochemical markers associated with yield and protein accumulation (Christenson et al., 2016). At maturity, plant pigments, canopy structure, and biochemical compositions such as chlorophyll, carotenoids, and nitrogen-related compounds have stabilized, influencing spectral reflectance patterns that can be detected using multispectral imagery (Bazrafkan et al., 2023). This outcome aligns with expectations that later growth stages reflect more comprehensive and integrative information about crop development, leading to improved model accuracy. These findings are consistent with a study on cereal crops like wheat (Y. Zhang et al., 2018), which demonstrate that stages close to maturity often yield higher prediction accuracy for traits such as biomass, yield, and quality attributes, as these traits become more pronounced and stable during the later phases of growth. Similarly, research on legumes, including soybeans (Rajah et al., 2017), has observed higher prediction accuracies at or near harvest time, further corroborating the results of this study. However, discrepancies may arise due to crop-specific differences, as well as variations in spectral preprocessing and modeling approaches. Therefore, while the maturity stage consistently demonstrates its utility, further research is needed to determine whether these trends are universally applicable or influenced by external factors such as environmental conditions or management practices. Future work should explore cross-validation strategies that exclude entire years or locations from training sets, to better evaluate model generalizability across unseen environments and assess genotype-by-environment (G × E) interactions more explicitly.

Translating UAS-based multispectral yield prediction models to satellite platforms holds great promise for large-scale agricultural monitoring. While satellites such as Sentinel-2 provide consistent and scalable data, challenges remain due to their coarser spatial resolution and differences in spectral band configurations (Bazrafkan et al., 2025). These differences may particularly impact the accuracy of trait predictions like protein content, which is less directly captured by canopy-level reflectance. Nonetheless, with proper recalibration and data harmonization techniques (e.g., transfer learning, multisource fusion), satellite data could effectively support generalized yield prediction frameworks. Future studies should explore model portability and performance trade-offs between drone and satellite platforms.

5 | CONCLUSION

This study aimed to address key research questions on improving the accuracy of yield and protein content predictions in dry peas, using multispectral data and stacked machine learning models. Our results demonstrate that the integration of multispectral data with meta-models significantly enhances the prediction accuracy for yield (RMSE = 396.28 kg ha^{-1}, R^2 = 0.77), compared to traditional and single-model approaches. This supports the hypothesis that combining multiple base models into a meta-model effectively captures the complex, nonlinear relationships in the data, thus enhancing prediction robustness and generalizability. However, the performance of the meta-model in predicting protein content was moderate (RMSE = 1.28%, R^2 = 0.54), suggesting that the spectral features used are more strongly correlated with yield than with protein content. These findings partially confirm our hypothesis by highlighting the differential predictive capabilities of spectral data across traits. The maturity stages emerged as the most critical growth phases for influencing both yield and protein content, further validating the importance of stage-specific spectral data in modeling. Developing models tailored to specific growth stages, particularly during anthesis and grain filling, may address the variability in model performance observed across stages. In addition, expanding the application of stacked models to other pulse crops and diverse environmental conditions would validate the generalizability of the proposed approach. Future studies could

combine UAS-acquired spectral data with real-time decision-support systems, enabling adaptive management practices to optimize yield and protein content. While the results confirm the value of multispectral data and ensemble learning for predicting dry pea yield with high accuracy, protein content estimation remains a more challenging task. The moderate performance for protein prediction suggests that spectral features alone may not capture all relevant biochemical variations. Therefore, protein content should be viewed as a secondary target in this study, with preliminary evidence of feasibility but not yet robust accuracy. Future research should prioritize the integration of additional data types, such as soil nutrient profiles, biochemical markers, or genomic information, and test broader VI sets to improve the reliability of protein content prediction models.

AUTHOR CONTRIBUTIONS

Aliasghar Bazrafkan: Conceptualization; data curation; formal analysis; investigation; methodology; software; validation; visualization; writing—original draft; writing—review and editing. **Hannah Worral**: Data curation; investigation; writing—review and editing. **Nonoy Bandillo**: Funding acquisition; project administration; writing—review and editing. **Paulo Flores**: Funding acquisition; project administration; resources; supervision; writing—review and editing.

ACKNOWLEDGMENTS

The authors sincerely thank the Northern Pulse Growers Association for their ongoing financial support in advancing NDSU's pea breeding lines, as well as the U.S. Department of Agriculture's Agricultural Research Service for supplying drones and sensors. The opinions, findings, conclusions, or recommendations presented here are those of the author(s) and do not necessarily reflect the views of the funding agencies.

DATA AVAILABILITY STATEMENT

The datasets generated during and/or analyzed during the current research are available from the following link. http://datadryad.org/share/YxTZ3IC4_S4yCi6LW6TG66NeqtuZcC-Dftjas1DHnOU

ORCID

Aliasghar Bazrafkan https://orcid.org/0000-0002-7862-0520

Paulo Flores https://orcid.org/0000-0002-3964-6904

REFERENCES

Abdi, H. (2007). Z-scores. *Encyclopedia of Measurement and Statistics*, *3*, 1055–1058.

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*, 433–459. https://doi.org/10.1002/wics.101

Astaoui, G., Dadaiss, J. E., Sebari, I., Benmansour, S., & Mohamed, E. (2021). Mapping wheat dry matter and nitrogen content dynamics and estimation of wheat yield using UAV multispectral imagery machine learning and a variety-based approach: Case study of Morocco. *AgriEngineering*, *3*, 29–49. https://doi.org/10.3390/agriengineering3010003

Astrauskas, P., Staugaitis, G., Masevičienė, A., & Žičkienė, L. (2022). Variation of winter wheat grain yield and its quality in fields with different soil cover. *Zemdirbyste-Agriculture*, *109*, 297–304. https://doi.org/10.13080/z-a.2022.109.038

Atla, A., Tada, R., Sheng, V., & Singireddy, N. (2011). Sensitivity of different machine learning algorithms to noise. *Journal of Computing Sciences in Colleges*, *26*, 96–103.

Basso, B., Cammarano, D., Chen, D., Cafiero, G., Amato, M., Bitella, G., Rossi, R., & Basso, F. (2009). Landscape position and precipitation effects on spatial variability of wheat yield and grain protein in Southern Italy. *Journal of Agronomy and Crop Science*, *195*, 301–312. https://doi.org/10.1111/j.1439-037X.2008.00351.x

Bazrafkan, A., Cannayen, I., Bandillo, N., & Flores, P. (2025). Optimizing integration techniques for UAS and satellite image data in precision agriculture—A review. *Frontiers in Remote Sensing*, *6*, Article 1622884. https://doi.org/10.3389/frsen.2025.1622884

Bazrafkan, A., Navasca, H., Kim, J.-H., Morales, M., Johnson, J. P., Delavarpour, N., Fareed, N., Bandillo, N., & Flores, P. (2023). Predicting dry pea maturity using machine learning and advanced sensor fusion with unmanned aerial systems (UASs). *Remote Sensing*, *15*, Article 2758. https://doi.org/10.3390/rs15112758

Bazrafkan, A., Navasca, H., Worral, H., Oduor, P., Delavarpour, N., Morales, M., Bandillo, N., & Flores, P. (2024). Predicting lodging severity in dry peas using UAS-mounted RGB, LIDAR, and multispectral sensors. *Remote Sensing Applications: Society and Environment*, *34*, Article 101157. https://doi.org/10.1016/j.rsase.2024.101157

Bernard, J. (2016). Python data analysis with pandas. In J. Bernard (Ed.), *Python recipes handbook: A problem-solution approach* (pp. 37–48). Apress. https://doi.org/10.1007/978-1-4842-0241-8

Boye, J., Zare, F., & Pletch, A. (2010). Pulse proteins: Processing, characterization, functional properties and applications in food and feed. *Food Research International*, *43*, 414–431. https://doi.org/10.1016/J.FOODRES.2009.09.003

Burns, B. W., Green, V. S., Hashem, A. A., Massey, J. H., Shew, A. M., Adviento-Borbe, M. A. A., & Milad, M. (2022). Determining nitrogen deficiencies for maize using various remote sensing indices. *Precision Agriculture*, *23*, 791–811. https://doi.org/10.1007/s11119-021-09861-4

Cameron, A. C., & Windmeijer, F. A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, *77*, 329–342. https://doi.org/10.1016/S0304-4076(96)01818-0

Candiago, S., Remondino, F., De Giglio, M., Dubbini, M., & Gattelli, M. (2015). Evaluating multispectral images and vegetation indices for precision farming applications from UAV images. *Remote Sensing*, *7*, 4026–4047. https://doi.org/10.3390/rs70404026

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Christenson, B. S., Schapaugh, W. T., Jr, An, N., Price, K. P., Prasad, V., & Fritz, A. K. (2016). Predicting soybean relative maturity and seed yield using canopy reflectance. *Crop Science*, *56*, 625–643. https://doi.org/10.2135/cropsci2015.04.0237

Colovic, M., Yu, K., Todorovic, M., Cantore, V., Hamze, M., Albrizio, R., & Stellacci, A. M. (2022). Hyperspectral vegetation indices to assess water and nitrogen status of sweet maize crop. *Agronomy*, *12*, Article 2181. https://doi.org/10.3390/agronomy12092181

Darra, N., Garcia, B. E., Psomiadis, E., & Fountas, S. (2024). Optimizing crop yield prediction: Inter-dated vegetation indices & AutoML ensembles in processing tomato crop. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium* (4320–4324). IEEE.

Dhanabal, S., & Chandramathi, S. (2011). A review of various K-nearest neighbor query processing techniques. *International Journal of Computer Applications*, *31*, 14–22.

Dong, T., Liu, J., Shang, J., Qian, B., Ma, B., Kovacs, J. M., Walters, D., Jiao, X., Geng, X., & Shi, Y. (2019). Assessment of red-edge vegetation indices for crop leaf area index estimation. *Remote Sensing of Environment*, *222*, 133–143. https://doi.org/10.1016/j.rse.2018.12.032

Eng, L. S., Ismail, R., Hashim, W., & Baharum, A. (2019). The use of VARI, GLI, and VIgreen formulas in detecting vegetation in aerial images. *International Journal of Technology*, *10*, 1385–1394. https://doi.org/10.14716/ijtech.v10i7.3275

Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., & Zeng, W. (2019). Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural Water Management*, *225*, Article 105758. https://doi.org/10.1016/j.agwat.2019.105758

Fei, S., Hassan, M. A., He, Z., Chen, Z., Shu, M., Wang, J., Li, C., & Xiao, Y. (2021). Assessment of ensemble learning to predict wheat grain yield based on UAV-multispectral reflectance. *Remote Sensing*, *13*, Article 2338. https://doi.org/10.3390/rs13122338

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*, 137–146. https://doi.org/10.1007/s11222-009-9153-8

Gharakhanlou, N. M., & Perez, L. (2024). From data to harvest: Leveraging ensemble machine learning for enhanced crop yield predictions across Canada amidst climate change. *Science of the Total Environment*, *951*, Article 175764. https://doi.org/10.1016/j.scitotenv.2024.175764

Heil, K., Lehner, A., & Schmidhalter, U. (2020). Influence of climate conditions on the temporal development of wheat yields in a long-term experiment in an area with Pleistocene loess. *Climate*, *8*, Article 100. https://doi.org/10.3390/cli8090100

Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, *2022*, 1–10.

Huang, J., Hartemink, A. E., & Kucharik, C. J. (2021). Soil-dependent responses of US crop yields to climate variability and depth to groundwater. *Agricultural Systems*, *190*, Article 103085. https://doi.org/10.1016/j.agsy.2021.103085

Jiang, J., Cai, W., Zheng, H., Cheng, T., Tian, Y., Zhu, Y., Ehsani, R., Hu, Y., Niu, Q., Gui, L., & Yao, X. (2019). Using digital cameras on an unmanned aerial vehicle to derive optimum color vegetation indices for leaf nitrogen concentration monitoring in winter wheat. *Remote Sensing*, *11*, Article 2667. https://doi.org/10.3390/rs11222667

Large, E. C. (1954). Growth stages in cereals illustration of the Feekes scale. *Plant Pathology*, *3*, 128–129. https://doi.org/10.1111/j.1365-3059.1954.tb00716.x

Li, Z., Jin, X., Zhao, C., Wang, J., Xu, X., Yang, G., Li, C., & Shen, J. (2015). Estimating wheat yield and quality by coupling the DSSAT-CERES model and proximal remote sensing. *European Journal of Agronomy*, *71*, 53–62. https://doi.org/10.1016/J.EJA.2015.08.006

Li, Z., Wang, J., Xu, X., Zhao, C., Jin, X., Yang, G., & Feng, H. (2015). Assimilation of two variables derived from hyperspectral data into the DSSAT-CERES model for grain yield and quality estimation. *Remote Sensing*, *7*, 12400–12418. https://doi.org/10.3390/rs70912400

Li, W., Wu, W., Yu, M., Tao, H., Yao, X., Cheng, T., Zhu, Y., Cao, W., & Tian, Y. (2023). Monitoring rice grain protein accumulation dynamics based on UAV multispectral data. *Field Crops Research*, *294*, Article 108858. https://doi.org/10.1016/j.fcr.2023.108858

Li, Q.-C., Xu, S.-W., Zhuang, J.-Y., Liu, J.-J., Yi, Z., & Zhang, Z.-X. (2023). Ensemble learning prediction of soybean yields in China based on meteorological data. *Journal of Integrative Agriculture*, *22*, 1909–1927. https://doi.org/10.1016/j.jia.2023.02.011

Li-Hong, X., Wei-Xing, C., & Lin-Zhang, Y. (2007). Predicting grain yield and protein content in winter wheat at different N supply levels using canopy reflectance spectra. *Pedosphere*, *17*, 646–653.

Liu, K., Bandara, M., Hamel, C., Knight, J. D., & Gan, Y. (2020). Intensifying crop rotations with pulse crops enhances system productivity and soil organic carbon in semi-arid environments. *Field Crops Research*, *248*, Article 107657. https://doi.org/10.1016/j.fcr.2019.107657

Liu, L., Knight, J., Lemke, R., & Farrell, R. (2021). Type of pulse crop included in a 2-year rotation with wheat affects total N_2O loss and intensity. *Biology and Fertility of Soils*, *57*, 699–713. https://doi.org/10.1007/s00374-021-01562-4

Lobell, D. B., & Field, C. B. (2007). Global scale climate–crop yield relationships and the impacts of recent warming. *Environmental Research Letters*, *2*, Article 014002. https://doi.org/10.1088/1748-9326/2/1/014002

Lu, J., Cheng, D., Geng, C., Zhang, Z., Xiang, Y., & Hu, T. (2021). Combining plant height, canopy coverage and vegetation index from UAV-based RGB images to estimate leaf nitrogen concentration of summer maize. *Biosystems Engineering*, *202*, 42–54. https://doi.org/10.1016/j.biosystemseng.2020.11.010

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, *10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 109–139) Springer. https://doi.org/10.1007/978-3-030-89010-0

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing*, *2*, 183–197. https://doi.org/10.1016/0925-2312(91)90023-5

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*, Article 21. https://doi.org/10.3389/fnbot.2013.00021

Parashar, N., Johri, P., Khan, A. A., Gaur, N., & Kadry, S. (2024). An integrated analysis of yield prediction models: A comprehensive review of advancements and challenges. *Computers, Materials & Continua*, *80*, 389–425.

Parganiha, V., & Verma, M. (2024). An effective soil analysis and crop yield prediction based on optimised light GBM in smart agriculture. *Journal of Agronomy and Crop Science*, *210*, Article e12726. https://doi.org/10.1111/jac.12726

Pranga, J., Borra-Serrano, I., Aper, J., De Swaef, T., Ghesquiere, A., Quataert, P., Roldán-Ruiz, I., Janssens, I. A., Ruysschaert, G., & Lootens, P. (2021). Improving accuracy of herbage yield predictions in perennial ryegrass with UAV-based structural and spectral data fusion and machine learning. *Remote Sensing*, *13*, Article 3459. https://doi.org/10.3390/rs13173459

Prey, L., Ramgraber, L., Seidl-Schulz, J., Hanemann, A., & Noack, P. O. (2023). The transferability of spectral grain yield prediction in wheat breeding across years and trial locations. *Sensors*, *23*, Article 4177. https://doi.org/10.3390/s23084177

Příkopa, M., Richter, R., Zimolka, J., & Cerkal, R. (2005). The influence of the year, fore-crops and fertilisation on yield and content of crude protein in spring barley. *Plant, Soil and Environment*, *51*, 144–150.

Rajah, P., Odindi, J., Abdel-Rahman, E., & Mutanga, O. (2017). Determining the optimal phenological stage for predicting common dry bean (*Phaseolus vulgaris*) yield using field spectroscopy. *South African Journal of Plant and Soil*, *34*, 379–388. https://doi.org/10.1080/02571862.2017.1317854

Ren, P., Li, H., Han, S., Chen, R., Yang, G., Yang, H., Feng, H., & Zhao, C. (2023). Estimation of soybean yield by combining maturity group information and unmanned aerial vehicle multi-sensor data using machine learning. *Remote Sensing*, *15*(17), Article 4286. https://doi.org/10.3390/rs15174286

Rong, L., Duan, X., Gu, Z., & Feng, D. (2021). Climatic and environmental drivers on temporal-spatial variations of grain meteorological yield in high mountainous region. *Archives of Agronomy and Soil Science*, *67*, 2000–2014. https://doi.org/10.1080/03650340.2020.1819531

Sandhu, K., Patil, S. S., Pumphrey, M., & Carter, A. (2021). Multitrait machine-and deep-learning models for genomic selection using spectral information in a wheat breeding program. *The Plant Genome*, *14*, Article e20119. https://doi.org/10.1002/tpg2.20119

Sarkar, T. K., Roy, D. K., Kang, Y. S., Jun, S. R., Park, J. W., & Ryu, C. S. (2024). Ensemble of machine learning algorithms for rice grain yield prediction using UAV-based remote sensing. *Journal of Biosystems Engineering*, *49*, 1–19. https://doi.org/10.1007/s42853-023-00209-6

Segal, M. R. (2004). *Machine learning benchmarks and random forest regression*. https://escholarship.org/uc/item/35x3v9t4

Seni, G., & Elder, J. (2010). *Ensemble methods in data mining: Improving accuracy through combining predictions*. Morgan & Claypool Publishers.

Shafian, S., Rajan, N., Schnell, R., Bagavathiannan, M., Valasek, J., Shi, Y., & Olsenholler, J. (2018). Unmanned aerial systems-based remote sensing for monitoring sorghum growth and development. *PLoS ONE*, *13*, Article e0196605. https://doi.org/10.1371/journal.pone.0196605

Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science*, *11*, Article 1120. https://doi.org/10.3389/fpls.2020.01120

Shanthakumar, P., Klepacka, J., Bains, A., Chawla, P., Dhull, S. B., & Najda, A. (2022). The current situation of pea protein and its application in the food industry. *Molecules*, *27*, Article 5354. https://doi.org/10.3390/molecules27165354

Sharifi, A. (2021). Yield prediction with machine learning algorithms and satellite images. *Journal of the Science of Food and Agriculture*, *101*(3), 891–896. https://doi.org/10.1002/jsfa.10696

Sims, D. A., & Gamon, J. A. (2002). Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sensing of Environment*, *81*, 337–354. https://doi.org/10.1016/S0034-4257(02)00010-X

Solomatine, D. P., & Shrestha, D. L. (2004). AdaBoost. RT: A boosting algorithm for regression problems. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)* (1163–1168). IEEE. https://doi.org/10.1109/IJCNN.2004.1380102

Stanton, C., Starek, M. J., Elliott, N., Brewer, M., Maeda, M. M., & Chu, T. (2017). Unmanned aircraft system-derived crop height and normalized difference vegetation index metrics for sorghum yield and aphid stress assessment. *Journal of Applied Remote Sensing*, *11*, Article 026035. https://doi.org/10.1117/1.JRS.11.026035

Stepanov, A., Dubrovin, K., Sorokin, A., & Aseeva, T. (2020). Predicting soybean yield at the regional scale using remote sensing and climatic data. *Remote Sensing*, *12*, Article 1936. https://doi.org/10.3390/rs12121936

Thavarajah, D., Lawrence, T., Boatwright, L., Windsor, N., Johnson, N., Kay, J., Shipe, E., Kumar, S., & Thavarajah, P. (2023). Organic dry pea (*Pisum sativum* L.): A sustainable alternative pulse-based protein for human health. *PLoS ONE*, *18*, Article e0284380. https://doi.org/10.1371/journal.pone.0284380

Thenkabail, P. S., Smith, R. B., & De Pauw, E. (2000). Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. *Remote Sensing of Environment*, *71*, 158–182. https://doi.org/10.1016/S0034-4257(99)00067-X

Wang, M., Yue, L., Yang, X., Wang, X., Han, Y., & Yu, B. (2021). Fertility-LightGBM: A fertility-related protein prediction model by multi-information fusion and light gradient boosting machine. *Biomedical Signal Processing and Control*, *68*, Article 102630. https://doi.org/10.1016/j.bspc.2021.102630

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*, 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Wu, H., Yan, Y., Lin, G., Yang, M., Ng, M. K., & Wu, Q. (2020). Iterative refinement for multi-source visual domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, *34*, 2810–2823.

Xie, Q., Dash, J., Huang, W., Peng, D., Qin, Q., Mortimer, H., Casa, R., Pignatti, S., Laneve, G., Pascucci, S., Dong, Y., & Ye, H. (2018). Vegetation indices combining the red and red-edge spectral information for leaf area index retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*, 1482–1493. https://doi.org/10.1109/JSTARS.2018.2813281

Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. (2005). Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, *97*, 322–336. https://doi.org/10.1016/j.rse.2005.05.008

Yang, S., Li, L., Fei, S., Yang, M., Tao, Z., Meng, Y., & Xiao, Y. (2024). Wheat yield prediction using machine learning method based on UAV remote sensing data. *Drones*, *8*, Article 284. https://doi.org/10.3390/drones8070284

Yunlu, S., Ming, Z., Lianlu, L., Zhaomin, Z., & Zhiyong, W. (2005). Analysis on crude protein content of silage corn in agriculture-animal husbandry ecotone and hyperspectral remote sensing estimation. *Journal of Maize Sciences*, *13*(1), 64–68.

Zha, H., Miao, Y., Wang, T., Li, Y., Zhang, J., Sun, W.-W., Feng, Z., & Kusnierek, K. (2020). Improving unmanned aerial vehicle remote sensing-based rice nitrogen nutrition index prediction with machine learning. *Remote Sensing*, *12*, Article 215. https://doi.org/10.3390/rs12020215

Zhang, Y., Qin, Q., Ren, H., Sun, Y., Li, M., Zhang, T., & Ren, S. (2018). Optimal hyperspectral characteristics determination for winter wheat yield prediction. *Remote Sensing*, *10*, Article 2015. https://doi.org/10.3390/rs10122015

Zhang, J., Qiu, X., Wu, Y., Zhu, Y., Cao, Q., Liu, X., & Cao, W. (2021). Combining texture, color, and vegetation indices from fixed-wing UAS imagery to estimate wheat growth parameters using multivariate regression methods. *Computers and Electronics in Agriculture*, *185*, Article 106138. https://doi.org/10.1016/j.compag.2021.106138

Zhang, Y., Wang, L., Chen, X., Liu, Y., Wang, S., & Wang, L. (2022). Prediction of winter wheat yield at county level in China using ensemble learning. *Progress in Physical Geography: Earth and Environment*, *46*, 676–696. https://doi.org/10.1177/03091333221088018

Zhang, C., Yi, Y., Wang, L., Zhang, X., Chen, S., Su, Z., Zhang, S., & Xue, Y. (2024). Estimation of the bio-parameters of winter wheat by combining feature selection with machine learning using multi-temporal unmanned aerial vehicle multispectral images. *Remote Sensing*, *16*, Article 469. https://doi.org/10.3390/rs16030469

Zhou, X., Kono, Y., Win, A., Matsui, T., & Tanaka, T. S. (2021). Predicting within-field variability in grain yield and protein content of winter wheat using UAV-based multispectral imagery and machine learning approaches. *Plant Production Science*, *24*, 137–151. https://doi.org/10.1080/1343943X.2020.1819165

Zhu, B., Chen, S., Xu, Z., Ye, Y., Han, C., Lu, P., & Song, K. (2023). The estimation of maize grain protein content and yield by assimilating LAI and LNA, retrieved from canopy remote sensing data, into the DSSAT model. *Remote Sensing*, *15*, Article 2576. https://doi.org/10.3390/rs15102576

Zhu, X., Li, X., Liu, X., Li, J., Zeng, X.-A., Li, Y., Yuan, Y., & Teng, Y.-X. (2023). Pulse protein isolates as competitive food ingredients: Origin, composition, functionalities, and the state-of-the-art manufacturing. *Foods*, *13*(1), Article 6. https://doi.org/10.3390/foods13010006

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bazrafkan, A., Worral, H., Bandillo, N., & Flores, P. (2025). Enhancing yield and protein content estimation in dry peas through multispectral data integration and ensemble learning model. *The Plant Phenome Journal*, *8*, e70037. https://doi.org/10.1002/ppj2.70037